Text Mining for Metabolic Reaction Extraction from Scientific Literature

Judith E. Risse

Thesis committee

Promotors
Prof. Dr A.H.J. Bisseling
Professor of Molecular Biology
Wageningen University

Prof. Dr J.A.M. Leunissen†
Professor of Bioinformatics
Wageningen University

Co-promotor
Dr P.E. van der Vet
Assistant professor, Human Media Interaction Group
University of Twente, Enschede

Other members
Prof. Dr R.D. Hall, Wageningen University
Prof. Dr U. Hahn, Jena University, Germany
Dr C. Evelo, Maastricht University
Dr P. Moerland, Academic Medical Centre, Amsterdam

This research was conducted under the auspices of the Graduate School of Experimental Plant Sciences

† Deceased

Text Mining for Metabolic Reaction Extraction from Scientific Literature

Judith E. Risse

Thesis
submitted in fulfillment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. Dr M.J. Kropff,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Monday 7 April 2014
at 1:30 p.m. in the Aula.

Judith E. Risse

Text Mining for Metabolic Reaction Extraction from Scientific Literature
138 pages.

PhD thesis Wageningen University, Wageningen, NL (2014)
With references, with summaries in English and Dutch

ISBN: 987-90-6173-900-1

"They think written words are even more powerful", whispered the toad.
"They think all writing is magic. Words worry them."

— *Terry Pratchett, The Wee Free Men*

Contents

Chapter 1

General Introduction

Data: structured and unstructured

Science relies on data, whether it is to generate hypotheses from observations or to verify predictions through experiments. In molecular biology and bioinformatics large scale data generation has taken centre stage in the last 20 years (Kanehisa&Bork 2003; 2007). Data in these fields are found in two major forms, in databases and in scientific literature. Despite calls to treat both kinds as equal (Bourne 2005) there are significant differences in both structure and perception.
Most scientists are familiar with classical structured data, usually in tabular form, accessible in online databases like GenBank (Benson et al. 2013), PDB (Berman et al. 2000) or UniProt (UniProt 2013). Those data can be accessed with tools like BLAST (Altschul et al. 1990) or web services and classical data mining techniques. Parallel to the exponential increase of biological data, mainly due to large scale genomic experiments, has been the almost exponential growth of scientific literature. Scientific literature is mostly used as information source, accessed with search engines like PubMed (NIH) and Scopus (Elsevier). The used queries are mainly based on structured fields like keywords, assigned or found in title or abstract, and on authors, journal title or publication date, but not so much the content itself.
Conventional data mining techniques do not work for the natural language part of a publication beyond simple retrieval queries based on pattern matching. Attempts have been made to make the content of publications accessible for data mining in the form of structured digital abstracts (Gerstein et al. 2007; Seringhaus&Gerstein 2008). Even if these were feasible for new publications (Hahn et al. 2007) there remains the backlog of more than 22 million past publications in PubMed alone (NIH).

Text Mining

To make the most of the articles' content requires a dedicated set of techniques and approaches tailored to the unstructured nature of free text. Analogous to the field of data mining for the analysis of structured data, the field of text mining has emerged for unstructured text. Scientific literature has been viewed as source of new knowledge in biomedicine since the 1960s (Baruch 1965) and in molecular biology since 1990 (Swanson 1990). Since then the research into text mining applications for that purpose has taken flight (Zhu et al. 2012).
Text mining is the analysis of free text to gain new insights and knowledge (Hearst 1999). It goes beyond the simple retrieval of facts and requires the combination of information from different textual sources to capture new relations, insights or hypotheses. More generally the term is used for everything pertaining to extraction of information from text and its preparatory steps (Zweigenbaum et al. 2007). It is the first definition we would like to achieve, but the second is the one we will use in the scope of this thesis. Regardless which definition is followed, text mining is a modular task (De Bruijn&Martin 2002). A first step to text analysis is the selection

of target documents. This selection determines to a great extent the further analysis and ranges from selecting a suitable corpus based on content to selecting the type of documents to be studied. For our study we initially consider all biological text relevant and we use Medline (NIH), a downloadable version of PubMed, as our main input as well as publications in the open source collection BioMed Central (Springer).

A crucial step for text mining is named entity recognition (NER), in which entities of interest are identified within text. Biology revolves around entities, their relations, the conditions under which these relations occur and the techniques that measure both relations and conditions. All these are referred to with named entities and these named entities are in natural language described in constituents, making their identification a first text mining step in understanding free text. Challenges of NER are the identification of the correct entity boundaries, especially considering the complexity of some biological multi-word terms (e.g. 'cytoplasmic nad-dependent glycerol-3-phosphate dehydrogenase') and the required specificity (e.g. 'cytoplasmic nad-dependent glycerol-3-phosphate dehydrogenase' vs. 'glycerol-3-phosphate dehydrogenase'). Any system should also have a strategy to handle homonyms, where one constituent has different meanings depending on its context (e.g. 'sequoia' can refer to a fruit fly gene or a North American tree). Further ambiguity is caused by ambiguous acronyms and abbreviations (e.g. 'pdh' can stand for pyruvate dehydrogenase, prephenate dehydrogenase or proline dehydrogenase, *see Chapter 3*). Separate tools have been developed to specifically resolve this problem (Schwartz&Hearst 2003). Also challenging is the fact that single entities are often referred to by multiple names, so-called synonyms (e.g. alpha-ketoglutarate dehydrogenase and oxoglutarate decarboxylase refer to the same enzyme). All NER approaches have to overcome the dynamic nature of language in so far as they have to be able to cope with the continuous invention of new terms by authors and shifts in meaning over time for existing terms.

Several different approaches have been taken to label named entities in text. Some like AbGene (Tanabe&Wilbur 2002) collect a set of rules to identify, in this case, gene names in text. GAPSCORE (Chang et al. 2004) uses a statistical model for gene names, quantifying appearance, morphology and context. Others like ABNER (Settles 2005) and BANNER (Leaman&Gonzalez 2008) use machine-learning techniques to train the application with manually annotated text to recognise gene names and other biological entities. A last method, employed by us (Chapter 3), is dictionary- or lexicon-based NER. Here a list of names of a given class of entities, in our case enzymes and metabolites, is collected to be identified in text with string matching techniques (Hanisch et al. 2005). The dictionary-based approach is set apart from the other methods in that it facilitates named entity identification (NEI). NEI is the identification of a specific entity like one gene or enzyme and its linkage to external database identifiers like GenBank ID (Benson et al. 2013) or enzyme number (Enzyme Nomenclature). This sets it apart from NER which only allows identification of the class of the entity as gene or enzyme (Hettne et al. 2009).

While systems with NER and particularly NEI are of great value for annotators and curators of scientific databases (Rebholz-Schuhmann et al. 2005), biologists require information beyond labelled entities from the text to gain a better understanding of the relationships between entities for a more complete picture. This is a task commonly referred to as relation extraction (RE). The challenges here lie in defining the nature of the relationship to extract and in linguistic nuances writers apply in their texts. Relationships can be described outright (e.g. 'protein A binds to protein B'), they can be of a more hypothetical nature (e.g. 'protein A probably binds to protein B') or downright negated (e.g. 'protein A does not bind to protein B'). Again there are different approaches to identifying relations. The most basic form is co-occurrence of two or more entities in the same piece of text, be it document, section, paragraph or sentence. The unit size of text determines to a large extent the accuracy of such an approach, with smaller units like sentences giving more precision. But generally co-occurrence is not very accurate as it cannot distinguish between the types of relations described above or random co-occurrences. Similar to the NER task, rule- or pattern-based solutions define rules or regular expressions to extract relations (Huang et al. 2004; Lee et al. 2012). Machine learning and statistical applications are also applied to relation extraction. A large number of the RE approaches focussed on protein-protein interactions have been tested in the BioCreative challenge (Krallinger et al. 2008). The BioCreative Challenge (2004 – now, with the BioCreative IV Challenge run in October 2013) (Hirschman et al. 2005), together with the genomics track by TREC (2003-2007) (Hersh&Voorhees 2009) provide a platform to boost and evaluate research for text mining in molecular biology with topics ranging from gene name identification to protein-protein interaction extraction and assisted curation.

A more sophisticated approach to relation extraction involves natural language processing (NLP) techniques. Here the goal is to computationally understand syntax and semantics of the text to extract relations (Jusoh&Alfawareh 2012). It usually involves pre-processing steps to split text into sentences, part of speech (POS) tagging of the constituents and then full or partial syntactic analysis of each sentence resulting in a parse tree (Allen 1994). The depth of syntactic analysis and semantic interpretation systems invoke varies. Chilibot (Chen&Sharp 2004) uses partial parsing with CASS (Abney 1996) to determine the interaction type. Pyysalo *et al.* (Pyysalo et al. 2004) use Link grammar (Sleator&Temperley 1995) to identify interaction subtrees in the syntactic parse tree. RelEx (Fundel et al. 2007) uses dependency parse trees to extract gene and protein relations. The AGFL grammar work lab (Koster&Verbruggen 2002) with the EP4IR grammar used by us in Chapter 4 is a dependency parser as well. The parsing strategy of dependency grammars focusses on the verb as head of a syntactic parse tree in which all other constituents in the sentence are linked by one-to-one relations labelled with their syntactic roles. Dependency grammars are particularly suited to the complex nature of biomedical text as they are able to identify relations between terms in sentences spanning large distances (Fundel et al. 2007).

Metabolomics

Not surprisingly, in the wake of large scale genomics experiments, most text mining applications in molecular biology have focussed on genes and proteins and their relations as well. However these experiments and other ~omics research have opened the field for more integrated approaches to solving biological questions (Kell 2004). No longer are we required to study single entities and their roles, we can now investigate complete systems (Hollywood et al. 2006). Central to the understanding of the system of an organism and its internal processes is the metabolome, the small compounds generated and converted by enzymatic processes (Oliver et al. 1998). Metabolomics, the study of those metabolites, provides an important link between genotype and phenotype (Fiehn 2002).

Text mining for metabolomics

Information about metabolites, the core entities in metabolomics, is to a large extent not (yet) stored in databases but in scientific literature (Nobata et al. 2011). Ma et al. (Ma et al. 2007) have shown that the available metabolic databases are far from complete and additional knowledge is still to be recovered from literature. Examples of such metabolic pathway databases are KEGG (Kanehisa et al. 2006), Reactome (Croft et al. 2011), HMDB (Wishart et al. 2009) and BioCyc (Caspi et al. 2012). Stobbe et al. (Stobbe et al. 2011) and Soh et al. (Soh et al. 2010) show that overlap between different pathway databases is limited, indicating that none of them covers all existent knowledge about metabolites. Despite this, most text mining approaches are focussing on other entities. Manual efforts have been made to transfer information about metabolites found in literature to databases. Recent large-scale and mostly manual efforts to establish a complete view of the human metabolome are RECON2 (Thiele et al. 2013) and the Edinburgh human metabolic network (Ma et al. 2007). In the WikiPathways project researchers are trying to harness the knowledge of the community to generate a comprehensive pathway database (Pico et al. 2008).

The lack of attention to the field of metabolomics by the text mining community and the clear need for text mining assistance to the field offers opportunities. Few have taken up this challenge. Hettne et al. (Hettne et al. 2009) published Jochem, a dictionary for small molecules to facilitate text mining, Nobata et al. (Nobata et al. 2011) attempt to extract the yeast metabolome from literature, Czarnecki et al. (Czarnecki et al. 2012) developed a hybrid rule-based method to extract metabolic reactions from literature, and EmPathIE (Humphreys et al. 2000) had the goal to extract metabolic reactions together with contextual information. Knox et al. (Knox et al. 2007) developed BioSpider, a web-based application for metabolome annotation incorporating text mining in their analysis.

None of the examples above use a full grammar approach for relation extraction like our attempt. Close in technique but not target relations is GENIES (Friedman et al.

2001) which extracts signal transduction pathways from literature. All in all we are not aware of a full-grammar approach to extract metabolic pathways from scientific literature as described by us in Chapter 4.

Scope of this thesis

The scope of this thesis is to investigate a text mining application capable of extracting metabolic reactions from scientific literature. With this we have taken all necessary preparatory steps to combine those reactions into metabolic pathways. This allows us to gather evidence for known pathways and potentially uncover missing links between metabolites not yet stored in a scientific database. To this length we describe the preliminary steps to reach this goal in the next two chapters and the prototype product in the fourth chapter.

In Chapter 2 we present an overview of performance and capabilities of classic data mining solutions for handling large amounts of XML-type data and full-text searches. Chapter 3 describes the creation of two thesauri, one for enzymes and one for metabolites, from existing data sources by a combination of computational and manual steps. The thesauri were designed specifically for the purpose of high quality named entity recognition and subsequent named entity identification for metabolic reaction extraction. They fill an important gap in the available resources required for text mining of metabolic reactions with high precision.

In Chapter 4 we describe the design of our core text mining application capable of identifying and extracting metabolic reactions from scientific literature, both abstracts and full-text, and compare its performance to a different state-of-the-art approach (Czarnecki et al. 2012). With the implementation of our text mining approach we provide a high precision attempt to metabolic reaction extraction making use of the full information content of scientific literature. In the final chapter we discuss and summarise the findings of this thesis, the broader implications of this research and its integration into a broader scope. We also briefly give our view on the future role of text mining in the biomedical world.

References

Abney S. 1996. Partial parsing via finite-state cascades. *Natural Language Engineering* **2**(4): 337-344.

Allen J. 1994. *Natural language understanding*. Addison Wesley.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of molecular biology* **215**(3): 403-410.

Baruch J. 1965. Progress in programming for processing English language medical records. *Annals of the New York Academy of Sciences* **126**(2): 795-804.

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ et al. 2013. GenBank. *Nucleic Acids Res* **41**(Database issue): D36-42.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN et al. 2000. The Protein Data Bank. *Nucleic Acids Research* **28**(1): 235-242.

