A dictionary to identify small molecules and drugs in free text

Hettne, K.M.; Stierum, R.H.; Schuemie, M.J.; Hendriksen, P.J.M.; Schijvenaars, B.J.A.; Mulligen, E.M.; ... ; Kors, J.A.

BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 22 2009, pages 2983–2991
doi:10.1093/bioinformatics/btp535

Data and text mining

A dictionary to identify small molecules and drugs in free text

Kristina M. Hettne[1,2,3,*], Rob H. Stierum[3,4], Martijn J. Schuemie[2], Peter J. M. Hendriksen[5], Bob J. A. Schijvenaars[6], Erik M. van Mulligen[2], Jos Kleinjans[1,3] and Jan A. Kors[2]

[1]Department of Health Risk Analysis and Toxicology, Maastricht University, Maastricht, [2]Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, [3]Department of Toxicoinformatics, Netherlands Toxicogenomics Centre, Maastricht, [4]Business unit Biosciences, Physiological Genomics, TNO Quality of Life, Zeist, [5]Safety and Health, RIKILT Institute of Food Safety, Wageningen, The Netherlands and [6]Collexis Holdings Inc., Columbia SC, USA

Received on May 13, 2009; revised on September 3, 2009; accepted on September 7, 2009

Advance Access publication September 16, 2009

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: From the scientific community, a lot of effort has been spent on the correct identification of gene and protein names in text, while less effort has been spent on the correct identification of chemical names. Dictionary-based term identification has the power to recognize the diverse representation of chemical information in the literature and map the chemicals to their database identifiers.

Results: We developed a dictionary for the identification of small molecules and drugs in text, combining information from UMLS, MeSH, ChEBI, DrugBank, KEGG, HMDB and ChemIDplus. Rule-based term filtering, manual check of highly frequent terms and disambiguation rules were applied. We tested the combined dictionary and the dictionaries derived from the individual resources on an annotated corpus, and conclude the following: (i) each of the different processing steps increase precision with a minor loss of recall; (ii) the overall performance of the combined dictionary is acceptable (precision 0.67, recall 0.40 (0.80 for trivial names); (iii) the combined dictionary performed better than the dictionary in the chemical recognizer OSCAR3; (iv) the performance of a dictionary based on ChemIDplus alone is comparable to the performance of the combined dictionary.

Availability: The combined dictionary is freely available as an XML file in Simple Knowledge Organization System format on the web site http://www.biosemantics.org/chemlist.

Contact: k.hettne@erasmusmc.nl

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Biomedical text mining has been shown to be valuable for diverse applications in the domains of molecular biology, toxicogenomics, and medicine. The techniques behind current text mining applications focus, however, for a great part on the ability of the system to correctly identify gene and protein names in text, while less effort has been spent on the correct identification of chemical names (Erhardt *et al.*, 2006; Kemp and Michael, 1998). Indeed, the domains of genomics and chemistry have developed quite separate from each other, until now, with the important difference that genomic databases and the bioinformatics tools used to mine them arise from an open-source and open-access friendly community while chemistry has a long tradition of closedness and restricted access to data (Murray-Rust *et al.*, 2005; Zimmermann *et al.*, 2005). This is, however, about to change as more and more chemical resources are becoming freely available [e.g. the chemistry search engine ChemSpider (http://www.chemspider.com)], giving rise to the new research field of chemical genomics (Murray-Rust, 2008; Williams, 2008a, b). ChemSpider has an internal dictionary containing links to many public chemical databases and provides web services to access the data. The dictionary is, however, not downloadable and there is no information published on how the dictionary was created and evaluated, which makes it difficult to include it in text mining applications.

Finding biomedical terms in natural language is essential for biomedical text mining. Biomedical named entity recognition (NER) is the task of identifying the boundary of a substring and then map the substring to a predefined category (Zweigenbaum *et al.*, 2007). Approaches to NER generally fall into three categories: dictionary-based systems, rule-based systems and statistically based systems making use of different machine learning techniques (Cohen and Hersh, 2005). The challenges of chemical name identification differ from the ones in the genomics field in the sense that the exact placement of tokens such as commas, spaces, hyphens and parentheses plays a much larger role. Chemical NER in general has been reviewed by Banville (2006) and methods for confidence-based chemical NER have been evaluated by Corbett and Copestake (2008). According to Klinger *et al.* (2008), the only chemical NER software freely available to the academic community is OSCAR3 (http://sourceforge.net/projects/oscar3-chem) (Corbett and Murray-Rust, 2006). OSCAR3 uses a combined NER approach of overlapping 4 g together with a dictionary based on the Chemical Entities (CM) of Biological Interest (ChEBI) ontology (Degtyarenko *et al.*, 2008).

In this article, we focus on the task of term *identification*, which goes beyond NER to also include term *mapping*, i.e. the linking of terms to referent data sources. To achieve this, a dictionary with

*To whom correspondence should be addressed.

database links is essential. For instance, the Whatizit system is able to directly link protein names to their respective UniProt-ID using a dictionary generated from the UniProt database (Rebholz-Schuhmann *et al.*, 2008). Naturally, the usefulness of the dictionary approach depends on the coverage of terms in the dictionary for the particular domain and how well the terms are suited for natural language processing. Recently, resources such as DrugBank (Wishart *et al.*, 2008) and the Unified Medical Language System (UMLS) metathesaurus (Bodenreider, 2004) have been applied for the identification of drug names in text (Kolarik *et al.*, 2007; Segura-Bedmar *et al.*, 2008) [for a recent review of literature mining in support of drug discovery, see Agarwal and Searls (2008)]. In this work, we aim at a broader level of chemical identification where also the organism's own biomolecules such as metabolites and signaling molecules are included, referred to as small molecules in the rest of this article. Dictionary-based approaches aiming at identifying small molecules in text have used different proprietary resources to create their dictionary: Singh *et al.* (2003) used the proprietary Compound Knowledge Base system (Walker *et al.*, 2002), and Zhu *et al.* (2005) used the proprietary Chemical Abstracts Services (CAS) Registry numbers (Weisgerber, 1997). Due to a lack of annotated chemical compound test corpora, before the year 2008 only one study reported the recall and precision of a small-molecule dictionary: Zimmermann *et al.* (2005) evaluated a dictionary consisting of the chemical part of the Medical Subject Headings (MeSH) (Lipscomb, 2000) together with ChEBI using the ProMiner system (Hanisch *et al.*, 2005) on a modified version of the GENIA corpus (Kim *et al.*, 2003), and reported 80% precision and 99% recall. They, however, made it very clear that the high recall was due to the artificial nature of the test corpus (since CM in the GENIA corpus mostly consist of ion names (e.g. Ca^{+}), these entities were removed and replaced by compounds randomly picked from their small-molecule dictionary). Recently, Kolarik *et al.* (2008) created a test corpus consisting of 100 manually annotated PubMed abstracts. Using a simple case insensitive string search, ignoring hyphens, they tested the recall and precision for MeSH headings, MeSH supplementary concept records, ChEBI, PubChem (Wheeler *et al.*, 2008), DrugBank, KEGG drug (Kanehisa *et al.*, 2008), KEGG compound (Goto *et al.*, 2002), Human Metabolome database (HMDB) (Wishart *et al.*, 2009), and for a combined version of the dictionaries, with the goal of gaining knowledge about the suitability for a dictionary with curation effort. The best recall was achieved using the a combination of all resources (precision 13%, recall 49%) and the best precision was achieved using KEGG drug (precision 59%, recall 12%).