Bourne P. 2005. Will a biological database be different from a biological journal? *PLoS Computational Biology* **1**(3): e34.

Caspi R, Altman T, Dreher K, Fulcher CA, Subhraveti P et al. 2012. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research* **40**(D1): D742-D753.

Chang JT, Schütze H, Altman RB. 2004. GAPSCORE: finding gene and protein names one word at a time. *Bioinformatics* **20**(2): 216-225.

Chen H, Sharp BM. 2004. Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics* **5**.

Croft D, O'Kelly G, Wu G, Haw R, Gillespie M et al. 2011. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Research* **39**(suppl 1): D691-D697.

Czarnecki J, Nobeli I, Smith A, Shepherd A. 2012. A text-mining system for extracting metabolic reactions from full-text articles. *BMC Bioinformatics* **13**(1): 172.

De Bruijn B, Martin J. 2002. Getting to the (c)ore of knowledge: Mining biomedical literature. *International Journal of Medical Informatics* **67**(1-3): 7-18.

Editorial. 2007. The database revolution. *Nature* **445**(7125): 229-230.

Scopus [http://www.elsevier.com/online-tools/scopus]

Enzyme Nomenclature [http://www.chem.qmul.ac.uk/iubmb/enzyme/]

Fiehn O. 2002. Metabolomics - The link between genotypes and phenotypes. *Plant Molecular Biology* **48**(1-2): 155-171.

Friedman C, Kra P, Yu H, Krauthammer M, Rzhetsky A. 2001. GENIES: A natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics* **17**(SUPPL. 1).

Fundel K, Küffner R, Zimmer R. 2007. RelEx—Relation extraction using dependency parse trees. *Bioinformatics* **23**(3): 365-371.

Gerstein M, Seringhaus M, Fields S. 2007. Structured digital abstract makes text mining easy. *Nature* **447**(7141): 142.

Hahn U, Wermter J, Blasczyk R, Horn PA. 2007. Text mining: powering the database

revolution. *Nature* **448**(7150): 130-130.

Hanisch D, Fundel K, Mevissen H-T, Zimmer R, Fluck J. 2005. ProMiner: Organism-specific protein name detection using approximate string matching. *BMC Bioinformatics* **6**(Suppl 1): S14.

Hearst MA. 1999. Untangling text data mining. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pp. 3-10. Association for Computational Linguistics.

Hersh W, Voorhees E. 2009. TREC genomics special issue overview. *Information Retrieval* **12**(1): 1-15.

Hettne KM, Stierum RH, Schuemie MJ, Hendriksen PJM, Schijvenaars BJA et al. 2009. A dictionary to identify small molecules and drugs in free text. *Bioinformatics* **25**(22): 2983-2991.

Hirschman L, Yeh A, Blaschke C, Valencia A. 2005. Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics* **6**(Suppl 1): S1.

Hollywood K, Brison DR, Goodacre R. 2006. Metabolomics: Current technologies and future trends. *Proteomics* **6**(17): 4716-4723.

Huang M, Zhu X, Hao Y, Payan DG, Qu K et al. 2004. Discovering patterns to extract protein–protein interactions from full texts. *Bioinformatics* **20**(18): 3604-3612.

Humphreys K, Demetriou G, Gaizauskas R. 2000. Bioinformatics applications of information extraction from scientific journal articles. *Journal of Information Science* **26**(2): 75-85.

Jusoh S, Alfawareh HM. 2012. Techniques, Applications and Challenging Issue in Text Mining. *International Journal of Computer Science Issues(IJCSI)* **9**(6).

Kanehisa M, Bork P. 2003. Bioinformatics in the post-sequence era. *nature genetics* **33**: 305-310.

Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M et al. 2006. From genomics to chemical genomics: new developments in KEGG. *Nucl Acids Res* **34**(suppl_1): D354-357.

Kell DB. 2004. Metabolomics and systems biology: making sense of the soup. *Current Opinion in Microbiology* **7**(3): 296-307.

Knox C, Shrivastava S, Stothard P, Eisner R, Wishart DS. 2007. Biospider: A web server for automating metabolome annotations. pp. 145-156.

Koster CHA, Verbruggen E. 2002. The AGFL Grammar Work Lab. In *Proceedings of the FREENIX Track: 2002 USENIX Annual Technical Conference*. USENIX Association.

Krallinger M, Leitner F, Rodriguez-Penagos C, Valencia A. 2008. Overview of the protein-protein interaction annotation extraction task of BioCreative II. *Genome Biology* **9**(Suppl 2): S4.

Leaman R, Gonzalez G. 2008. BANNER: an executable survey of advances in biomedical named entity recognition. In *Pacific Symposium on Biocomputing*, Vol 13, pp. 652-663.

Lee J, Kim S, Lee S, Lee K, Kang J. 2012. High precision rule based PPI extraction and per-pair basis performance evaluation. In *Proceedings of the ACM sixth international workshop on Data and text mining in biomedical informatics*, pp. 69-76. ACM, Maui, Hawaii, USA.

Ma H, Sorokin A, Mazein A, Selkov A, Selkov E et al. 2007. The Edinburgh human metabolic network reconstruction and its functional analysis. *Mol Syst Biol* **3**.

Medline [http://www.nlm.nih.gov/pubs/factsheets/medline.html]

PubMed [http://www.ncbi.nlm.nih.gov/pubmed/]

Nobata C, Dobson P, Iqbal S, Mendes P, Tsujii Ji et al. 2011. Mining metabolites: extracting the yeast metabolome from the literature. *Metabolomics* **7**(1): 94-101-101.

Oliver SG, Winson MK, Kell DB, Baganz F. 1998. Systematic functional analysis of the yeast genome. *Trends in Biotechnology* **16**(9): 373-378.

Pico AR, Kelder T, van Iersel MP, Hanspers K, Conklin BR et al. 2008. WikiPathways: pathway editing for the people. *PLoS biology* **6**(7): e184.

Pyysalo S, Ginter F, Pahikkala T, Boberg J, J\ J et al. 2004. Analysis of link grammar on biomedical dependency corpus targeted at protein-protein interactions. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pp. 15-21. Association for Computational Linguistics, Geneva, Switzerland.

Rebholz-Schuhmann D, Kirsch H, Couto F. 2005. Facts from Text—Is Text Mining Ready to Deliver? *PLoS Biol* **3**(2): e65.

Schwartz AS, Hearst MA. 2003. A simple algorithm for identifying abbreviation definitions in biomedical text. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing*: 451-462.

Seringhaus M, Gerstein M. 2008. Manually structured digital abstracts: A scaffold for automatic text mining. *FEBS letters* **582**(8): 1170.

Settles B. 2005. ABNER: An open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics* **21**(14): 3191-3192.

Sleator DD, Temperley D. 1995. Parsing English with a link grammar. *Technical Report CMU-CS-91-196, Department of Computer Science, Carnegie Mellon University, Pittsburgh, PA*.

Soh D, Dong D, Guo Y, Wong L. 2010. Consistency, comprehensiveness, and compatibility of pathway databases. *BMC Bioinformatics* **11**(1): 449.

BioMed Central [http://www.biomedcentral.com/about/datamining]

Stobbe M, Houten S, Jansen G, van Kampen A, Moerland P. 2011. Critical assessment of human metabolic pathway databases: a stepping stone for future integration. *BMC Systems Biology* **5**(1): 165.

Swanson DR. 1990. Medical literature as a potential source of new knowledge. *Bulletin of the Medical Library Association* **78**(1): 29.

Tanabe L, Wilbur WJ. 2002. Tagging gene and protein names in biomedical text. *Bioinformatics* **18**(8): 1124-1132.

Thiele I, Swainston N, Fleming RMT, Hoppe A, Sahoo S et al. 2013. A community-driven global reconstruction of human metabolism. *Nat Biotech* **31**(5): 419-425.

UniProt. 2013. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Research* **41**(D1): D43-D47.

Wishart DS, Knox C, Guo AC, Eisner R, Young N et al. 2009. HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Research* **37**(suppl 1): D603-D610.

Zhu F, Patumcharoenpol P, Zhang C, Yang Y, Chan J et al. 2012. Biomedical text mining and its applications in cancer research. *Journal of Biomedical Informatics*.

Zweigenbaum P, Demner-Fushman D, Yu H, Cohen KB. 2007. Frontiers of biomedical text mining: current progress. *Briefings in Bioinformatics* **8**(5): 358-375.

Chapter 2

A Comparison of Database Systems for XML type Data

Judith E Risse, Jack AM Leunissen†

Laboratory for Bioinformatics, Plant Sciences Group, Wageningen University, Wageningen, Netherlands

†Author deceased

Published in *In Silico Biology*, 2010:**10**(3,4):pp193-205

Abstract

Background

In the field of bioinformatics interchangeable data formats based on XML are widely used. XML type data is also at the core of most web services. With the increasing amount of data stored in XML comes the need for storing and accessing the data. In this paper we analyse the suitability of different database systems for storing and querying large datasets in general and Medline in particular.

Results

All reviewed database systems perform well when tested with small to medium sized datasets, however when the full Medline dataset is queried the times start to vary greatly.

Conclusions

There no one system that is vastly superior to the others in this comparison and depending on the database size and the query requirements different systems are most suitable. The best all-round solution is the Oracle 11g database system using the new binary storage option. Alias-i's Lingpipe is a more lightweight, customizable and sufficiently fast solution. It does however require more initial configuration steps. For data with a changing XML structure Sedna and BaseX as native XML database systems or MySQL with an XML-type column are suitable.

Background

In the last few years XML (eXtensible Markup Language (W3C 2006) has become the data format of choice in the field of bioinformatics (Strömbäck et al. 2007). Also increasing numbers of data standards are defined in XML schemas, for example MIBBI (Taylor et al. 2008) or Medline (NIH), the dataset used in this paper. With the advent of web services (Neerincx&Leunissen 2005) providing high throughput computational access to remote datasets XML has become ubiquitous.
The advantage of a structured, extensible, flat-text format for data exchange might be immediately obvious, but there are also some challenges in working with large XML files mainly caused by the sequential nature of the format. The problems centre on how to pass through the files to quickly access different levels within the XML-tree without consuming vast amounts of hardware resources.
Suitable database systems (dbs's) greatly assist in accessing large XML datasets and making them available to high throughput data mining applications. A common perception is that XML type dbs's are still lagging behind in performance compared to object-relational dbs's from major database vendors but tests are not readily available. Selecting a dbs that fits the requirements regarding maximum database size, power of the query language and, importantly, performance is not trivial. With this paper we review our own selection process and criteria and hope to aid others in choosing the appropriate system.
The data relevant to our research is the Medline database (NIH), which makes up a large part of PubMed, the NLM literature database. More importantly it is available for lease as a local copy allowing its application in a high throughput setting. However, the results of this study are applicable to most types of XML datasets.

Previous Research

In our selection process we looked at a variety of approaches to process and handle XML data.
In the classic database implementations data is stored in tables with columns which are then linked to each other via shared columns. This way of storing data is not well-suited for all XML structures, but is applicable to the Medline format used in this paper and to most other XML formats commonly used in the field of bioinformatics. The advantage of relational dbs's is that they have been developed and optimized for a long time by major database vendors, which should give them an advantage in the comparison.
Database systems specifically designed for XML have only been around since 1999 and most of the systems still in use appeared in 2000 (e.g. eXist (Meier 2009), BerkelyDB-XML (Berkeley DB)) or later (Sedna (Fomichev et al. 2006), MonetDB (Boncz et al. 2006)). These dbs's have been especially designed to handle the specificities of the XML format. In 1999 XPath was defined as a query language for XML based on the tree-like structure by the W3C (W3C 1999) and in 2007 augmented with the

XQuery standard (W3C 2010) that allows SQL-like queries. In January 2010 the W3C has recommended the XQuery Full-text and Update standards for addition to the XQuery standard, further increasing the possibilities for XML-type dbs's. It has long been thought that these dbs's work well with structure-centric XML, where conversion to a table-based format would result in many tables with complex relationships requiring extensive and expensive joins for queries, whereas data centric XML, with a simple data structure, warrants the conversion to table-based formats (Steegmans et al. 2004). The choice is much less clear for intermediate types of XML data. As most schemas in life-sciences are data-centric we expected the table-based databases systems to perform better than true XML-type databases on our data but wanted to see whether the gain in performance warrants the added overhead and loss of flexibility.

Flat-text dbs's have been around since at least 1989 when SRS was developed for the fast retrieval of annotation information and cross-referencing of different flat-file libraries (Etzold&Argos 1993). One of the main features is the indexing of links between different datasets creating a network of databases. Flat-text indexing systems like SRS and MRS (Hekkelman&Vriend 2005) have been particularly designed for flat-text and should suit XML-type data, hence warranting a place in this comparison.

This research was performed in order to evaluate the performance of different database systems when applied to mining data from large XML type datasets. From this high throughput point of view the performance of a native XML database was compared to more classical database approaches. Compared to a relational database model where XML data has to be parsed into different columns and tables before it can be loaded into the database, native XML databases can work directly with the XML data. To broaden the comparison we included two other types of database systems namely text-based and hybrid systems.

In this research we compare an Oracle 11g relational database as representative for the classical database systems to a selection of other database types. As representatives of the true XML type dbs's we chose Sedna and BaseX (BaseX). Some other XML-type systems were considered, but discarded as they could not hold all data (MonetDB, 10Gb data limit) or could not be installed on our system (Timber). MySQL is included as it is the most widely used dbs in the academic world. Furthermore we compare two further implementations of the Oracle 11g database system, namely object-relational, XML type with CLOB (character large object) type storage and XML type with binary storage with Alias-i LingPipe (LingPipe), Sedna, SRS 7.1 and MRS 4.0. The SRS 7.1 version of SRS was chosen as it is still widely used in the bioinformatics community although the current version is 8.3 (biowisdom website). This study is focused on the more data-centric XML-type data usually found in the field of bioinformatics and the results can be transferred to data of a similar complexity. The results of the performance comparison however cannot be unequivocally extended to structure centric data but the other conclusions regarding ease of use and power of query language hold true regardless the structure.