The objectives of this study are (i) to create a combined dictionary to identify small molecules and drugs in free text, and (ii) to study the impact on precision and recall of term rewrite and suppress rules, manual check of highly frequent terms and disambiguation rules.

2 METHODS

2.1 Choice of chemical resources

We focused on freely available and downloadable terminology resources containing small molecules from the context of human studies. A description of resources included is provided below.

2.1.1 Chemicals from a broad chemical space The UMLS (http://www.nlm.nih.gov/research/umls/) contains information about biomedical and health-related concepts, their various names and the relationships among them. It is provided by the US National Library of Medicine (NLM). All entities, henceforth referred to as concepts, in the UMLS are assigned a unique concept identifier (CUI); the terms belonging to the concept are, in turn, assigned a unique term identifier (LUI), a unique string identifier (SUI) and a unique atom identifier (AUI). We extracted concepts based on the CUI and terms based on the SUI. In addition, every concept has at least one semantic type from the Semantic Network (http://www.nlm.nih.gov/pubs/factsheets/umlssemn.html) assigned to it. These semantic types have been aggregated into semantic groups (McCray *et al.*, 2001b). Similar to Wilbur *et al.* (1999), we used the semantic types belonging to the semantic group 'Chemicals & Drugs' and removed the types T120 'Chemical Viewed Functionally', T122 'Biomedical and Dental Material' and T192 'Receptor'. In contrast, we excluded the semantic types T200 'Clinical Drug' (e.g. 'fluorescein 250 mg/ml injectable solution [fluorescein lite]'), T126 'Enzyme' (e.g. 'Kininase III'), T116 'Amino Acid, Peptide or Protein' (e.g. 'alpha 1-antitrypsin-leukocyte elastase complex') and T103 'Chemical' (e.g. 'Chemicals'), and in addition, we added the semantic type T129 'Immunologic Factor' (e.g. 'Efalizumab'). The different choice of removal or inclusion of semantic types compared with Wilbur *et al.* (1999) was determined based on a manual analysis of a random set of 100 terms from each semantic type, with the criteria that the terms should mainly represent small molecules or drugs and be likely to be found in text. Since the UMLS does not contain CAS numbers or InChI strings, the concepts were mapped to CAS numbers via the MeSH identifier in the UMLS. The resulting dictionary will be referred to as UMLSchem.

MeSH (http://www.nlm.nih.gov/mesh/) is a controlled vocabulary thesaurus from the NLM. The terms are organized in a hierarchy to which synonyms as well as inflectional term variants are assigned. Similar to the UMLS, every concept in MeSH has a semantic type attached to it. We extracted records concerning small molecules from MeSH by filtering for the same semantic types as we used for the UMLS. We will refer to this dictionary as MeSHchem.

MeSH supplemental concept records (http://www.nlm.nih.gov/mesh/) are used to index chemicals, drugs and other concepts for MEDLINE and are searchable by Substance Name [NM] in PubMed. We extracted records concerning small molecules from MeSH by filtering for the same semantic types as we used for the UMLS and MeSH. We will refer to this dictionary as MeSHsupp.

ChEBI (http://www.ebi.ac.uk/chebi/) is an ontology of molecular entities, hosted by the European Bioinformatics Institute.

PubChem (http://pubchem.ncbi.nlm.nih.gov/) is a component of the US National Institutes of Health's Molecular Libraries Roadmap Initiative and is organized as three linked databases (PubChem Substance, PubChem Compound and PubChem BioAssay) within the NCBI's Entrez information retrieval system. PubChem Substance is a chemical repository with little or no manual check and curation of the records. PubChem Compound is a subset of PubChem Substance which contains validated chemical depiction information but no chemical synonyms. In order to retrieve high-quality information while at the same time incorporating as many synonyms as possible, a PubChem subset dictionary was made consisting of the PubChem Substance records that contain a link to a PubChem Compound entry.

2.1.2 Drug terminology DrugBank (http://www.drugbank.ca/) combines detailed drug data with drug target information. It is provided by the University of Alberta.

KEGG drug (http://www.genome.jp/kegg/drug/) is a chemical structure-based information resource for all approved drugs in the US and Japan. It is maintained by the Kanehisa Laboratories. We will refer to this dictionary as KEGGd.

2.1.3 Metabolic substances KEGG compound (http://www.genome.jp/kegg/compound/) is a database for metabolic compounds and other chemical substances that are relevant to biological systems. It is maintained by the Kanehisa Laboratories. We will refer to this dictionary as KEGGc.

HMDB (http://www.hmdb.ca/) contains detailed information about small molecule metabolites found in the human body. HMDB is provided by the University of Alberta.

2.1.4 Toxic substances ChemIDplus (http://www.nlm.nih.gov/pubs/factsheets/chemidplusfs.html) is a web-based search system that provides access to structure and nomenclature authority files used for the identification of chemical substances cited in NLM databases, including the TOXNET system. NLM provides a ChemIDplus subset for download which does not include the structure or the toxicity data available from the NLM's online version of the database.

2.2 Data extraction

All data was downloaded on November 4, 2008. Since we aim to create a dictionary for small molecules and drugs, it is desirable that each separate record in the dictionary represents a unique substance. There are currently two accepted standards that provide unique identifiers for chemical substances: CAS Registry Numbers [proprietary, assigned by the CAS registry (http://www.cas.org/)] and InChI strings [non-proprietary, developed by International Union of Pure and Applied Chemistry (IUPAC) (http://www.iupac.org/inchi/)]. Only records containing CAS numbers or InChI strings were included in the extracted versions of the databases. Non-English terms [term contained a non-English language or a non-English country at the end of the term, e.g. 3,4-Benzopirene (Italian)] and terms longer than 255 characters were removed. If a term contained the name of the original vocabulary or pharmaceutical company (for drugs) at the end of the term [e.g. Goserelin acetate (JAN/USP), Wellferon (GlaxoSmithKline)], this part was removed. For each database, we extracted the data from the fields used for entry term, synonyms, summary structures and identifiers (Supplementary Material 1). If available, the entry term was set as preferred term, otherwise the first synonym was used. After extraction, all resources were transformed into the Simple Knowledge Organization System (SKOS) thesaurus format (http://www.w3.org/TR/skos-reference/). SKOS provides a standard way to represent knowledge organization systems using the Resource Description Framework.

2.3 Dictionary pre-processing

We have previously investigated the effect of a number of rewrite and suppress rules, collectively called *filtering rules*, on the terms in the UMLS (K.M.Hettne *et al*., submitted for publication). The number of uniquely identified terms and their frequency in MEDLINE were computed before and after applying the rules. The 50 most frequently found terms together with a sample of 100 randomly selected terms were evaluated per rule. Using the rewrite rules that passed our evaluation, we were able to identify 1 117 772 new occurrences of 14 784 rewritten terms, and using the suppress rules that passed our evaluation, a total of 257 118 were suppressed in the UMLS. We also implemented a software tool to apply these rules to the UMLS (http://biosemantics.org/casper). We decided to use the rules suitable for chemical terms to rewrite and suppress terms in the chemical dictionaries. The rules are listed and explained below together with references to the original sources.

Short token filter rule (McCray *et al*., 2001a; Rogers and Aronson, 2008): remove term if the whole term after tokenization and removal of stop words is a single character, or is an Arabic or Roman number (e.g. 'T' as an abbreviation for 'Tritium'). For this rule, the stop word list from PubMed (http://www.ncbi.nlm.nih.gov/books/bv.fcgi?highlight=stopwords&rid=helppubmed.table.pubmedhelp.T43) was used. This rule resembles the one mentioned in McCray *et al.* (2001a) and Rogers and Aronson (2008) with the difference that it takes each token into account separately.

Dosages rule (McCray *et al.*, 2001a): the original rule addressed terms belonging to certain term types in the UMLS, namely BD (fully specified drug brand name that can be prescribed), CD (Clinical Drug) or MS (Multiple names of branded and generic supplies or supplements). This rule was further refined by us to remove all terms that contain a dosage in percent, gram, microgram or milliliter (e.g. 'Theophylline 0.4% and dextrose 5% in plastic container' as a synonym for 'Theophylline').

At-sign rule: this rule was implemented by us to remove terms that contain the @-character (e.g. 'sNqDLLQxbRvuUQX@' as a synonym for '1,4-dibromobutan-2-ol').

Any underspecification rule (McCray *et al.*, 2001a; Rogers and Aronson, 2008): remove terms that contain any of the following features: 'not otherwise specified', 'not specified' or 'unspecified'; 'NOS' at the end of a term and preceded by a comma, or 'NOS' within parentheses or brackets at the end of a term and preceded by a space (e.g. 'unspecified phosphate of chloroquine diphosphate' as synonym for 'chloroquine diphosphate').

Miscellaneous rule (McCray *et al.*, 2001a; Rogers and Aronson, 2008): remove terms that contain the following features: 'other' at the beginning of a term and followed by a space character or at the end of a term and preceded by a space character, 'deprecated', 'unknown', 'obsolete', 'miscellaneous' or 'no' at the beginning of a term and followed by a space character (e.g. 'no stereochem' as synonym for 'Encainide').

Syntactic inversion rule (McCray *et al.*, 2001a; Rogers and Aronson, 2008): add syntactic inversion of term if a term contains a comma followed by a space and does not contain a preposition or conjunction (e.g. 'acid, gamma-vinyl-gamma-aminobutyric' is rewritten to 'gamma-vinyl-gamma-aminobutyric acid'). We added the condition that only one such pattern of a comma followed by a space is to be found in a term for the rule to be executed.

Possessives rule (McCray *et al.*, 2001a; Rogers and Aronson, 2008): remove the possessive 's' at the end of a term (e.g. 'Ringer's lactate' rewritten as 'Ringer lactate') and add the rewritten term.

Short form/long form rule (Schwartz and Hearst, 2003): add short form and long form of term [e.g. 'Hydrogen chloride (HCL)' is split into 'Hydrogen chloride' and 'HCL']. The rule is based on the abbreviation finding algorithm described by Schwartz and Hearst (2003). The algorithm achieved 96% precision and 82% recall on a standard test collection, which was as good as existing approaches at the time (Schwartz and Hearst, 2003) and still competitive according to recent comparison studies (Torii *et al.*, 2007; Xu *et al.*, 2009). An advantage of the algorithm is that, unlike other approaches, it does not require any training data. Two extra conditions were added to the original rule by Schwartz and Hearst: (i) the short form must be found at the end of the term, and (ii) the first letter of the short form should be the same as the first letter of the long form. These conditions were added in order to adjust the rule to extract abbreviations from a dictionary instead of from biomedical text.

2.3.1 Manual check of highly frequent terms A set of 100 000 randomly selected MEDLINE abstracts were indexed (see Section 2.5) with each dictionary, and the top 500 most frequent terms found in the set per dictionary were selected for manual evaluation. If they corresponded to a general English term (e.g. 'access'), they were added to a master list of unwanted terms. This master list was then used to filter all dictionaries separately.

2.4 Data resource combination

We merged entries if they had the same CAS numbers [similar to Zimmerman *et al.* (2005)], database identifier, or InChI string.

2.5 Identification of chemical names

For the term and concept identification, we used our concept recognition software Peregrine (Schuemie *et al.*, 2007a). The Peregrine system was designed with two goals in mind. First of all, it should be easy to maintain. There is only a single step (manual check of highly frequent terms) that requires human involvement when implementing a new lexicon. The second

goal was speed. Because Peregrine does not rely on part-of-speech tagging or natural language parsing, it is very fast: 100 000 MEDLINE records can be processed in 213 s on a standard PC. The whole of MEDLINE can be processed within a single day (Morgan *et al.*, 2008). The Peregrine system translates the terms in the dictionary into sequences of tokens or words. When such a sequence of tokens is found in a document, the term, and thus the chemical associated with that term, is recognized in the text. Some tokens are ignored, since these are considered to be non-informative ('of', 'the', 'and' and 'in'). We used Peregrine with the following settings: case-insensitive, word-order sensitive and largest match. In its default setting, the tokenizer in Peregrine considers everything that is not a letter or a digit to be a word delimiter. To fine-tune the tokenizer for chemical concept recognition we made the following adjustments: full stops, commas, plus signs, hyphens, single quotation marks and all types of parentheses ((, {, [) were excluded from the word delimiter list. After tokenization, the tokens were stripped of trailing full stops, commas and non-matching parentheses. Parentheses were also removed if they surrounded the whole token. In addition, a list of common suffixes was used to remove these suffixes at the end of tokens (Supplementary Material 2). The suffix list was obtained by scanning the whole UMLS (i.e. not just the chemical part) for suffixes that were English verbs or adjectives.