The reviewed systems were tested with the Medline dataset (2008 Baseline).

The database systems will be compared on different criteria namely usability, performance, power of the query language and scalability. Furthermore we will provide a brief introduction to the integration of each database into an application.

Results and Discussion

When evaluating the criteria set for a suitable system the overall impression is positive. The performance of the dbs's with the two smaller databases is sufficiently close together to allow other criteria to influence the choice of system. Performance does however become the most important factor in the full dataset. In this section we will discuss the results of the performance analysis and other criteria important for choosing a dbs for a large-scale high throughput setting.

Installation

With every piece of software the time and effort required for installation and maintenance have to weigh up against the gain in using it. This is especially true for Linux systems.
For the Oracle 11g databases installation is straightforward as long as one of the Oracle certified operating systems is being used and installation should not take longer than two hours.
MySQL is already included in the Linux installation, therefore there are no problems with dependencies. Using a version different from the one included in the Linux distribution is slightly more complicated as MySQL is deeply embedded in the system. Here we use MySQL sandbox (MySQL Sandbox), a perl module wrapped around the MySQL installation to manage the environment.
Lingpipe and BaseX are Java based and therefore run immediately after extracting the downloaded file and being Java based are also platform independent. BaseX can be used in a client-server model configuration which allows the database to be accessed from within any application.
Sedna can be used directly with a pre-built binary or compiled from source, but to include the full-text indexing and search capabilities a licence for dtSearch has to be obtained and installed. Integrating dtSearch into the Sedna installation does require a dtSearch-ready build but is otherwise straightforward.
SRS 7 is easy to install as it has little or no dependencies; installation should not take longer than two hours including the web server, which is strictly speaking not necessary for a high throughput approach accessing the database from within an application using getz, the SRS query parser.
MRS is a specialised text search engine which makes use of many external libraries and, depending on the Linux flavour used, can be very time-consuming to install and configure. Currently it officially only supports Debian.

Requirements

When working with large amounts of data, the disk space required is a factor to consider. Here the databases storing the complete XML obviously perform worse than the databases only storing the content of nodes. The least amount of disk space is required by SRS7 with 65MB for the small dataset and 10.3GB for the full set. In comparison Sedna requires 1.5GB for the small set and more than 300GB for the full set. Lingpipe requires 139MB for the small set and MRS 4.0 326MB. Of the Oracle 11g dbs's the relational database requires 166MB and the XML dbs's 583 and 631MB respectively, the binary storage model taking up more disk space. MySQL database sizes are 43Mb for the small set and 81Gb for the full set. BaseX requires 413Mb for the small set and 84Gb for the full set.

Updating

On the part of keeping the database up-to-date, apart from MRS SRS and BaseX, the indexes of all systems can be updated allowing the addition of new data without the need to re-index the complete dataset. It is also possible to update the content of individual records. In our case it allows for the incorporation of the daily updates provided by Medline. As it takes more than 24 hours to re-build both MRS and SRS indexes, daily updates cannot be used but a weekly update cycle is possible. BaseX allows queries across multiple databases offering the possibility to simply add new data in a new database and building indexes only on the new subset. MySQL is easily updated by adding new records to the table but requires a full rebuild of the full-text index.

Flexibility

Regarding data stored in XML two types are distinguished, i.c. formats defined by a schema or document type definition and those without. In a schema-based format the structure and elements of the data are predefined. If no schema information is present the structure is determined by the XML document itself. This allows for greater flexibility in mixing data from different sources and expanding and combining existing data structures without having to completely re-index the database.
Oracle 11g XML requires a schema to be bound to the XML type column of the table in order to index the data and allow full query possibilities. Similarly lingpipe uses the schema information to parse the Medline files during indexing. Lingpipe already provides the proper definitions for indexing Medline, but in case of other XML formats, all searchable fields have to be defined beforehand.
Oracle 11g rdbms, MRS and SRS also require predefined structures as the data have to be parsed before they can be imported into the database and indexed. SRS 7 does provide the

	Oracle 11g			MySQL	Lingpipe	Sedna	BaseX	MRS	SRS
	RDBMS	XML Type							
		CLOB	Binary						
version	11.1.0.6.0			5.1.45	3.1.2	3.0.145	6.1	4.1	7.1
language	C/C++			C/C++	Java	C/C++	Java	Perl/C++	Icarus/C
storage	Object-relational		binary	MyISAM	Lucene index files	B-tree	baseX/B-tree	cmp	flat-file
Max db size	none			256TB	none	none(3)	2^31 nodes	none(3)	none(3)
updateable index	yes			yes	yes	yes	yes	no	no
Query language	SQL	SQL/XPath		SQL/XPath	Lucene/XQuery	XQuery	XQuery	custom	custom
Load + index 100000 records	~3min	~1min	~5min	~3min	~3min	~3min	~2.5min	~2min	~5min

Table 1 – Characteristics per database system
Shows the characteristics per database system. (1) Customizable with thesauri, stoplists; (2) Requires dtSearch (license required); (3) not encountered; (4) No proximity searches possible

possibility to automatically extract the database structure from a provided doctype file, but this does not work flawless for complex XML-structures.
Sedna and BaseX as true XML databases can load and index data without prior knowledge of the structure using the structure defined in the XML document. MySQL also uses the structure of the XML document without schema to access nodes and the accompanying full-text index covers the whole document including XML tags. If the structure of the data in question is static and well defined, like Medline in this example, the faster database systems, like Oracle 11g and MRS, can be used. If on the other hand data structures continue to change and do not adhere to a schema or doctype, Sedna, BaseX and MySQL are the options from this comparison.

Query Possibilities

All databases under investigation offer full-text search options although the implementations and the number of possibilities provided vary. This is also strongly linked to the type of query language used.
The general query language for Oracle is SQL, augmented with text-search syntax for full-text columns. In the Oracle 11g systems columns can be covered by a full-text index that can be configured to a large extent considering the types of queries used and incorporating e.g. a stop-list or word boundaries. To search within the data in XML-type columns XPath expressions have to be used to extract the values of nodes. This requires knowledge of the XML data structure, but because the data is extracted from the XML before it is used in a query, the query possibilities are not limited to XQuery and the complete array of PL/SQL queries can be used. On the other hand when queries are performed on a column-based database, knowledge of the table structure is required as well.
Querying the MySQL database requires a similar approach as the data within the XML-column has to be accessed through XPath expressions, it is however necessary to combine this with a full-text query to greatly decrease the processing time.
Sedna requires combination with a 3rd party commercial text-indexing system (dtSearch) (dtSearch), of which the trial version for Linux was used in this study. This system allows for a large variety of queries. Furthermore Sedna complies with the XQuery standard (January 2007 specification). Using XQuery (combined with XPath) can be unfamiliar for users of SQL but it is never the less a powerful query language which emulates the classical SQL *SELECT-FROM-WHERE* with the FLWOR statement (W3C 2010), fully taking into account the peculiarities of XML-type data. The one drawback in the setting of this paper is that XQuery in the 2007 version does not allow full-text searches, proximity queries in particular (i.e. wordA NEAR wordB), but this is remedied in Sedna by using full-text indices from dtSearch. BaseX also uses XQuery combined but does comply with the 2010 XQuery recommendations and fully incorporates the XQuery Full-Text recommendations amended with fuzzy queries. Lingpipe is built around Lucene, a text indexing

system from the Apache Foundation. This dbs can be queried with the Lucene query language, which is a key-value type of language (Lucene) offering large flexibility with a wide variety of term modifiers and which does not require any knowledge of the data structure other than identifiers of nodes. When integrated into a program the query objects can be accessed directly and do not require the parser.
SRS 7 does not allow proximity searches, as it was not designed for this kind of use. On the other hand it offers the advantage of searching across a combination of databases with a single query. This is especially of use in any kind of ~omics research, where large collections of databases are provided in SRS. Apart from that the query-language is of a key-value type. SRS can be accessed from the command-line with getz, the SRS query parser or from within the integrated web-interface.
MRS, although a very fast database, has the most limited set of query options allowing only boolean searches with wildcards. It does not have a positional index so does not allow proximity queries. If it is run from the command line all output fields have to be pre-defined and linked to flags. Running it from within Perl offers a larger flexibility and direct access to the query procedures.

	Boolean	Wildcards	Fuzzy	Proximity
Oracle RDBMS	x	x	x	x
Oracle XML	x	x	x	x
Oracle XML binary	x	x	x	x
MySQL	x	x		
Alias-I Lingpipe	x	x	x	x
Sedna (with dt-search)	x	x		x
BaseX	x	x	x	x
MRS	x	x		
SRS	x	x		

Table 2 - Database query possibilities
Shows the absence or presence of different query types for each database system studied.

Performance

We tested the performance of the dbs's with 5 queries of increasing complexity and averaged the resulting times over replicates (see Materials and Methods). As individual query times are influenced by database and system caching as well as the size of the result set, these averages only give an indication for performance. We find, however, that the mean presents an appropriate and informative result. Across all investigated dbs's performance for the small dataset is very good (Figure 1). All queries are returned in less than one second regardless the type of query and other

factors should take the forefront for choosing a system for databases of this size (see Tables 1 and 2). Stepping up to the large dataset the performances are still as expected (see scalability) but differences between the dbs's increase (
Figure 2). Looking at query 3 (Table 3), the retrieval of a single complete record, most dbs's perform similarly well and differences between the best and the worst performer are less than 0.5 seconds for the full set. Notable exceptions are MySQL and BaseX. The query time for query 3 in the large set in MySQL takes more than ten times longer than in any other dbs. This lack of performance appears to be related to the presence of the full-text index on the XML-column as without it the query time is 0.15 seconds for the large dataset. But leaving out this index slows down query 2 for the same dataset by factor three to about 100 seconds and therefore about 30 times slower than the nearest competitor. Depending on which type of query takes priority a choice for or against the full-text index should be made. For BaseX the times for query 3 in the small and large dataset are in the same league as Sedna, but fall behind for the full set where querying across the different databases significantly impacts performance. Query 2 favours the relational database and the key/value approaches with MRS, SRS, Oracle 11g rdbms and lingpipe performing more than six times better than both Oracle 11g XML-type dbs's and Sedna and BaseX. Exception to this rule is MySQL. These differences are more prominent for the full dataset (9 times better for query 2) (Figure 3).
Overall Sedna, BaseX and MySQL are the weakest dbs's regarding performance and in the full set have query times of up to 5200sec (query 2, MySQL), which can hardly be considered useful in a high throughput approach or even for a single query. However performance is competitive for the small and large dataset.
Overall MRS 4.0 is the fastest dbs across all database sizes for all queries but query 5.

Query 1	SELECT pmid, abstract WHERE author.lastname = "Leunissen"
Query 2	COUNT entries WHERE title CONTAINS "DNA"
Query 3	SELECT title, abstract WHERE pmid = "12954768"
Query 4	SELECT pmid, title, abstract WHERE publication.year = "1992" AND abstract CONTAINS "protein"
Query 5	SELECT pmid, abstract WHERE abstract CONTAINS "virus" NEAR "protein"

Table 3 - Queries in pseudo-code

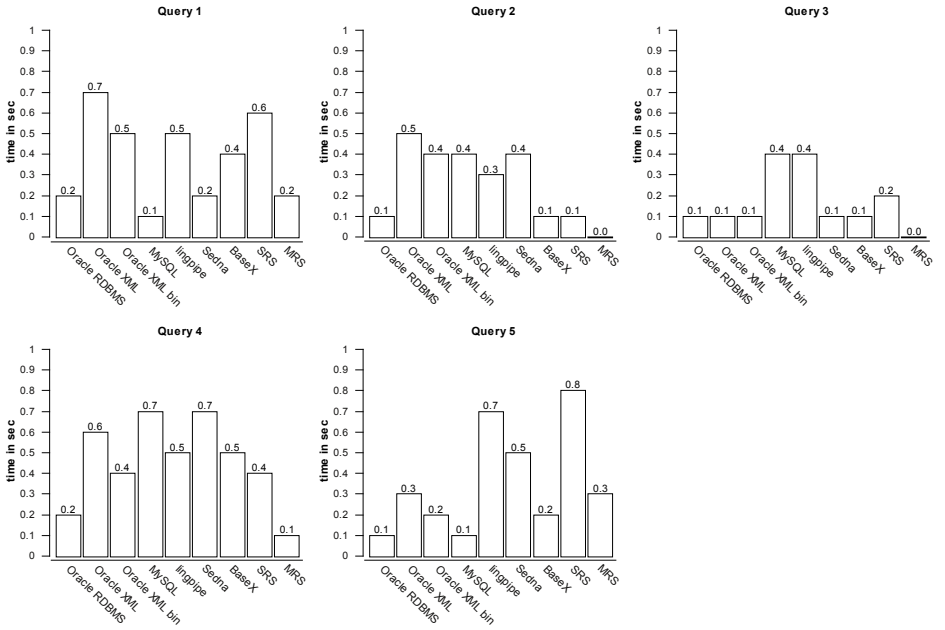

Figure 1 - Query times for the small dataset
Query times per database system per query in seconds. The given times are an average of 25 repeats per query with different search terms. The dataset consisted of 77169 records.

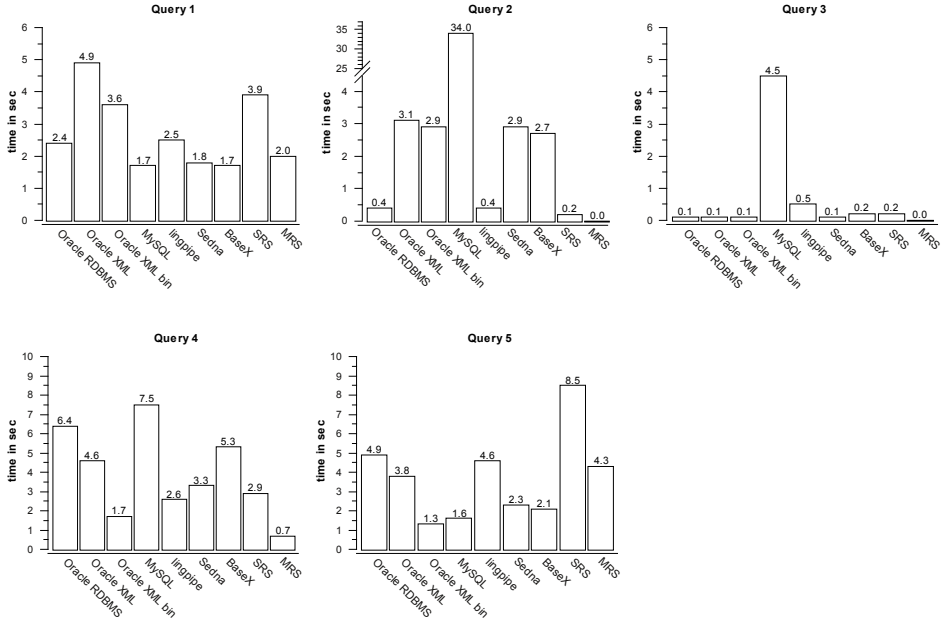

Figure 2 - Query times for the medium sized dataset
Query times per database system per query in seconds. The given times are an average of 25 repeats per query with different search terms. The dataset consisted of 990000 records.

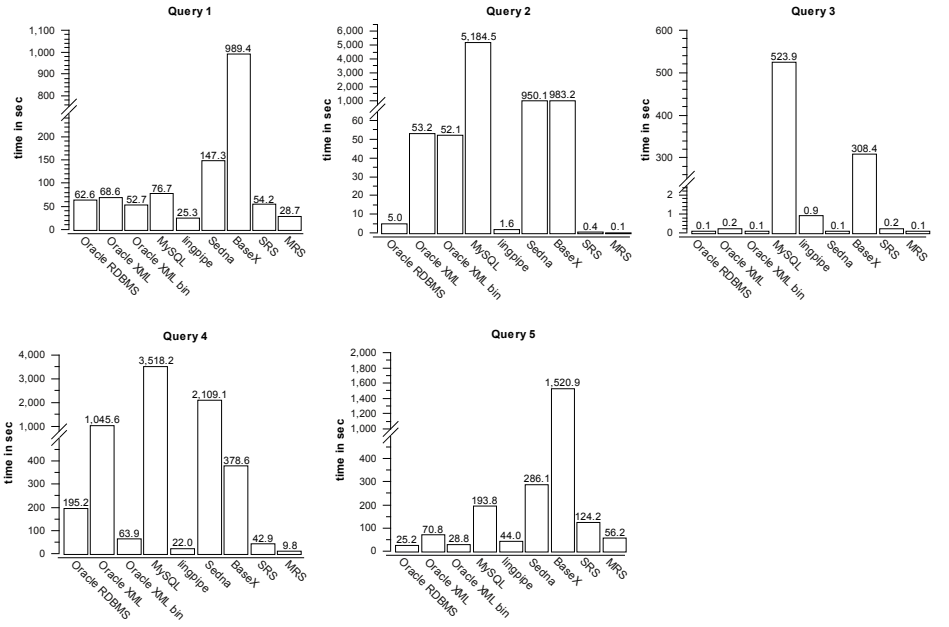

Figure 3 – Query times for the full Medline dataset
Query times per database system per query in seconds. The given times are an average of 25 repeats per query with different search terms. The dataset consisted of the complete Medline dataset containing 16.8 million records.

Scalability

Scalability is an important issue for users with large datasets, or datasets that have the potential to grow. The choice of a database system has to be made with the final database size in mind as a database system performing well with smaller datasets might not do so with larger ones.
All dbs's apart from BaseX in the comparison scale close to linear with increasing database size. Scaling from the large to the full database size is in most cases slightly better than from small to large indicating the overhead of simply executing a query. This is the most obvious in Alias-I's lingpipe implementation where the average slope decreases from 1.39 to 0.83. We assume this is due to the overhead of starting the Java Virtual Machine for each query. Query 3 with its fixed result set scales best across all investigated dbs's showing that the amount of output generated has a large influence on performance. It is also an indication for the quality of indexes as query times hardly increase with increasing datasets. On this level most dbs's perform similar with close to no increase in query time, with the exception of lingpipe (double) MySQL and BaseX (1000 fold) (see above).

Integration

Although even in a high throughput setting command line calls like the ones used in this study play a role, integration of database access into an application should be the aim. The Oracle databases can be integrated into applications through drivers for all major programming languages or the application can be developed around the database using Oracle Application Developer. MySQL also provides access from within major programming languages. BaseX and Sedna provide a.o. drivers, bindings or APIs for C, C#, Java, Python and Ruby.
Lingpipe is developed in Java and consequently integration into Java applications is seamless. Lucene, the underlying indexer is also available for a.o. C# and .NET, C, Perl, Ruby and Python. SRS 7 can only be accessed through wgetz and the web interface or getz the command line interface, integration into an application is therefore only possible through system calls or as a REST-type web service. MRS4.0 provides a Perl interface and integration into a Perl application is thus straightforward, integrating it into applications developed in another programming language could proof to be more laborious. Sedna provides a.o. drivers or APIs for C, C#, Java, Python and Ruby.

Conclusions

The performances and possibilities of the different database types vary greatly, also depending on the type of query. There is no single system that outperforms the others; instead different circumstances can lead to a different optimal solution.
The best all-round solutions are the Oracle 11g dbs's, with the new binary storage option providing the best adaptation to XML combined with good speed. It also provides the most customisable indexing options and query possibilities. On the downside, Oracle is a rather large application and not available under any kind of open source license. If only speed is required and not the exhaustive features of Oracle, MRS 4.0 is a suitable option, although support is limited.
The strong point of SRS 7, apart from its good speed is the fact that it allows integration with numerous already existing databases in the field of molecular biology.
If complete integration in software, a good performance and customisability are required, Lucene is a very good option, as it can be completely integrated into an application and provides implementations for the major programming languages.
Sedna and BaseX, the XML databases in contention, and MySQL (with XML-type columns) are not suitable for very large datasets (> 1 million records) but do offer true support for XML and flexibility for schema-less XML data, especially BaseX with full support for full-text queries based on W3C standards.

Methods

Data

The XML data used is the 2008 baseline of Medline. To evaluate scalability and performance we used three sets of increasing amounts of data. A small set comprised 77169 records, resulting from a random selection of 100000 records from Medline which were subsequently filtered for OLD Medline entries (lacking an abstract) and allowing only English as language. The large set consists of 990000 records, retrieved from 33 randomly selected Medline release files. Lastly we tested the complete 2008 baseline containing about 16.8 million records.

Hardware

All database systems were tested on one server with an Intel Xeon Quadcore CPU (E5335 Clovertown, 2Ghz), 8 Gb memory and a 320Gb Sata disk array in RAID 0 (3Ware RAID controller).

Software

The review focuses on seven different database implementations which all adhere at least to the criterion of being able to contain the complete Medline dataset. Other criteria under evaluation are update capability, full-text indexing capability, ease-of-use and real time querying.
As reference dbs we used an Oracle 11g (version 11.1.0.6.0, 64bit) relational database consisting of a citation and an author table. This we compared to two different Oracle 11g XML database tables containing one primary key column with the PubMed id and one XML type column. In the first implementation we used the older CLOB (Character Large OBject) storage option for the XML and for the second implementation the new binary storage option. Furthermore we evaluated MySQL (5.1.45, with MySQL Sandbox) using XML, two native XML dbs's, Sedna (3.0.145dt), in combination with the dt full-text indexing system and BaseX (6.1), lingpipe (3.1.2) from Alias-i, which is built in Java around the Apache Lucene text-indexing system (2.1.0), and the text-based indexing systems MRS 4.0 and SRS 7.1.

Preparation

All dbs's where installed according to their respective installation instructions and where necessary additional libraries were added.
All databases database systems require either the table or the database to be created and the data to be loaded. The MySQL database was run from within MySQL Sandbox, a Perl module facilitating parallel installations of MySQL.
Once the data is loaded or in some cases already during loading the data is indexed (Table 4). In addition there are database system specific actions that need to be

performed prior to being able to load the data. For the Oracle 11g rdbms this means parsing the data into the required fields in a tab-delimited format. For the Oracle XML type databases the schema had to be bound to the XML type column. MRS required writing of a parser specific for the Medline XML format, the same holds for SRS although there is a tool to convert XML doctype information to Icarus (the SRS parser language) files the indexer can read. Sedna and Lingpipe did not require any specific preparation. However if any other XML data source is used not with Lingpipe but Lucene all fields have to be defined within the code prior to indexing.

	PubMed id	Abstract	Title	Author last name	Publication year	Complete citation
Oracle 11g RDBMS	pk	ft	ft	id	id	
Oracle 11g XML	pk					ft
Oracle 11g XML binary	pk					ft
MySQL	pk					ft
Alias-i Lingpipe	ft	ft	ft	ft	ft	
Sedna	id	ft	ft	id	id	
BaseX						ft/path/att
MRS	pk	ft	ft	ft	id	
SRS	pk	ft	ft	ft	date	

Table 4 - Indexes created per database system
Fields indexed and type of index per database system. In the Alias-i Lingpipe database system all fields in CitationSet are indexed pk = primary key, id = index, ft = full-text index, date = date index, path = path index, att = attribute index

Queries

In order to evaluate the performance of the database systems a representative set of five queries, representing simple retrieval and simple text-mining tasks, was devised (Table 3). The first and third queries represent simple retrieval on different levels of the XML tree. The second query requires a text search within a node. The fourth query requires searching within two different types of nodes, and therefore a join for the relational database. The fifth query tests the full-text search capabilities by using a NEAR operator for proximity searches.

Analysis

To compare the query times all queries where submitted from shell scripts and all generated output was written to /dev/null which was done to limit the influence of disk write speed but still force the dbs's to generate all output. Although in a high

throughput environment one would normally integrate the database access within an application we chose this approach as to not test the integration of the database but the performance of the database itself. The query times were measured with the Linux /usr/bin/time tool, each query was repeated 25 times with different query terms for a consistent measurement and to eliminate bias due to caching of data by the database system itself and by the Linux operating system. All recorded times were then averaged per query.

Authors' contributions

JR did most of the analysis, wrote the manuscript and prepared all figures and tables. JL provided the initial idea, feedback and proof-reading.

All authors agreed on the manuscripts.

Acknowledgements

John van Hall and Martijn Liebrand set up the Oracle System and Harm Nijveen installed most of the other systems and provided the MRS and SRS parsers.

This research has been conducted with funding of NBIC/BioRange.

References

Alias-i LingPipe [http://alias-i.com/lingpipe/]

Apache Lucene Java [http://lucene.apache.org/java]

BaseX the XML Database [http://www.inf.uni-konstanz.de/dbis/basex/index]

Publicly accessible SRS servers [http://downloads.biowisdomsrs.com/publicsrs.html]

Boncz P, Grust T, Keulen Mv, Manegold S, Rittinger J et al. 2006. MonetDB/XQuery: a fast XQuery processor powered by a relational engine. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*. ACM, Chicago, IL, USA.

dtSearch [http://www.dtsearch.com/]

Etzold T, Argos P. 1993. SRS--an indexing and retrieval tool for flat file data libraries. *Computational Applications in the Biosciences* **9**(1): 49-57.

Fomichev A, Grinev M, Kuznetsov S. 2006. Sedna: A Native XML DBMS. In *SOFSEM 2006: Theory and Practice of Computer Science*, pp. 272-281.

Hekkelman ML, Vriend G. 2005. MRS: a fast and compact retrieval system for biological data. *Nucleic Acids Research* **33**(suppl_2): W766-769.

Meier W. 2009. eXist: An Open Source Native XML Database. In *Web, Web-Services,*

and Database Systems, pp. 169-183.
MySQL Sandbox [http://mysqlsandbox.net/index.html]
Neerincx PBT, Leunissen JAM. 2005. Evolution of web services in bioinformatics. *Briefings in Bioinformatics* **6**(2): 178-188.
Medline [http://www.nlm.nih.gov/pubs/factsheets/medline.html]
Oracle Berkeley DB XML [http://www.oracle.com/database/berkeley-db/xml/index.html]
Steegmans B, Bourret R, Guyennet O, Kulkarni S, Priestley S et al. 2004. XML for DB2 Information Integration.
Strömbäck L, Hall D, Lambrix P. 2007. A review of standards for data exchange within systems biology. *Proteomics* **7**(6): 857-867.
Taylor CF, Field D, Sansone S-A, Aerts J, Apweiler R et al. 2008. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nat Biotech* **26**(8): 889-896.
XML Path Language (XPath) [http://www.w3.org/TR/1999/REC-xpath-19991116]
Extensible Markup Language (XML) 1.1 (Second Edition) [http://www.w3.org/TR/2006/REC-xml11-20060816]
XQuery 1.0: An XML Query Language [http://www.w3.org/TR/2007/REC-xquery-20070123/]

Chapter 3

Two Thesauri for Identifying Enzymes and Metabolites in Text

Judith E Risse[1,2], Harm Nijveen[1,2], Paul E van der Vet[1,3]

[1]Laboratory of Bioinformatics, Wageningen University and Research Centre, P.O. Box 569, 6700 AN Wageningen, the Netherlands
[2]Netherlands Bioinformatics Centre (NBIC), P.O. Box 9101, 6500 HB Nijmegen, the Netherlands
[3] Human Media Interaction Group, Department of Computer Science, University of Twente, P.O. Box 217, 7500 AE Enschede, the Netherlands

Submitted for publication

Abstract

Background

Literature databases like PubMed provide access to large amounts of biomedical knowledge, yet the format of natural language makes these databases difficult to access by computational means. The field of text mining provides solutions to extracting information from free text. Most of the approaches require expert knowledge in the form of thesauri in order to identify specialist terms like enzyme or metabolite names in the text. The creation of these thesauri is very labour-intensive and they are therefore not readily available.