2.5.1 Disambiguation rules Disambiguation of terms is important since terms not only can have different meanings ('word senses') in a dictionary but also in text (e.g. 'BAP' is a shared synonym between the two chemicals 'Benzo(a)pyrene' and 'Benzyladenine' and has an additional 44 meanings according to Acronym Finder (http://www.acronymfinder.com), including 'Blood Agar Plate', 'BiP-Associated Protein' and 'British Association of Psychotherapists'). Word-sense disambiguation algorithms can be distinguished as supervised, unsupervised or using established knowledge (Alexopoulou *et al.*, 2009; Edmonds and Agirre, 2006). Peregrine uses established knowledge to disambiguate terms on the fly during the indexation process. Specifically, Schuemie *et al.* (2007a, 2007b) evaluated a number of rules to disambiguate gene names found in text. These disambiguation rules are potentially also applicable to chemical names. Disambiguation of terms found in text was carried out as follows (Fig. 1). We first determine whether a term is a dictionary homonym, i.e. if it refers to more than one entity in the dictionary. If the term is a dictionary homonym, but it is the preferred term of that entity, it is further handled as if it is not a dictionary homonym. If the term is not a dictionary homonym it still needs further processing since it can have many meanings in text. Therefore, terms that are not complex (i.e. longer than five characters or containing a number) are also considered potential homonyms, and require extra information to be assigned. A (potential) homonym is only kept if (i) another synonym of the entity is found in the same piece of text; (ii) a keyword (i.e. a word or 'token' that occurs in any of the long-form names of the small molecule, and appears less than 1000 times in the dictionary as a whole) is found in the same piece of text.

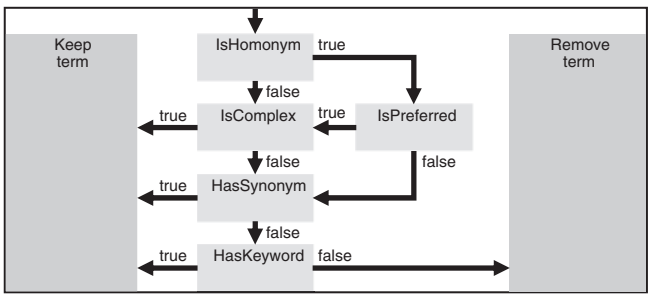

Fig. 1. Term disambiguation scheme.

2.6 Annotated test corpus

The annotated corpus (http://www.scai.fraunhofer.de/chem-corpora.html) from Kolarik *et al.* (2008) was used to test the chemical dictionaries. The corpus consists of 100 MEDLINE abstracts with 1206 annotated chemical occurrences divided into the following groups: multi-word systematic names (IUPAC, 391 occurrences), partial chemical names (PART, 92 occurrences), sum formulas (SUM, 49 occurrences), trivial names (including single word IUPAC names) (TRIV, 414 occurrences), abbreviations (ABB, 161 occurrences) and chemical family names (FAM, 99 occurrences). Larger drug molecules such as protein drugs were not annotated. See Kolarik *et al.* (2008) for details on the creation of the corpus.

3 RESULTS

3.1 Dictionary characteristics

The number of concepts in the dictionaries before any processing and removal of concepts that did not have a CAS number or InChI string were the following: ChEBI 20 606; ChemIDplus 367 358; DrugBank 4776; HMDB 6892; KEGGc 13 543; KEGGd 7737; MeSHchem 6831; MeSHSupp 100 198; PubChem 3 987 338; UMLSchem 197 578. Table 1 shows the characteristics of the different dictionaries after applying filtering and manual check of highly frequent terms. No dictionary was completely covered by another which justifies a combination of all dictionaries (Supplementary Material 3). Most dictionaries contain non-unique records, i.e. two or more records with the same CAS number or InChI string. These records were merged when the combined dictionary was created. The number of terms affected by the filtering rules and manual check of highly frequent terms per dictionary can be found in Supplementary Material 4. The master list of unwanted terms from the manual check of highly frequent terms that was used to filter all the dictionaries (258 terms) can be found in Supplementary Material 5.

Table 1. Contents of the different vocabularies after removal of concepts lacking a CAS number or InChI string and application of filter rules and manual check of highly frequent terms[a]

Dictionary	Concepts	Terms	CAS numbers	InChI strings
ChEBI	11 428	65 409	6436 (6295)	11 212 (11 152)
ChemIDplus	260 393	1 378 808	260 393 (260 393)	–
DrugBank	4540	37 508	2240 (2218)	4381 (4208)
HMDB	6859	75 957	2683 (2537)	6857 (6734)
KEGGc	11 976	31 143	7695 (7661)	11 875 (11 738)
KEGGd	6927	18 697	6769 (6670)	6140 (6083)
MeSHchem	2897	29 023	2897 (2897)	–
MeSHsupp	19 137	92 918	19 137 (19 137)	–
PubChem	383 043	2 121 960	420 737 (395 108)	16 222 (16 108)
UMLSchem	47 508	126 470	47 509 (18 703)	–
Combined	377 849	2 600 445	400 899 (400 899)	50 254 (50 254)

Notably, PubChem contains more unique CAS numbers than unique concepts. There can be various reasons for the 'extra' CAS numbers for a compound. For example, the CAS registry may assign different CAS numbers for the same compound based on properties such as purity, polymorphism, or country of registration.

[a]The numbers in parentheses refer to the number of unique CAS numbers or InChI strings.

Table 2. Precision (P), recall (R) and *F*-score (F) of the dictionaries for the annotated corpus

Dictionary	Unprocessed			Filtered			Curated			Disambiguation			Kolarik		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
ChEBI	0.21	0.28	0.24	0.58	0.28	0.38	0.63	0.28	0.39	0.71	0.25	0.37	0.13	0.27	0.18
ChemIDplus	0.27	0.41	0.33	0.43	0.40	0.41	0.60	0.40	0.48	0.71	0.37	0.49	–	–	–
DrugBank	0.40	0.22	0.28	0.50	0.22	0.31	0.70	0.21	0.32	0.77	0.19	0.30	0.33	0.13	0.19
HMDB	0.21	0.22	0.21	0.57	0.21	0.31	0.66	0.21	0.32	0.71	0.18	0.29	0.21	0.16	0.18
KEGGc	0.43	0.25	0.32	0.58	0.25	0.35	0.70	0.25	0.37	0.72	0.23	0.35	0.30	0.24	0.27
KEGGd	0.63	0.16	0.26	0.73	0.16	0.26	0.76	0.16	0.26	0.78	0.16	0.27	0.59	0.12	0.20
MeSHchem	0.70	0.23	0.35	0.70	0.23	0.35	0.74	0.23	0.35	0.75	0.22	0.34	0.34	0.27	0.30
MeSHsupp	0.75	0.08	0.14	0.75	0.08	0.14	0.82	0.08	0.15	0.83	0.07	0.13	0.15	0.10	0.12
PubChem	0.24	0.47	0.32	0.39	0.47	0.43	0.58	0.47	0.52	0.73	0.35	0.47	0.15	0.33	0.21
UMLSchem	0.43	0.32	0.37	0.62	0.32	0.42	0.74	0.32	0.45	0.78	0.29	0.42	–	–	–
Combined (PubChem included)	0.18	0.49	0.26	0.36	0.49	0.42	0.51	0.49	0.50	0.62	0.39	0.48	0.13	0.49	0.21
Combined (PubChem excluded)	0.20	0.47	0.28	0.39	0.46	0.42	0.55	0.46	0.50	0.67	0.40	0.50	–	–	–

For comparison, the results from Kolarik *et al.* (2008) have also been included.

3.2 Dictionary performance

Dictionary term strings that matched the start and end positions of the chemical term strings in the corpus constituted true positives (TP), dictionary term strings that did not match were false positives (FP) and chemical term strings in the corpus that were not matched were false negatives (FN). Recall (R), precision (P) and *F*-score were computed in the usual way:

- Recall = TP/(TP + FN)
- Precision = TP/(TP + FP)
- *F*-score = (2 × P × R)/(P + R)

Table 2 shows the effect of preprocessing and disambiguation on precision and recall for each of the dictionaries. The values reported by Kolarik *et al.* (2008) are also shown, if available. It is clear that the preprocessing steps and the disambiguation rules have a strong positive influence on the precision of all dictionaries. We also achieve higher recall and precision than Kolarik *et al.* (2008) for most dictionaries even in the unprocessed stage, which may be explained by updates of the dictionaries since the study by Kolarik *et al.* (2008), by our additional criteria to only include entities with a CAS number or InChI string, and by our refined search strategy. The combined version of all dictionaries after executing all the preprocessing steps and disambiguation rules had the highest recall (0.39) but the lowest precision (0.62) compared with all the separate dictionaries using disambiguation (Table 2), which led us to investigate the possibility to exclude resources with low precision to further improve the precision of the combined dictionary without loss of recall. PubChem had the lowest precision (0.58) of all dictionaries before application of the disambiguation rules, which raises questions about the quality of the data. Indeed, in a recent publication by Richard *et al.* (2006) concerning chemical information available in databases and through search engines, the quality of chemical information in PubChem was described as 'user beware'. Also Williams (2008b) expressed concerns about the accuracy of some of the identifiers associated with PubChem compounds. In addition, all resources that we used claim to perform manual curation of the data except for PubChem. When PubChem was left out of the combined dictionary it achieved a precision of 0.67 and a recall of 0.40, both higher than for the combined dictionary without PubChem. When removing the dictionary with the second lowest precision before disambiguation (ChemIDplus: 0.60), the precision of the combined dictionary rose to 0.69 but at the cost of lower recall (0.37) (for comparison, ChemIDplus alone achieved better precision with the same recall; Table 2). Since the removal of PubChem from the combined dictionary improved both the recall and the precision, the combined dictionary without PubChem was used for further analysis. Notably, the curated dictionary with disambiguation rules applied has much higher precision (0.67) than the combined dictionary reported by Kolarik *et al.* (2008) (0.13), with a difference in recall of 9 percentage points. The combined dictionary without PubChem contains 1 692 020 terms belonging to 278 577 concepts. Of these concepts, 266 705 have a CAS number and 34 146 have an InChI string. The curated combined dictionary (PubChem excluded) with disambiguation rules applied had the highest *F*-score ($F = 0.50$) at a reasonable precision (0.67), closely followed by the curated version of ChemIDplus with disambiguation rules applied ($F = 0.49$, precision = 0.71). Overall, the recall was best for the TRIV class of entities (Supplementary Material 6), with ChemIDplus as the best performing dictionary (recall 0.82) and the combined dictionary (PubChem excluded) as a close number two (0.80). The PART class of entities had the lowest recall of all classes (0.00) with the combined dictionary (PubChem excluded) and PubChem as the best performing dictionary (0.04). The PART class is, however, more relevant when the corpus is going to be used for machine learning purposes since parts of chemical names are not expected to be found in dictionaries. This class was, therefore, left out of the error analysis in Section 3.3.

To investigate the effect of a general normalization procedure on chemical terms, we ran an analysis using the normalization program *norm* that comes with the LVG normalizer (McCray *et al.*, 1994). The LVG normalizer operates after the tokenization has taken place but before disambiguation of terms. The normalization procedure constitutes lower casing each token, converting each token to its base form, ignoring punctuation and sorting the tokens

in a multi-token term into alphabetic order. The analysis run resulted in one percentage point lower precision and one percentage point higher recall. The additional terms resulting in the higher recall for the combined dictionary, however, all corresponded to family names being mapped to a single chemical in the dictionary (e.g. diphenols mapped to diphenol), which for the purpose of term identification is to be considered an error. The lower precision was caused by the removal of punctuation, a very important feature of chemical terms, which introduces unnecessary homonyms in the dictionary [e.g. '(-)-Catechol' (CAS 18829-70-4) becomes the same as 'Catechol' (CAS 120-80-9)]. To further illustrate the importance of punctuation in chemical term identification, we ran an analysis using the original tokenizer in Peregrine (Section 2.5). This run resulted in a precision of 0.42 and a recall of 0.40, the much lower precision mainly arising from erroneous partial mapping of terms. In contrast, the original tokenizer in Peregrine has produced good results (precision 0.75, recall 0.76) for a combined dictionary of gene names on the BioCreAtIvE 2 test set (Schuemie *et al.*, 2007a). We used an updated version of the combined dictionary of gene names on the same BioCreAtIvE 2 test set with the two different tokenizers, resulting in a precision of 0.74 and recall of 0.81 (F = 0.77) for the original tokenizer and a precision of 0.76 and a recall of 0.79 (F = 0.77) for the modified tokenizer. Judged by these results, punctuation is less important for gene names than for chemical names.