Results

In this paper we present two thesauri, created from existing data sources by semi-automated means, suitable for text mining: an enzyme thesaurus containing 4,588 EC-numbers as categories with 85,248 synonyms and a compound thesaurus containing 16,003 biologically relevant compounds as categories with 73,055 synonyms. These thesauri are specifically tailored to identify metabolic reactions in text.
When testing the capability to recoverenzymes with the thesaurus from a manually annotated set of PubMed abstracts the enzyme thesaurus performs well (F-score 0.63). The compound thesaurus has been compared with Jochem, a chemical thesaurus reported in literature. Our thesaurus is much smaller and therefore computationally favourable, and has the same global performance on a manually annotated set of PubMed abstracts (F-scores 0.64 vs. 0.61) but at higher precision (0.71 vs. 0.60).

Conclusions

The enzyme and compound thesauri described here are compact and of high quality. To our knowledge the enzyme thesaurus is the first in its kind that is publicly available. Both our thesauri fill an important niche in the field of text mining and are valuable resources for reconstructing metabolic pathways from literature.

Availability

Both thesauri and the scripts used in the automated steps are freely available for download on the website http://www.bioinformatics.nl/thesauri

Background

The large collections of publicly available biological text (PubMed, PMC) provide a treasure trove of information for biologists and bioinformaticians but the format of free text hampers automated extraction of that information. Highlighting and extracting information from free text is a task for the text mining domain and a large set of text mining services is available for this purpose (Friedman et al. 2001; Liu et al. 2004; Settles 2005). Biomedical terms found in the texts and in particular their interrelationships can aid in scientific hypothesis formation and also add to data already stored in biological databases (Waagmeester et al. 2009). A thesaurus, an ordered collection of terms, can be used to identify entities of interest. We investigate enzymes and metabolites for text mining for metabolomics.

Metabolic reactions and pathways have not been the focus of text mining-based relationship extraction as it requires the identification of two very different types of entities, namely enzymes and metabolites (Czarnecki et al. 2012) and their relationships. In this paper we provide a solution to named entity identification (NEI) of enzymes and metabolites in a dictionary-based approach by providing an enzyme and a metabolite thesaurus suitable for this task.

In the related task of named entity recognition (NER), terms of a category of interest (such as enzymes and metabolites) are highlighted in or extracted from text and then assigned to the appropriate category (Cohen&Hersh 2005; Zweigenbaum et al. 2007). Named entity identification is the ability to subsequently link recognised terms to database identifiers (Hettne et al. 2009). NEI is a prerequisite to integrate information found in text with existing database sources. There are three main approaches to NER and NEI: rule-based, statistical (or machine learning) and dictionary-based (Cohen&Hersh 2005). Rule-based approaches make use of regular expressions and other rules to identify specific terms or phrases in text. Statistical approaches use statistical pattern learning and supervised machine learning. Dictionary-based approaches make use of terms collected in a dictionary or thesaurus to identify these in text. Rule-based and statistical methods are very good at NER but require additional steps to provide the links to other data-sources, leaving dictionary-based methods as the only option capable of performing both NER and NEI directly.

The specialized thesauri necessary for a dictionary-based approach enable us to group different terms with the same meaning (synonyms) into categories (synsets) and therefore allow us to link seemingly disparate mentions of terms. The references to other data sources can either be through category identifiers or as part of the list of synonyms. The quality of a thesaurus is determined by coverage, both in categories and synonyms, granularity, and synset purity. Granularity describes the synset boundaries: what is grouped together into one synset. For example, are citrate and citric acid grouped together into one synset or not? In our case each synset should comprise terms representing a single enzyme or compound as used in literature. This means that we group “citrate” and “citric acid” together into one synset because many authors do not distinguish between the two. Synset purity is defined here as

the proportion of terms belonging to exactly one category with respect to all terms, expressed as a percentage. If the synset purity is low, the probability that an identified term maps to multiple synsets and therefore multiple identifiers is high. In domains with many homonyms, high synset purity cannot be attained. Fortunately, in the metabolomics domain the proportion of homonyms is very low for compounds and still quite low for enzymes.

For our purpose of identifying metabolic reactions in text we have chosen to pursue a dictionary-based approach. We have constructed the thesauri ourselves because thesauri suitable for metabolomics are not readily available in the public domain. Two efforts in providing thesauri to the bioinformatics community are Biothesaurus (Liu et al. 2006) and BioLexicon (BioLexicon). Biothesaurus is a source for gene and protein names. BioLexicon currently contains terminologies for enzymes (IUPAC-IUBMB (Enzyme Nomenclature)), chemical compounds (3-star compounds of ChEBI (Hastings et al. 2013)), genes and proteins, and species. Biothesaurus and the gene/protein terminologies of BioLexicon do not cover our domain. The enzyme terminology of BioLexicon is limited to the IUPAC-IUBMB enzyme nomenclature list and at the time of writing (26 Feb 2013) contains 4288 names and an additional 8082 synonyms. This means an average of only three synonyms per enzyme, this does not reflect realistic usage (e.g. EC 1.1.1.92 has 9 synonyms) making this set is too limited for our purposes.

The Jochem thesaurus of Hettne et al. (Hettne et al. 2009) contains chemical entities and is made by a largely automated process from a number of data sources, among which the ones we used in our thesaurus, ChEBI and KEGG. We use Jochem as comparison to our thesaurus.

Nobata et al. (Nobata et al. 2011) employ an entirely automated two-step method to solve the problem machine learning approaches have in NEI by combining the machine learning technique of conditional random fields (CRF) for the NER task with the mapping of the resulting terms to chemical structures in the ChemSpider database in the subsequent term identification (NEI) step (ChemSpider). Although the machine learning step does not require a thesaurus, they apply a protein thesaurus when training the CRF algorithm to boost performance (Sasaki et al. 2008). Their NER step has very high precision and recall (0.83 and 0.74), while in the subsequent NEI step 26% of the recognised terms are not mapped to ChemSpider structures.

By way of an interesting alternative to providing a static thesaurus, two groups propose recognition on the fly. Engelken et al. (Engelken et al. 2009) have formalized the IUPAC-IUBMB rewrite rules for chemicals in order to match a broader set of terms. Unfortunately there is to our knowledge no software code available in the public domain, only a web application that performs the normalization of the query term internally and only displays the matching terms. As there is no API, it is not possible to utilize this system as part of a processing pipeline. In a similar vein, Corbett and Murray-Rust propose Oscar3 (Corbett&Murray-Rust 2006), an approach that identifies chemical entities while the parser analyses the natural-language text. Oscar3 can also be used as a standalone application. It makes use of

machine learning to identify potential chemical entities. It maps the terms it finds in two ways to categories: through simple lookup and through parsing systematic names. Lookup is based on a table that was initially populated using ChEBI. Both proposals work for narrow domains and are not generally applicable. Hettne et al. compare their Jochem thesaurus to Oscar3.

Manual construction of thesauri requires expert knowledge and is very time-consuming. It also implies an organization that can guarantee sustainable maintenance. We therefore have chosen an approach that minimizes the manual effort. We have constructed our two thesauri in two steps. In the first step relevant information was harvested from source databases by automated text processing. For the enzyme thesaurus we used the KEGG enzyme and Brenda databases, and for the compound thesaurus we used ChEBI and KEGG compound. In the second step the resulting lists were manually curated to obtain high-quality thesauri. We thus combine the best of both worlds without excessive effort.

Materials and methods

For the creation of the thesauri we merged different data sources to obtain a preliminary thesaurus and then applied common automated filtering steps. We manually curated the result to obtain the final thesauri.

Enzyme thesaurus

The thesaurus was obtained by merging KEGG's enzyme table (downloaded 26-05-2011, the last freely available version) with Brenda (RN, SN and SY lines, release 11-2011, downloaded 29-09-2012) using EC numbers as identifiers. We decided to follow the most recent IUPAC/IUBMB release in those cases in which Brenda flagged entries as 'deleted', as 'transferred', or as 'preliminary supplied entries'. In a further automated step described below, a number of filters were applied to remove unsuitable entries and duplicates. The output of the first step was curated manually as described below.

To compare the performance of our enzyme thesaurus with that of its individual sources, one thesaurus was built from Brenda alone and one from KEGG alone using only the automated steps.

Compound thesaurus

The compound thesaurus was created by merging the KEGG compound table with the ChEBI dataset. The KEGG table was downloaded on 26-5-2011 and is the last freely available version; ChEBI was downloaded on 31-5-2012. The ChEBI terms were mapped to KEGG identifiers based on cross-references and the children of cross-referenced ChEBI identifiers, resulting in a flat thesaurus with KEGG compound ids as term identifiers. We chose KEGG's granularity over that of ChEBI as ChEBI, although chemically more accurate, is in our eyes too fine-grained for text

mining. For example, in ChEBI conjugate bases and acids are preferably stored as compounds separate from the base form, whereas KEGG has one form and authors are generally not that precise in their use of language. “Citric acid” (ChEBI 30769) and the conjugate base “citrate” (ChEBI 35804), for instance, both map to KEGG C00158.

An automated filtering step was followed by manual curation. Both are described below. The result before the manual cleaning steps is a thesaurus with 94,678 entries, of which 38,403 stem uniquely from KEGG, 44,227 uniquely from ChEBI, and the remaining 12,058 make up an again small number of overlapping entries (see also Figure 3b). The final compound thesaurus consists of 73,055 entries in 16,003 categories.

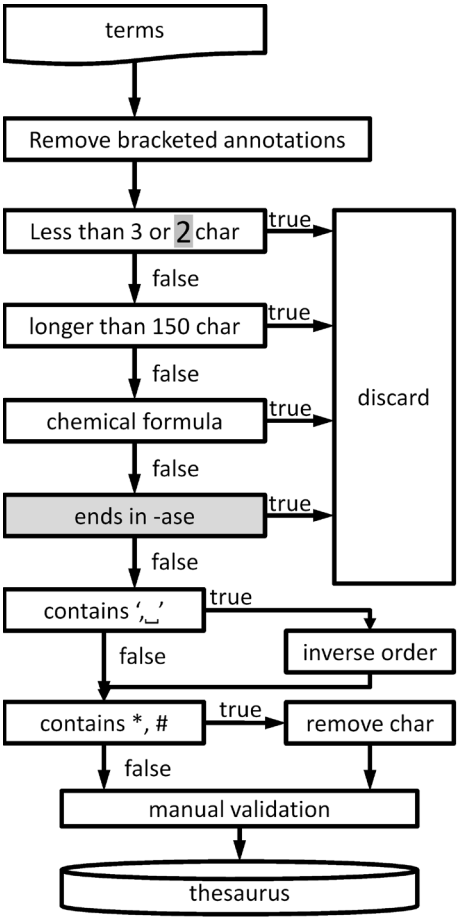

Figure 1– Filtering steps

Schematic of steps taken to filter and clean the terms before they enter the final thesaurus. (in grey: this step or figure is only relevant to the compound thesaurus; the step of inverting order refers to e.g. ‘dehydrogenase, alcohol’ -> ‘alcohol dehydrogenase’). In the discarded box are the rounded percentages of terms discarded from the enzyme thesaurus and the compound thesaurus (in grey).

As Hettne et al. (Hettne et al. 2009) have already reported on the performance of thesauri built from KEGG compound and ChEBI sources, we did not prepare versions of the compound thesaurus from KEGG or ChEBI alone.

Automated filtering steps

The automated filtering steps of the enzyme thesaurus process and the compound thesaurus process are largely shared (Figure 1). We have profited from earlier work by Schuemie et al. (Schuemie et al. 2007). The steps are, per term:

- Remove semicolons at the end of a term.
- Delete XML-tags from ChEBI notations such as "OH<small-sub>2</smallsub>" to obtain "OH2".
- Remove bracketed annotations. For example, "(isozyme)" is often added to an enzyme entry in both KEGG and BRENDA (Table 1). The compound thesaurus also has comma-separated add-ons such as "inner salt" that are removed as well. The list can be found in Table 2.
- If the number of characters (*n)* that make up the term, is smaller than 3 (n<2 for the compound thesaurus), remove the term. If n>150, remove the term.
- If the term is a chemical formula that covers many isomers, remove it. Such a term is not suited to identify a single compound. We have implemented this by using a regular expression to identify all chemical formulae of the general appearance C_nH_nR, where "*n*" stands for any integer number and "R" stands for any continuation.
- (For the compound thesaurus only:) If the term ends in "-ase", remove it. If the term is an EC-number, remove it. Enzymes are listed in the enzyme thesaurus.
- (For the compound thesaurus only:) If a term contains "Same as:", "Source" or "see", remove it (Table 2).
- (For the enzyme thesaurus only:) If the term contains an indication that the entry is invalid, remove the line. An example is "Deleted entry". The complete list can be found in Table1.
- If the term contains a comma followed by space character, we surmise that the order of term constituents has been inverted. For example, sources may well list "dehydrogenase, malate" rather than "malate dehydrogenase". In this step the original order is restored. If there is more than one comma-space occurrence, this step is skipped. Such occurrences are dealt with in the manual curation step.

- If the term contains one or more of the characters "#" and/or "*", remove those characters. For example, BRENDA delimits references with hash-signs, and ChEBI has the asterisk in some radicals.