To compare a pure dictionary-based term identification approach with a combined NER approach, we ran OSCAR3 on the corpus. To make the comparison as fair as possible, only CM were counted, thus excluding the other entity classes in OSCAR3 (ASE = enzyme, CPR = chemical prefix, RN = reaction, CJ = chemical adjective, ONT = ontology term). Using this approach, OSCAR3 had a precision of 0.45 and a recall of 0.82 on the corpus, giving an F-score of 0.58. If only entities that had been mapped to the dictionary in OSCAR3 (we will refer to these as OSCAR3_dict) were taken into account, the system achieved a precision of 0.68 and a recall of 0.25, giving an F-score of 0.37, comparable to the curated version of ChEBI in our approach with disambiguation rules applied. Recall values for the different entity classes are presented in Table 3. The curated combined dictionary had the highest recall value for the TRIV class of entities, which also was the highest for that class for all approaches. OSCAR3_dict scored higher than the curated combined dictionary for the PART and FAM classes of entities. OSCAR3 had a high recall over all entity classes.

Table 3. Recall values for the entity classes as defined by Kolarik *et al.* (2008) using the curated combined dictionary with disambiguation rules applied (=Combined), OSCAR3 (=OSCAR3) and the dictionary in OSCAR3 (=OSCAR3_dict)

Entity class	Combined	OSCAR3	OSCAR3_dict
IUPAC (391)	0.21	0.82	0.08
PART (92)	0.04	0.84	0.10
SUM (49)	0.29	0.82	0.00
TRIV (414)	0.80	0.79	0.50
ABB (161)	0.22	0.84	0.08
FAM (99)	0.19	0.84	0.44

3.3 Error analysis

We performed a manual error analysis for the combined curated dictionary with disambiguation rules applied and the results from OSCAR3 and OSCAR3_dict. A random set of maximum 25 false negatives from each class (Table 4) and a random set of 50 false positives (Table 5) were analyzed for each approach. We defined six error categories for the false negatives: *partial match* (e.g. only 'azaline' in 'azaline B' was recognized); *annotation error* (e.g. only part of the chemical name has been marked in the text: 'thiophen' in 'thiophene'); *not in dictionary*; *removed by disambiguation* (e.g. single letter 'T'); *removed by manual check of highly frequent terms* (e.g. 'acid'); and *tokenization error* [e.g. 'Ca(2+)' will not be found in the sentence '… free calcium concentration ([Ca(2+)]i) of human peripheral blood lymphocytes …' due to the positioning of the 'i' that does not allow the surrounding brackets to be removed from the entity]. For the false positives, we defined four error categories: *partial match*; *annotation error*; *out of corpus scope* (e.g. larger drug molecules such as protein drugs); *not a chemical* (e.g. 'n = 34' was tokenized and mapped to 'N 34', which is a synonym for Calcium Carbonate). The major reason that entities were not found (i.e. were false negatives) was that they simply were not in the combined curated dictionary or the dictionary in OSCAR3_dict, or for OSCAR3, were not recognized by the NER algorithm (Table 4). For the combined curated dictionary, this holds true for all classes except ABB, for which a larger part was removed during the disambiguation step. This is not surprising since abbreviations are notoriously ambiguous and difficult to resolve. For OSCAR3, the

Table 4. Error analysis of a random sample of max 25 false negatives from each class for the combined curated dictionary (PubChem excluded) with disambiguation rules applied (=Comb.), OSCAR3 (=OSC) and the dictionary part of OSCAR3 (=OSC_d)

Error type	TRIV			SUM			IUPAC			FAM			ABB		
	Comb	OSC	OSC_d	Comb	OSC	OSC_d	Comb	OSC	OSC_d	Comb	OSC	OSC_d	Comb	OSC	OSC_d
Partial match	3	1	3	0	0	0	0	23	3	0	4	2	0	3	0
Annotation error	2	2	3	0	0	1	1	2	1	0	0	0	0	0	0
Not in dictionary/recognized	15	21	19	16	0	22	24	0	21	24	12	23	8	18	25
Removed by disambiguation	5	0	0	7	0	0	0	0	0	1	0	0	12	0	0
Removed by manual check of highly frequent terms	0	0	0	1	0	0	0	0	0	0	0	0	2	0	0
Tokenization error	0	1	0	1	9	2	0	0	0	0	0	0	3	4	0

Table 5. Error analysis of a random sample of 50 false positives for the combined curated dictionary (PubChem excluded) (=Combined) with disambiguation rules applied, OSCAR3 (=OSCAR3) and the dictionary part of OSCAR3 (OSCAR3_dict)

Error type	False positives		
	Combined	OSCAR3	OSCAR3_dict
Partial match	15	9	20
Annotation error	6	6	9
Out of corpus scope	21	13	16
Not a chemical	8	22	5

exceptions are instead the IUPAC class, where a majority of the false negatives were only partially found and the SUM class for which the tokenizer performed poorly. For the false positives, it was clear that the corpus is not optimal for a dictionary that aims at both small molecules and drugs, since larger drug molecules have not been annotated in the corpus. This was true for 42% of the entities in the random set for the combined dictionary approach, 32% of the entities in the random set for the OSCAR3_dict and 26% of the entities in the random set for OSCAR3 (Table 5). Another major source for the false positives using all approaches was partial matches of longer chemical names. For OSCAR3, it can be noted that it recognized a higher percentage of non-CM than the combined dictionary and OSCAR3_dict.

4 DISCUSSION

For all dictionaries, the best *F*-scores in combination with high precision are reached with the disambiguation rules applied. Disambiguation is, therefore, of high importance when the dictionaries are to be used for text mining purposes. The combined curated dictionary (excluding PubChem) with disambiguation rules applied had the best *F*-score of all of the separate curated dictionaries with disambiguation rules applied, even better than PubChem and ChemIDplus which themselves are made up of combinations of different resources. Still, the good performance of the combined dictionary can be weighted against the time-consuming process of downloading, curating and combining all the different resources. The best alternative to a combined dictionary would be ChemIDplus, which showed a minor difference in performance compared with the combined dictionary. The downloadable version of ChemIDplus does, however, not contain InChI strings.

The largest part of the false positives could be contributed to the fact that not all chemicals were tagged in the corpus. Even though the corpus is a welcome initiative, it is not ideal for the testing of a dictionary that is a combination of small molecules and drugs since large drug molecules such as protein drugs are not annotated. The other major factor that caused false positives was that parts of chemical terms were recognized as whole entities. This happened because the dictionary did not contain the larger term. A way around this would be to first determine the boundaries of a chemical and then map it to a dictionary. This, however, seems to only partly solve the problem since even though OSCAR3 uses such an approach, it scored high in the partial match error category for both the false negatives (class IUPAC) and the false positives. The fact that more than half of the false positives were caused by problems that have nothing to do with the dictionary (entity out of corpus scope, or annotation error), put the relatively low precision of 0.67 in a different light. If these false positives would be excluded from the analysis, the combined dictionary would have a precision of 0.90.

According to our study, a recall of 0.49 (at a precision of 0.51) would be the highest achievable recall for a pure dictionary approach to term recognition and mapping. This is the recall reached by the curated combined dictionary (PubChem included) without disambiguation rules applied. Kolarik *et al.* (2008) reached the same recall at a precision of 0.13. If higher recall is desired, an approach such as has been implemented in OSCAR3, i.e. a combined NER approach using machine learning together with a dictionary, would be the better choice. This approach has, however, the disadvantages of lower precision (at least on the corpus used in this study) and an incomplete mapping of entities to external data sources. The precision of OSCAR3 on the corpus (0.45) is lower than what has been reported by Corbett *et al.* (2007) on a non-public PubMed corpus (0.75), but the recall (0.82) is better (Corbett *et al.* reported a recall of 0.74). Notably, many (44%) of the false positives arising from OSCAR3 fell under the non-CM error category. These were mainly abbreviations of entities such as 'CNS' for 'Central Nervous System' or 'AD' for 'Alzheimer Disease' or text structures that resemble CM such as '11*a–c*' or 'IC(50)'. The false positives arising from non-chemical abbreviations could possibly be removed with the use of the disambiguation rules described in this study. If the false positives that were due to corpus mismatch and annotation errors are removed from the calculation, the precision is still lower (57.3%) but at least closer to the one earlier reported. The difference in precision and recall can be due to differences in the annotation scheme of CM underlying the training corpus used in OSCAR3 and the corpus by Kolarik *et al.* (2008). The dictionary in OSCAR3 had a lower recall than the combined dictionary [precision 0.68 (0.84 when corrected for corpus mismatch and annotation errors) and recall 0.25 versus precision 0.67 (0.90 when corrected for corpus mismatch and annotation errors) and recall 0.40], which suggests that the dictionary in OSCAR3 would benefit from a combined dictionary approach. However, embedding the combined dictionary from this study in OSCAR3 is out of the scope of this article and we suggest this for future research.

In our study and in the study by Kolarik *et al.* (2008), an important class with low recall was IUPAC. The main reason for not finding these entries was that they simply were not present in the combined dictionary, even though IUPAC-like names had been added when available. Clearly, dictionary-based term identification is not capable of identifying multiple-term systematic names to a satisfactory extent since not enough of these types of names are available in current resources. If only a synonym for an entity is missing, this might be solved by term variant generation but if the whole entity is missing from the dictionary it can only be solved by adding the entity to the dictionary. Spelling errors might be helped by fuzzy matching (Bingjun *et al.*, 2007, 2008; Chen *et al.*, 2007; Schulz *et al.*, 2006), with a possible cost to precision. In contrast, machine-learning or rule-based systems have reported good performance for the recognition of multiple-term systematic names [e.g. Klinger *et al.* reported an *F*-score of 0.82 on a PubMed corpus for their method based on conditional random fields (CRFs) and CRFs was also used in a high proportion of entries in the latest BioCreative

evaluation (Smith *et al.*, 2008), OSCAR3 had a recall of 0.82 for the IUPAC class of entities on the corpus used in this study, Corbett and Copestake an *F*-score of 0.83 for a system of cascaded classifiers on a PubMed corpus and Wren (2006) a recall of 0.93 with an average precision of 0.83 (depending upon the cutoff score used) for a first order Markov model on a PubMed corpus], but then the problem remains of mapping a term to its referent data source.

The lower recall of 0.40 for the combined dictionary with the disambiguation rules applied compared to without disambiguation is foremost due to the problem associated with the disambiguation of abbreviations and summary structures. Yu *et al.* (2007) divided the problem of disambiguating abbreviations into two types. First, abbreviations may be disambiguated ('defined') near their occurrence in the text. The second type of abbreviation appears without the intended full form nearby. This second type of abbreviation is more prevalent and harder to disambiguate (Yu *et al.*, 2002, 2007). Abbreviations and summary structures of chemicals are of the second type, in the sense that they are used in abstracts to a large extent without the long form of the term, which will cause these entities to be removed since there is not enough extra information to make sure that they actually represent a CM. Using full text articles instead of abstracts might be an answer but unfortunately there has been a report of high (75%) occurrence of abbreviations without their long forms also in full text articles (Yu *et al.*, 2002). To resolve this, another way of taking the context into account is needed, using for example document labeling. If a document is labeled, a term could be assigned directly if it was not an in-dictionary homonym.

5 CONCLUSIONS

In this article, we present a method to prepare a chemical dictionary for dictionary-based text mining. We conclude that preprocessing of terms with limited manual check of highly frequent terms together with disambiguation rules increase precision with a minor loss of recall, leading to a an acceptable overall performance for a combined dictionary. In addition, the combined dictionary performed better than the dictionary in the state-of-the-art chemical recognizer OSCAR3. We also conclude that ChemIDplus performs almost as well as a combined version of all dictionaries.

Funding: Dutch Technology Foundation STW, applied science division of NWO; Technology Program of the Ministry of Economic Affairs.

Conflict of Interest: none declared.

REFERENCES

Agarwal,P. and Searls,D.B. (2008) Literature mining in support of drug discovery. *Brief. Bioinform.*, **9**, 479–492.

Alexopoulou,D. *et al*. (2009) Biomedical word sense disambiguation with ontologies and metadata: automation meets accuracy. *BMC Bioinformatics*, **10**, 28.

Banville,D.L. (2006) Mining chemical structural information from the drug literature. *Drug. Discov. Today*, **11**, 35–42.

Bingjun,S. *et al*. (2007) Extraction and search of chemical formulae in text documents on the web. In *Proceedings of the 16th International Conference on World Wide Web*. ACM, Banff, Alberta, Canada.

Bingjun,S. *et al*. (2008) Mining, indexing, and searching for textual chemical molecule information on the web. In *Proceeding of the 17th International Conference on World Wide Web*. ACM, Beijing, China.

Bodenreider,O. (2004) The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, **32**, D267–D270.

Chen,J.H. *et al*. (2007) ChemDB update–full-text search and virtual chemical space. *Bioinformatics*, **23**, 2348–2351.

Cohen,A.M. and Hersh,W.R. (2005) A survey of current work in biomedical text mining. *Brief. Bioinform.*, **6**, 57–71.

Corbett,P. *et al*. (2007) Annotation of chemical named entities. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*. Prague, pp. 57–64.

Corbett,P. and Copestake,A. (2008) Cascaded classifiers for confidence-based chemical named entity recognition. *BMC Bioinformatics*, **9**(Suppl. 11), S4.

Corbett,P. and Murray-Rust,P. (2006) High-throughput identification of chemistry in life science texts. In Berthold,M.R. *et al*. (eds), *CompLife 2006*. Springer Berlin/Heidelberg, Cambridge, UK, pp. 107–118.