Removal of bracketed terms		
wild type	acylating	... hydroxylating
mutant enzyme	acetylating	... cleaving
... forming	cytochrome	... eliminating
... specific	... transferring	isozyme
decarboxylating	dearomatizing	Swissprot
incorrect	demethylating	… decyclizing
ambiguous	... cyclizing	dehydrating
acceptor	... hydrolising	bifunctional enzyme
phosphorylating	... hydrolyzing	isoform
Removal of semicolon at end of line		
Removal of lines containing		
entries shorter than 3 or longer than 150 characters	Deleted entry	deleted, included in
Transferred to	formerly	Preliminary Brenda EC numbers
Inversion of "comma space" separated terms containing on occurrence of comma followed by a space		

Table 1 – Details of automated steps for Enzyme Thesaurus

Removal of bracketed terms		
[cpd…]	including	tn
Removal of comma separated add-ons		
inner salt	human	nitrogenous compound
normal	mouse	
Removal of lines containing		
lines shorter than 2 or longer than 150 characters	Same as:	see
enzymes (ending in –ase)	EC numbers	
generic compound in reaction hierarchy	Source	
Removal or conversion of xml tags from ChEBI database		
Inversion of "comma space" separated terms containing on occurrence of comma followed by a space		

Table 2 – Details of automated steps for Compound Thesaurus

Manual curation

The results of the automatic filtering steps were manually curated. We inspected each term, paying attention to the following issues:

- Incorrect terms that were included in the course of the automated steps were removed. Examples of incorrect terms are too generic terms such as "protein" and "nucleotide", and terms that do not even denote an enzyme or compound at all such as "name" and "data".
- All duplicates were inspected. A few duplicates are homonyms and thus were retained. Examples are ambiguous abbreviations and double function enzymes such as "3-dehydroquinate_dehydratase/shikimate_dehydrogenase" that is in the synset of EC 4.2.1.10 and in that of EC 1.1.1.25. Most duplicates, however, are redundant, incorrect or too general, unnecessarily lowering the synset purity. Such duplicates were always removed. For example, for the enzyme thesaurus some duplicates arose from the different versions of IUPAC nomenclature, where in one data source an entry had been moved or deleted and in the other data source it had not, resulting in two different EC numbers for one enzyme. Here we followed the most recent IUPAC release and removed the old entry. Other duplicate terms in the enzyme thesaurus only occur in Brenda and do not have a literature reference associated with them. We considered them to be too tentative to merit inclusion in the thesaurus. An example is "20-alpha-HSD" for EC 1.1.1.21: aldehyde reductase. Still other duplicates arose because of the use of general terms to denote more specialised terms, for example "alcohol dehydrogenase" as synonym for the more specialised alcohol dehydrogenases such as cinnamyl alcohol dehydrogenase. Duplicates also occur in the compound thesaurus after automated filtering. Here, too, generic terms occur that are used to denote more specialized terms. For example, "alcohol" for all alcohols; in this particular case, "alcohol" was kept as synonym for "ethanol" and removed from other synsets. In the case of stereomeres where the generic term had been assigned to both the L and the DL form we manually assigned the generic term to the L-form. For example, glutamate is assigned to both L-glutamate and DL-glutamate while in literature it usually refers to L-glutamate.
- Remaining bracketed annotations were checked. Sometimes the annotation contained a typing mistake and belonged in one of the lists of Tables 1 and 3. Other erroneous cases were also encountered. For example, in the enzyme thesaurus the string "(L-idonate-forming)" was added in front of the enzyme name. In the compound thesaurus

we encountered, for example, ""An acid" also means a carboxylic acid (see [CPD:C00060])". Such annotations were removed.

- Expressions with multiple comma-space occurrences, indicating inverted constituent order. Such expressions occurred quite frequently. There are no generally valid rules to handle such cases and they therefore have to be normalised by hand. For example, in the enzyme thesaurus we found "isomerase, 3-carboxy-cis, cis-muconate cyclo-" that we converted into "3-carboxy-cis, cis-muconate cyclo-isomerase". Likewise, in the compound thesaurus we converted "2-Butenoic acid, 2-methyl-, ethyl ester, (2E)-" into "(2E)-2-methyl 2-butenoic acid ethyl ester".
- Lines with explanations were removed. In the enzyme thesaurus we found sentences such as "the purified AdhE exhibits high enzymatic activity attributed to aldehyde dehydrogenase and low alcohol dehydrogenase activity".
- For the enzyme thesaurus, terms annotated by KEGG or BRENDA as ambiguous or incorrect were removed. However, those entries were retained if they are unique in the set, as they do not influence synset purity and increase coverage.

Validation

We assessed the performance of our enzyme thesaurus in comparison to the two original data sources Brenda and KEGG with the gold standard of 296 abstracts made available by the National Centre for Text Mining (NaCTeM) as the Enzyme and Metabolite corpus (http://www.nactem.ac.uk/metabolite-corpus/) (Nobata et al. 2011) (Figure 2). We will call this corpus the NaCTeM corpus from now on. We compared the performance of our compound thesaurus with the Jochem thesaurus of Hettne et al. (version 1.2 from http://www.biosemantics.org/chemlist) and the two-step method of Nobata et al. Here, too, we used the NaCTeM corpus as it is specifically geared towards identifying metabolites in text.

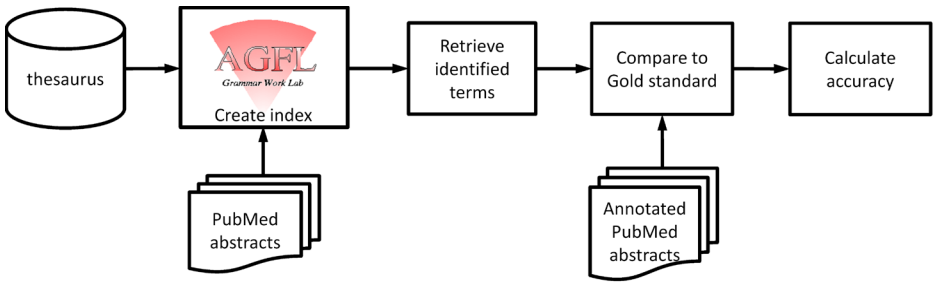

Figure 2 – Validation
Schematic of the steps taken to validate the performance of the thesauri.

To identify terms in the text we used the EP4IR parser/transducer which is generated by AGFL (version 2.8) (Koster&Verbruggen 2002) from the EP4IR grammar, a generic English grammar for information retrieval, and lexicon in combination with a lexicalized version of UMLS (UMLS) and the thesauri created for this paper. The system does a syntactic analysis of the text using a rule-based approach in combination with thesauri and lexicons indicating multi-word terms and part of speech. Information about terms is stored in dependency triples indicating a syntactic relationship allowing specific search queries. For example, the sentence "Malate dehydrogenase converts malate" is represented as "malate dehydrogenase SUBJ convert" and "convert OBJ malate". The terms stored in the triples are normalized to singular in case of nouns and infinitive in case of verbs, passive sentences ("malate was reduced by …") are stored in their active form. This triple extraction from the parse tree and the normalisation were done by using the unnester and normalizer from the Linguistic Classification System (LCS, version 2.4) (Koster&Seutter 2003). This parsing approach is copied from the PHASAR system (Koster et al. 2006).
For use in the parse-chain, all thesauri were converted to the required format and we parsed the NaCTeM corpus. For each thesaurus (enzyme, compound, Jochem) and the raw datasets (KEGG, BRENDA) the identified terms were then retrieved from the list of triples and compared to the respective gold standard.
For the performance analysis, we mapped identified terms to thesaurus terms and counted the number of occurrences. Occurrences can be counted in two ways: as a bag (five occurrences of term *X* successfully recognized by the thesaurus count as five when calculating recall and precision), or as a set (an arbitrary number of occurrences of term *X* successfully recognized by the thesaurus counts as one when calculating recall and precision). Counting occurrences as a bag may give distorted results when, for example, a few terms occur very often in the texts used for validation. We chose to count occurrences as a set to avoid this bias and get a better indication of the coverage of a given thesaurus. Hettne et al. (Hettne et al. 2009) and Nobata et al. (Nobata et al. 2011) count occurrences as a bag. For comparison we therefore also counted occurrences as a bag and included those figures in the tables with results.

Determining accuracy

Manually identified terms from the abstracts were matched with the terms identified computationally using the thesaurus. Identical terms were counted as true positives, falsely identified terms as false positives and non-identified terms that should have been identified as false negatives.
For both thesauri, partial matches may lead to double penalties. An example is the term "2-methyl butyraldehyde" mentioned in the text while the compound thesaurus only recognizes the partial "butyraldehyde". A partial can be interpreted as a false positive (for the partial) and as a false negative (for the complete term). To avoid double penalties, we manually removed the partials from the false positives.

Precision, recall and unweighted F-score have been calculated in the usual way (TP = number of true positives, FP = number of false positives, FN = number of false negatives):
Precision P = TP/(TP + FP)
Recall R = TP/(TP + FN)
F-score = (2·P·R)/(P+R) = 2·TP/(2·TP + FN + FP)

Results

In order to extract metabolic reactions from text we created two thesauri, one containing enzymes and one containing chemical compounds (i.e. metabolites). The thesauri are available for download at http://www.bioinformatics.nl/thesauri/ in a tab-delimited text format.

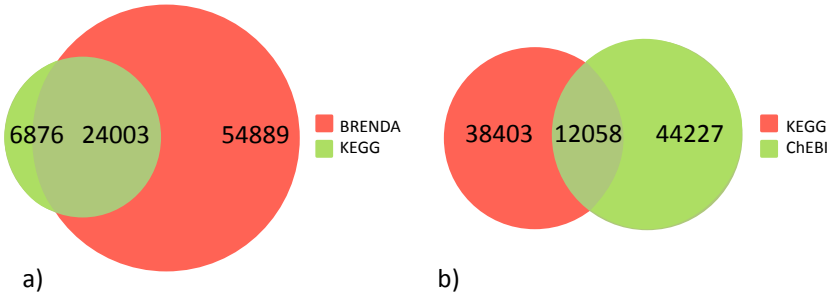

Figure 3 – Contribution of data sources
Shows the contribution of the original data source to the final complete thesauri: a) enzyme thesaurus with sources KEGG enzyme and BRENDA, b) compound thesaurus with sources KEGG compound and ChEBI.

Enzyme thesaurus

We created an enzyme thesaurus with 4,588 categories and a total of 85,248 terms. Of this total 6,876 terms were uniquely provided by KEGG and 54,889 by Brenda. The remaining 24,003 terms are found in both data sources (Figure 3a). The modest overlap between the two sources is remarkable, we expected Brenda as broader source to contain almost all terms from KEGG.
Our enzyme thesaurus has a synset purity of over 95.5%. In the validation run on the NaCTeM corpus and counting occurrences as a set, it has an F-score of 0.63 (precision 0.62, recall 0.62) (Table 3). The performance of the individual data sources on the same set with all manual and automatic steps applied was slightly below that of our complete thesaurus, with KEGG having an F-score of 0.57 (precision 0.85, recall 0.43) and BRENDA an F-score of 0.61 (precision 0.58, recall 0.65) (Table 3).

	enzyme	KEGG	BRENDA
categories	4588		
terms	85248	32063	82253
synset purity in %	95.5	95.1	87.5
true positives	371	243	371
false positives	232	43	273
false negatives	200	328	200
precision	0.62	0.85	0.58
recall	0.62	0.43	0.65
F-score	0.63	0.57	0.61

Table 3 – Performance of the enzyme thesaurus
Performance of the enzyme thesaurus in identifying enzyme mentions in the NaCTeM corpus and the performance of the contributing data sources after automatic curation.

Compound thesaurus

We created a compound thesaurus with 16,003 categories and 73,055 terms. The overlap of terms between Jochem and our compound thesaurus is 37,574 terms when taking into account only the unique terms in both thesauri and ignoring (upper- and lower-) case. The synset purity of the compound thesaurus is 99.8%.
On the NaCTeM corpus, our compound thesaurus has an F-score of 0.64 (precision 0.71, recall 0.59), again counting as a set. Jochem has an F-score of 0.61 (precision 0.60, recall 0.62). See Table 4, where one can also find the figures counted as a bag. Our data are difficult to compare with the data provided by Nobata et al. (Nobata et al. 2011) for their two-step method because they only provide recall, precision, and F-score for the NER step (recall 0.74, precision 0.83, F-score 0.78). However, from the data provided by Nobata et al. we calculate that 26% of the terms recognized in the NER step are subsequently lost in the NEI step. This makes their performance comparable to that of our thesaurus.
In the NEI task our compound thesaurus maps 90% of the identified terms in the test set to a single category and therefore one database identifier. Jochem maps 19% of the true positives identified in the test set to more than one synset. In total 6.4% of all synonyms (excluding chemical formulas) in the Jochem thesaurus map to more than one category compared to 1% in our compound thesaurus.

	as in Nobata et al.		unique mentions and excluding parse errors	
	compound	Jochem	compound	Jochem
categories	16003	278581		
terms	73055	1691507		
synset purity in %	99.8	93.6		
true positives	695	734	327	345
false positives	195	303	131	227
false negatives	1743	17044	231	213
precision	0.78	0.70	0.71	0.60
recall	0.29	0.30	0.59	0.62
F-score	0.42	0.42	0.64	0.61

Table 4 – Performance of the compound and Jochem thesauri
Performance of the compound thesaurus and Jochem in identifying metabolite mentions in the NaCTeM corpus.

Discussion

Manual creation of a thesaurus is labour-intensive and implies some form of organization that guarantees sustainable maintenance. The general perception is that this does not always warrant the effort. In creating a thesaurus completely by computational methods Hettne et al. provide a time saving alternative that also compares well to a manually curated system (Hettne et al. 2010). When creating such a large thesaurus from multiple data sources it is difficult to maintain control over granularity and synset purity. Synset purity is particularly important when a post-parsing disambiguation step is not appropriate for the applied parsing strategy. In our approach we maintain control over granularity and synset purity by using the best of both worlds: we use computational steps to create the thesauri from several data sources and we use manual curation to maintain even granularity and synset purity throughout the resulting merged thesaurus.
Another disadvantage of thesauri is their static nature and inability to recognise new terms requiring constant updates. To alleviate the amount of labour required to update or rebuild a thesaurus we suggest using a database that keeps track of all manual interventions. We propose that many steps that were performed manually in the first run can be done automatically, by creating custom rules, in the second and further runs, reducing the amount of manual work needed for maintenance. On the other hand, alternatives to dictionary-based NER methods like machine-learning require annotated training corpora to train their algorithms. Annotated corpora are also labour-intensive to construct and are necessary for each different type of text for the algorithm to maintain an acceptable level of performance.
Two major factors are responsible for the quality of a thesaurus, one being the coverage

of relevant terms in the area of interest and the other being the quality of synonym sets assigned to categories. In our particular use case of reconstructing metabolic networks through text mining, it is vitally important that there is a clear separation between categories and that the boundaries are not blurred by synonyms assigned to incorrect categories or to multiple categories. The synset purity of a thesaurus is seen to depend on the domain and on the way the thesaurus is constructed. Certain domains (for example gene names in the genomics field) are notorious for the high prevalence of homonyms. In such a domain, high synset purity cannot be achieved. If the thesaurus is constructed automatically from multiple sources, which also might have slightly varying granularities, we end up with relatively low synset purity as well. Hettne et al. partially solved the problem of low synset purity by introducing disambiguation rules to determine whether an identified term is the preferred name for an entity and whether there is corroborating evidence in the surrounding text for a particular meaning. This solution is inevitable for domains in which the occurrence of many homonyms makes high synset purity impossible but it has problems of its own. Retrieving disambiguation information from surrounding text slows down the parsing process because it necessitates an extra step. Worse, the information may not be present in the surrounding text and may even not be present in the text at all.

In the application for which we developed the thesauri described in this paper, we extract information about enzymatic reactions from literature and store them in syntactic triples with the identified term mapped to its corresponding identifier. Multiple triples are then connected and joined into a metabolic network based on the common identifiers. In this approach a lack of synset purity causes the number of triples to be multiplied and the resulting network to expand rapidly. A lack of synset purity also results in incorrectly assigned edges when a post-parsing disambiguation step is not performed.

While we joined two datasets for our thesaurus, the relatively small size makes it still feasible to manually clean ambiguous synonyms and assign them to the appropriate synset. The synonyms that are removed are in some cases plain mistakes in one of the source datasets. In other occasions the granularity of one dataset is different from that of another dataset as when, for instance, one dataset has only one entry for an entity and the other dataset has two. The duplicate terms that remain are mainly ambiguous abbreviations, where both meanings are used in literature. For a concrete example, “pdh” can stand for pyruvate dehydrogenase (EC 1.2.4.1), for prephenate dehydrogenase (EC 1.3.1.12), and for proline dehydrogenase (EC 1.5.99.8).

The results of the enzyme thesaurus show another form of ambiguity within the text: It is difficult to distinguish between gene names and enzyme name abbreviations in the text without additional context information. This results in a lower precision in the enzyme thesaurus benchmark compared to the compound benchmark.

Precision, recall and the resulting F-score are important factors when assessing the quality of thesauri. As expected, our smaller compound thesaurus has a lower recall than the bigger Jochem thesaurus, but where the size of our thesaurus is only 4.3% of that of Jochem, we are able to recover 95% of their true positives. Moreover, due

to the manual curation step our compound thesaurus compared to Jochem achieves a significantly higher precision on the test-set (Table 4). This large difference in size between our compound thesaurus and Jochem is largely due to their use of PubChem as data source. In comparison to the source datasets KEGG and ChEBI the Hettne et al. paper already concludes that both perform less well than Jochem and therefore must perform also worse than our compound thesaurus.

When comparing the NER performance of our compound thesaurus and Jochem to the two methods described by Nobata et al., our compound thesaurus and Jochem perform in the same league as the Nobata dictionary-based method and worse than the Nobata CRF-NER. As Nobata et al. (Nobata et al. 2011) already discuss, the difficulty for thesauri lies in the complex definition of metabolites employed, where also the context of a term determines whether it has the role of metabolite in that sentence.

Using thesauri for NER-tasks does not require an annotated training set to perform the machine learning task and is therefore more flexible in use. The costs of manual curation have to be compared to those of producing an annotated corpus for training a machine learning algorithm such as CRF. The excellent performance of the CRF algorithm in NER is somewhat mitigated by the loss of performance in the NEI task due to problems in mapping recognized terms to the appropriate chemical structures and, by consequence, database identifiers.

We conclude that our thesauri are compact and of high quality. We are not aware of another publicly available enzyme thesaurus. We expect that both our thesauri will fill an important niche in the field of text mining and reconstructing metabolic pathways from literature.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

JER created the thesauri, analysed the results and drafted the manuscript. HN participated in setting up the study, supported the technical aspects of the analysis and gave feedback on the manuscript. PEvdV participated in analysing the results and critically revised the manuscript.
All authors read and approved the final manuscript.

Acknowledgements

The authors would like to thank Sophia Ananiadou and Paul Thompson of the University of Manchester for publishing the enzyme tags of the NaCTeM corpus.
The authors would like to thank Sandra Smit for the feedback on the manuscript.
This research has been funded by the Netherlands Bioinformatics Centre as part of the BioRange Project SP 4.1.1.
The authors are also indebted to the late Jack Leunissen who sadly passed away before the manuscript was finished.

References

BioLexicon [http://www.ebi.ac.uk/Rebholz-srv/BioLexicon/biolexicon.html]

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

Cohen AM, Hersh WR. 2005. A survey of current work in biomedical text mining. *Briefings in Bioinformatics* **6**(1): 57-71.

Corbett P, Murray-Rust P. 2006. High-Throughput identification of chemistry in life science texts. In *Proceedings of the Second international conference on Computational Life Sciences*, pp. 107-118. Springer-Verlag, Cambridge, UK.

Czarnecki J, Nobeli I, Smith A, Shepherd A. 2012. A text-mining system for extracting metabolic reactions from full-text articles. *BMC Bioinformatics* **13**(1): 172.

Engelken H, Golebiewski M, Bittkowski M, Hamm F, Saric J et al. 2009. Flache und semantische Verarbeitung von Namen biochemischer Verbindungen. *INFORMATIK - Im Focus das Leben*: 687-692.

Enzyme Nomenclature [http://www.chem.qmul.ac.uk/iubmb/enzyme/]

Friedman C, Kra P, Yu H, Krauthammer M, Rzhetsky A. 2001. GENIES: A natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics* **17**(SUPPL. 1).

Hastings J, de Matos P, Dekker A, Ennis M, Harsha B et al. 2013. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Research* **41**(D1): D456-D463.

Hettne K, Williams A, van Mulligen E, Kleinjans J, Tkachenko V et al. 2010. Automatic vs. manual curation of a multi-source chemical dictionary: the impact on text mining. *Journal of Cheminformatics* **2**(1): 4.

Hettne KM, Stierum RH, Schuemie MJ, Hendriksen PJM, Schijvenaars BJA et al. 2009. A dictionary to identify small molecules and drugs in free text. *Bioinformatics* **25**(22): 2983-2991.

Koster C, Seibert O, Seutter M. 2006. The PHASAR Search Engine. In *Natural Language Processing and Information Systems*, Vol 3999, pp. 141-152-152. Springer Berlin / Heidelberg.

Koster CA, Seutter M. 2003. Taming Wild Phrases. In *Advances in Information Retrieval*, Vol 2633 (ed. F Sebastiani), pp. 161-176. Springer Berlin Heidelberg.

Koster CHA, Verbruggen E. 2002. The AGFL Grammar Work Lab. In *Proceedings of the FREENIX Track: 2002 USENIX Annual Technical Conference*. USENIX Association.

Liu H, Hu Z-Z, Zhang J, Wu C. 2006. BioThesaurus: a web-based thesaurus of protein and gene names. *Bioinformatics* **22**(1): 103-105.

Liu H, Wu C, Friedman C. 2004. BioTagger: a biological entity tagging system. *BioCreative Workshop—A Critical Assessment of Text Mining Methods in Molecular Biology, Granada, Spain, March*: 28–31.

Nobata C, Dobson P, Iqbal S, Mendes P, Tsujii Ji et al. 2011. Mining metabolites: extracting the yeast metabolome from the literature. *Metabolomics* **7**(1): 94-101-101.

Sasaki Y, Tsuruoka Y, McNaught J, Ananiadou S. 2008. How to make the most of NE dictionaries in statistical NER. *BMC Bioinformatics* **9**(Suppl 11): S5.

Schuemie MJ, Mons B, Weeber M, Kors JA. 2007. Evaluation of techniques for increasing recall in a dictionary approach to gene and protein name identification. *Journal of Biomedical Informatics* **40**(3): 316-324.

Settles B. 2005. ABNER: An open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics* **21**(14): 3191-3192.

Unified Medical Language System (UMLS) [http://www.nlm.nih.gov/research/umls/]

Waagmeester A, Pezik P, Coort S, Tourniaire F, Evelo C et al. 2009. Pathway Enrichment Based on Text Mining and Its Validation on Carotenoid and Vitamin A Metabolism. *OMICS: A Journal of Integrative Biology* **13**(5): 367-379.

Zweigenbaum P, Demner-Fushman D, Yu H, Cohen KB. 2007. Frontiers of biomedical text mining: current progress. *Briefings in Bioinformatics* **8**(5): 358-375.

Chapter 4

Evaluation of Metabolic Reaction Extraction from Scientific Literature with a Deep Parsing Approach

Judith E Risse[1,2], Paul E van der Vet[1,3]

[1]Laboratory of Bioinformatics, Wageningen University and Research Centre, P.O. Box 569, 6700 AN Wageningen, the Netherlands
[2]Netherlands Bioinformatics Centre (NBIC), P.O. Box 9101, 6500 HB Nijmegen, the Netherlands
[3] Human Media Interaction Group, Department of Computer Science, University of Twente, P.O. Box 217, 7500 AE Enschede, the Netherlands

Submitted for publication

Abstract

Background

With the ever increasing amount of scientific publications comes the need for better access to the content of scientific papers. Text mining is the area of research geared towards unlocking the knowledge held within text. Two types of information contained within text are the most relevant to biological and bioinformatics researchers, the entities that make up their area of research and the relations between them. Here we focus on the possibilities to extract relations between enzymes and metabolites from text in the form of metabolic reactions. We apply a full parsing approach to analysis of scientific literature and compare the performance of our approach to a modified rule-based approach by Czarnecki et al. on three metabolic pathway corpora. Furthermore we present a number of experiments to assess the influence of various text-pre-processing steps on the overall parsing performance. We also assess the potential of our system to be applied to a large full-text corpus in BioMed Central.

Results

Our analysis approach to recover metabolic reactions from scientific literature has a slightly lower recall and (much) higher precision compared to Czarnecki et al. for all three studied pathways. The difference in recall for reaction recovery compared to Czarnecki et al. is limited to one or two reactions. A similar picture presents itself when looking at interaction pairs. In comparison, the lower recall for interaction pairs including the enzyme is illustrating the challenge for our approach of identifying the interacting enzyme.
The text-pre-processing experiments do not have much effect on the results, except for the conversion of schematic chemical reactions in the text to syntactically complete sentences.

Conclusions

We conclude that despite its current low recall our full parsing approach to metabolic reaction extraction has high precision and potential to be used to (re-)construct metabolic pathways in an automated setting. With future improvements to the grammar and relation extraction rules, reactions can be extracted with even higher specificity.

Introduction

Text mining can greatly assist the scientific community in collecting and cataloguing information from scientific literature, for example for confirmation of existing knowledge or for hypothesis generation. There are already many applications and approaches using text mining as tool for in-depth literature analysis, many of which focus on named entity recognition (NER) (Klinger et al. 2008; Ferreira&Couto 2010; Sayle et al. 2011) or relation extraction (RE) (Rindflesch et al. 2000; Sharma et al. 2010; Van Landeghem et al. 2012). In the wake of the genomics revolution most of these efforts focus on genes and proteins and their interactions (Settles 2005; Rinaldi et al. 2007; Miyao et al. 2009; Wang et al. 2009). We on the other hand focus on a field that was only recently discovered by the text mining community, namely that of metabolomics, including enzymes, metabolites and metabolic reactions in an attempt to identify and extract these from scientific literature.

Text mining

Text mining in its strictest definition has the purpose of knowledge discovery by making connections between different statements in texts explicit (Hearst 1999). More commonly text mining refers also to all sub-tasks required to reach this goal (Zweigenbaum et al. 2007). These subtasks range from document classification, the grouping of documents into categories, to named entity recognition (NER), the labelling and extraction of (biomedical) entities from text, to named entity identification (NEI), the identification of recognised entities as, for example, a particular enzyme, to relation extraction (RE), where RE presupposes NER. RE is the main focus of this chapter.

Relation extraction is the recognition, identification and extraction of entities of interest (*here* enzymes and metabolites) and their relationships (*here* metabolic reactions) from text, in this case scientific articles or abstracts. In our approach we map identified terms to specific entities (i.e. "alcohol dehydrogenase" to EC 1.1.1.1) instead of mere classes (i.e. "alcohol dehydrogenase *is an* enzyme") with the aid of specialist thesauri (*see* chapter 3). This allows us in a later stage to merge single reactions into complete pathways based on overlapping entities.