Degtyarenko,K. *et al*. (2008) ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.*, **36**, D344–D350.

Edmonds,P. and Agirre,E. (2006) *Word Sense Disambiguation: Algorithms and Applications*. Springer Verlag.

Erhardt,R.A.A. *et al*. (2006) Status of text-mining techniques applied to biomedical text. *Drug Discov. Today*, **11**, 315–325.

Goto,S. *et al*. (2002) LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res.*, **30**, 402–404.

Hanisch,D. *et al*. (2005) ProMiner: rule-based protein and gene entity recognition. *BMC Bioinformatics*, **6**(Suppl. 1), S14.

Kanehisa,M. *et al*. (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.

Kemp,N. and Michael,L. (1998) Extraction of information from the text of chemical patents. 1. Identification of specific chemical names. *J. Chem. Inf. Comput. Sci.*, **38**, 544–551.

Kim,J.D. *et al*. (2003) GENIA corpus–semantically annotated corpus for bio-textmining. *Bioinformatics*, **19** (Suppl. 1), i180–i182.

Klinger,R. *et al*. (2008) Detection of IUPAC and IUPAC-like chemical names. *Bioinformatics*, **24**, i268–i276.

Kolarik,C. *et al*. (2007) Identification of new drug classification terms in textual resources. *Bioinformatics*, **23**, i264–i272.

Kolarik,C. *et al*. (2008) Chemical names: terminological resources and corpora annotation. In *Proceedings of the Workshop on Building and evaluating resources for biomedical text mining (6th edition of the Language Resources and Evaluation Conference)*. Marrakech.

Lipscomb,C.E. (2000) Medical subject headings (MeSH). *Bull. Med. Libr. Assoc.*, **88**, 265–266.

McCray,A.T. *et al*. (1994) Lexical methods for managing variation in biomedical terminologies. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, Washington, DC, pp. 235–239.

McCray,A.T. *et al*. (2001a) Evaluating UMLS strings for natural language processing. In *Proceedings of the AMIA Symposium*. Washington, DC, pp. 448–452.

McCray,A.T. *et al*. (2001b) Aggregating UMLS semantic types for reducing conceptual complexity. *Stud. Health Technol. Inform.*, **84**, 216–220.

Morgan,A.A. *et al*. (2008) Overview of BioCreative II gene normalization. *Genome Biol.*, **9** (Suppl. 2), S3.

Murray-Rust,P. (2008) Chemistry for everyone. *Nature*, **451**, 648–651.

Murray-Rust,P. *et al*. (2005) Chemistry in bioinformatics. *BMC Bioinformatics*, **6**, 141.

Rebholz-Schuhmann,D. *et al*. (2008) Text processing through web services: calling Whatizit. *Bioinformatics*, **24**, 296–298.

Richard,A.M. *et al*. (2006) Chemical structure indexing of toxicity data on the internet: moving toward a flat world. *Curr. Opin. Drug Discov. Devel.*, **9**, 314–325.

Rogers,W.J. and Aronson,A.R. (2008) Filtering the UMLS Metathesaurus for MetaMap. Technical Report. Available at http://skr.nlm.nih.gov/papers/references/filtering07.pdf

Schuemie,M.J. *et al*. (2007a) Peregrine: lightweight gene name normalization by dictionary lookup. In *Proceedings of the Biocreative 2 workshop*. Madrid.

Schuemie,M.J. *et al*. (2007b) Evaluation of techniques for increasing recall in a dictionary approach to gene and protein name identification. *J. Biomed. Inform.*, **40**, 316–324.

Schulz,M. *et al*. (2006) SBMLmerge, a system for combining biochemical network models. *Genome Inform.*, **17**, 62–71.

Schwartz,A.S. and Hearst,M.A. (2003) A simple algorithm for identifying abbreviation definitions in biomedical text. *Pac. Symp. Biocomput.*, **8**, 451–462.

Segura-Bedmar,I. *et al*. (2008) Drug name recognition and classification in biomedical texts. A case study outlining approaches underpinning automated systems. *Drug Discov. Today*, **13**, 816–823.

Singh,S.B. *et al*. (2003) Text influenced molecular indexing (TIMI): a literature database mining approach that handles text and chemistry. *J. Chem. Inf. Comput. Sci.*, **43**, 743–752.

Smith,L. *et al*. (2008) Overview of BioCreative II gene mention recognition. *Genome Biol.*, **9** (Suppl. 2), S2.

Torii,M. *et al*. (2007) A comparison study on algorithms of detecting long forms for short forms in biomedical text. *BMC Bioinformatics*, **8** (Suppl. 9), S5.

Walker,M.J. *et al*. (2002) CKB - the compound knowledge base: a text based chemical search system. *J. Chem. Inf. Comput. Sci.*, **42**, 1293–1295.

Weisgerber,D.W. (1997) Chemical abstracts service chemical registry system: history, scope, and impacts. *J. Am. Soc. Inform. Sci.*, **48**, 349–360.

Wheeler,D.L. *et al*. (2008) Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, **36**, D13–D21.

Wilbur,W.J. *et al*. (1999) Analysis of biomedical text for chemical names: a comparison of three methods. *Proc. AMIA Symp*., 176–180.

Williams,A.J. (2008a) Internet-based tools for communication and collaboration in chemistry. *Drug Discov. Today*, **13**, 502–506.

Williams,A.J. (2008b) A perspective of publicly accessible/open-access chemistry databases. *Drug Discov. Today*, **13**, 495–501.

Wishart,D.S. *et al*. (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.*, **36**, D901–D906.

Wishart,D.S. *et al*. (2009) HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Res.*, **37**, D603–D610.

Wren,J. (2006) A scalable machine-learning approach to recognize chemical names within large text databases. *BMC Bioinformatics*, **7** (Suppl. 2), S3.

Xu,Y. *et al*. (2009) MBA: a literature mining system for extracting biomedical abbreviations. *BMC Bioinformatics*, **10**, 14.

Yu,H. *et al*. (2002) Mapping abbreviations to full forms in biomedical articles. *J. Am. Med. Inform. Assoc.*, **9**, 262–272.

Yu,H. *et al*. (2007) Using MEDLINE as a knowledge source for disambiguating abbreviations and acronyms in full-text biomedical journal articles. *J. Biomed. Inform.*, **40**, 150–159.

Zhu,S. *et al*. (2005) A probabilistic model for mining implicit 'chemical compound-gene' relations from literature. *Bioinformatics*, **21** (Suppl. 2), ii245–ii251.

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

Zweigenbaum,P. *et al*. (2007) Frontiers of biomedical text mining: current progress. *Brief. Bioinform.*, **8**, 358–375.