Relation extraction requires not only the recognition of entities of interest in the text but also the identification of the relationships that may exist between those entities. RE therefore is one of the more complex text mining tasks. Many different approaches and techniques have been applied to RE and they vary both on how difficult it is to implement them and on how specific the relations are that can be extracted. The performance of text mining approaches is measured in terms of two concepts borrowed from the Information Retrieval community, recall and precision. Briefly, recall measures the number of recovered hits as proportion in relation to the number of hits in the entire source; precision measures the number of recovered hits as proportion in relation to the number of recovered items. In the Information

Retrieval community, it is generally held that increase of recall results in a loss of precision and vice versa (Baeza-Yates&Ribeiro-Neto 2011). When we want to regard relations extracted by text mining as hypotheses, high precision is favourable, whereas usage for evidence gathering for existing knowledge benefits most from higher recall. When working with very large corpora high precision is to be preferred because it limits the number of results to be analysed (Lee et al. 2012).

A number of approaches to RE have been described in the literature. The simplest method to find relations is co-occurrence, the assumption that two or more entities within the same article, article section or sentence share a relationship (Jenssen et al. 2001). This low precision, high recall 'guilt-by association' approach cannot identify the type and direction of the relation and is also incapable of detecting any linguistic nuances or broader context. Rule-based approaches involve the usage of regular expression type patterns describing the relations. Depending on the type and number of rules, these approaches can extract relations with high precision (Lee et al. 2012) and can also often identify the type and directionality of the relations. Czarnecki et al. (Czarnecki et al. 2012) use a rule-based method combined with filtering based on a scoring system that is capable of discerning source and target of identified reactions. Machine learning and statistical approaches learn rules and/or probabilities associated with features in the text relevant to relations (Abi-Haidar et al. 2008; Alex et al. 2008; Bundschus et al. 2008). This type of approach featured strongly in the BioCreativeII PPI (protein-protein interaction) challenge (Krallinger et al. 2008). It generally is capable of high precision and recall but depends on the availability of high quality training corpora, the construction of which may take a large effort. Given the right annotations in the training corpora these approaches can also be used to identify directionality and type of the identified relations.

The text mining approach that theoretically has the highest potential of using the full range of information held within free text is the one analysing syntax and semantics. This involves complex natural language processing (NLP) methods that rely on a formal grammar of the language used, either generic or specific to the interactions of interest and full or partial parsing of the sentences in the text. Full or partial parsing approaches have been applied to biomedical RE with Link Grammar (Pyysalo et al. 2004), BioIE (Kim&Park 2004), MedScan (Novichkova et al. 2003), GenIE (Cimiano et al. 2005) and GENIES (Friedman et al. 2001).

In this study we make use of a full parsing approach using the dependency parser system AGFL with the EP4IR grammar (Koster&Verbruggen 2002), a generic English grammar, slightly modified by us to adapt it to the specific sublanguage of metabolic reactions.

Metabolomics

Metabolomics, the study of the small molecules (metabolites) in cells, is essential to understand the complex processes taking place within those cells and, as a result, in living organisms (Hollywood et al. 2006). Metabolic pathways describe series of

chemical reactions converting metabolites, each reaction catalysed by a particular enzyme. Insights into metabolic pathways allow us to interpret the findings of genomics experiments and to follow the link between a genomic mutation and its phenotypic effect (Fiehn 2002; Hall 2006). Detailed and comprehensive metabolic networks are also essential for the accurate modelling of metabolic fluxes. While knowledge about metabolites, enzymes and pathways is commonly stored in databases (e.g. KEGG (Kanehisa et al. 2006), Reactome (Croft et al. 2011), EcoCyc (Keseler et al. 2011)), studies show that this information is not complete (Stobbe et al. 2011). Labour-intensive manual efforts are underway to collect additional information from scientific literature (RECON2 (Thiele et al. 2013), The Edinburgh Human Metabolic Network (Ma et al. 2007)) and the scientific community (WikiPathways (Pico et al. 2008)). The effort needed to collect correct information in significant amounts combined with the continuous increase in scientific publications establishes a clear need for text mining solutions in this area. This we attempt with our approach. The unique challenges posed by metabolic pathways with their different entities and complex interactions provide a good use case for comparing different text mining approaches.

Evaluation

In order to evaluate the performance of our approach we compare it with that of a recent relation extraction effort in metabolomics by Czarnecki et al. (Czarnecki et al. 2012). Because we will refer to that paper frequently we abbreviate the paper and its results as CNSS, after the first letters of the author last names. CNSS uses a real-world sample collection of full-text papers referenced in three metabolic pathways in the EcoCyc database. Lacking annotated corpora for metabolic reactions, CNSS chose the second best option. The GENIA corpus (Kim et al. 2003), one of the representative corpora in the biomedical field, does not specifically tag enzymes and sentences pertaining to metabolic reactions are rare. The same holds true for the NacTem corpus (Nobata et al. 2011), despite it recently having been extended with enzyme tags.

Pre-processing experiments

We study the possibilities of a full parsing approach for the extraction of metabolic reactions from text. Our parsing approach is complex and along the whole analysis pipeline there are possible bottlenecks affecting performance. The earlier in the analysis errors occur, the greater their effect on precision and recall. We already evaluated the performance of the NEI-part of the analysis in the previous chapter with regards to the enzyme and compound thesauri. That performance can be translated to the part-of-speech (POS) labelling quality that is one of the first steps in the analysis (Figure 1). In order to only measure the performance of the parser as distinct from the effect of thesaurus quality we added the relevant terms from the test corpora to these thesauri if they were lacking (*see* Materials and Methods).

Other early parts in this process are the sentence splitting and syntax analysis of the input text. Errors occurring in this part of the analysis have a knock-on effect on the performance of subsequent steps and therefore warrant particular attention. In order to differentiate between short-comings of the grammar and easily remedied problems generated by the input texts we devised a number of experiments. The experiments show the effect of a number of text pre-processing steps, aimed at increasing recall, on the performance of our approach, again using the CNSS benchmark. Main aim is the improvement of the accuracy of the sentence splitting and syntax analysis with removal of literature and figure references from within sentences, conversion of Greek characters to words and conversion of chemical reaction symbols to syntactic constructs. We also study the impact of anaphoric references on recall. This particular experiment gives an indication about the amount of information about metabolic reactions hidden for NEI and RE approaches. The experiments together should give an overview of the potential of our full grammar approach to metabolic reaction extraction and the effects of textual features on precision and recall. It should also help to differentiate between the causes of loss of recall: text-formatting, obfuscation or grammar problems.

Material and Methods

Input data

To evaluate the performance and behaviour of our syntax parsing approach in comparison to the CNSS rule-based approach we used the same pathway literature as in CNSS. The full-text articles of the test corpora are papers that are referenced in three pathways in the EcoCyc database: the pantothenate and coenzyme A biosynthesis pathway (8 papers), the tetrahydrofolate biosynthesis pathway (13 papers) and the fatty acid beta-oxidation pathway (11 papers). We refer to the CNSS paper for details. The original papers are available as scanned pdf documents at the publishers' websites. We used the extracted text as kindly provided to us by the first author of the CNSS paper, J. Czarnecki (Birkbeck College, University of London, personal communication), ensuring the use of identical input data. While the articles used are full-text, the text analysis is, like that of CNSS, performed on the text up to but not including the Materials and Methods sections.
To give an impression of the performance of our system on a large corpus we also performed an analysis of the BioMed Central full-text corpus (including papers upto 13 September 2012).

Overview of text analysis

The texts are analysed by the dependency parser system AGFL (AGFL Grammar Work Lab 2.8) with the EP4IR grammar (Koster&Verbruggen 2002), a generic English grammar. We somewhat modified the EP4IR grammar to better suit the

peculiarities of our corpus. We will discuss those changes below. EP4IR makes use of a lexicon to determine part-of-speech, such as noun, verb or adjective. The list of recognised POS categories can be easily extended to include specialist terms. To enable the identification of compound nouns common in enzyme and metabolite names we extended the lexicon with all terms of the enzyme and metabolite thesauri (chapter 3). Because natural language is ambiguous, many sentences allow of more than one parse. Most parsers therefore generate a number of parses rather than just a single parse. The problem then becomes to identify the correct parse among the alternatives. Link-grammar (Pyysalo et al. 2004) requires a post-processing step to identify the correct parse. AGFL, by contrast, normally delivers a ranked list of parses where the rank of each parse relative to its alternatives is determined by a system of penalties and bonuses. This system can be easily adapted for individual grammar rules and lexicon terms. We have used AGFL's bonus and penalty system to adapt the parsing results to suit the specific syntactic constructs used in describing metabolic reactions. AGFL is capable of returning partial parses of sentences and syntactically incomplete utterings when a full sentence parse is not possible, making it very robust. The parser is able to convert passive to active sentences in a normalisation step, dispensing with the need for different relation extraction rules for both types of constructs.

The resulting parse tree is converted into so-called dependency triples. A dependency triple normally gathers two normalised constituents from the text and establishes a relation between them. The relations relevant to our study defined in AGFL are: SUBJ (subject), OBJ (object), PRED (predicate), ATTR (attribute), PREP (preposition) and MOD (modifier). For a complete list see Appendix Table A. 1

For example, the dependency triple "alcohol dehydrogenase, SUBJ, converts" means that in the source text "alcohol dehydrogenase" has the subject role with respect to (the verb) "convert". We subsequently query the dependency triples for patterns indicative of metabolic reactions. The integration with the two specialist thesauri allows us to map identified entities to thesaurus categories. We expect our approach to have very high precision for the extracted relations and tight control over the type of extracted relation because we are able to filter negations, to recognise tentative assertions, and to perform other high-level tasks. Like in many information retrieval experiments with very high precision we initially expect a rather low recall.

Our text analysis system consists of three major parts: text pre-processing, syntax analysis and extraction of interaction triples, each with multiple sub-steps. These steps are schematically outlined in Figure 1 with the numbers in the text referring to the corresponding step in the figure. Our approach to steps one to seven (Figure 1) of the text analysis follows that of the PHASAR system (Koster et al. 2006; Koster et al. 2007).

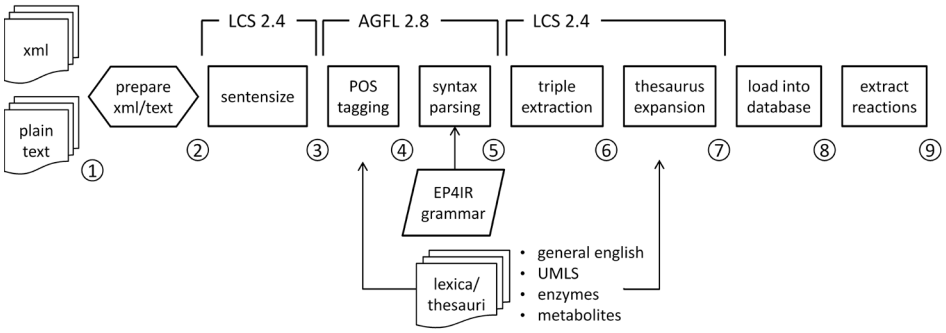

Figure 1 – Schematic view of text analysis steps

Text pre-processing

The system accepts PubMed XML, BioMed Central XBML, and flat text as input (1). The first step converts the various input formats into one uniform XML-type format suitable for subsequent analysis (2):

```
<Article>
        <DocID>…</DocID>
        <Title>…</Title>
        <Abstract>…</Abstract>
        <Body>…</Body>
</Article>
```

From any input text we extract (if possible) PubMed ID or DOI, title, abstract text, and the article text. If no PubMed ID or DOI can be identified the parser assigns an identifier assigned by a local counter that is unique within the same run. If the original text contains any XML-tags they are either removed or replaced by docid, title, abstract and body tags as above. Texts are made XML-safe by replacing XML reserved characters (<,>, &, %) with corresponding character entity references. Unicode character entity references are replaced or removed (e.g. *α* with *alpha*, for a complete list see the Appendix). The pre-processed input texts are then split into sentences using a sentence splitter from the LCS 2.4 package (Koster&Seutter 2003) (3).

Syntax analysis

In the next part syntax analysis is performed with applications from the AGFL 2.8 Grammar Work Lab and LCS 2.4. The constituents of each sentence are tagged with their part of speech (POS) using the lexica of the parser suite. Unknown words are parsed as strings and no POS is assigned (4). Where applicable, the grammar permits untagged words with a penalty.

The AGFL grammar suite is a parser generator and parser (5). The parser has been

generated from a version of the EP4IR grammar, slightly adapted to better parse the structure of biomedical language in general and metabolic reactions in particular (exact adaptations in Appendix Table A. 2 - Table A. 5) :

- EP4IR prefers 'bare' nouns over the combination of a determiner and a noun. We reversed this as the determiner was lost on some occasions.

- We added a rule specific to the identification of metabolic reaction patterns. The rule relies, among other things, on certain nouns in the lexicon annotated as denoting a metabolic conversion; examples of such nouns are "conversion" and "phosphorylation". We will call these nouns "metabolic nouns". The full list of such nouns can be found in the Appendix. Metabolites are marked as such in the lexicon. The rule recognises the pattern "m of x into y" (with some variants with respect to the prepositions), where "m" is a metabolic noun and "x" and "y" are metabolites.

- We added a rule to account for subordinate clauses occurring at the end of a sentence.

- A rich source of ambiguities in our corpus is the fact that in English "to form" can be either the indeterminate form of the infinitive of a verb or the preposition "to" followed by the noun "form". EP4IR appeared to prefer the second alternative, which sometimes created havoc in our results. We added a rule to shift parsing preference to the first alternative.

- EP4IR also appeared to handle phrases like "to form x and y" not very well. We changed infinitive clause parsing rules to give preference to 'and' as conjunction between noun phrases. As a result, "to form x and y" is analysed as the infinitive "to form" with as its object the conjunction "x and y".

To evaluate the performance on the test corpora we used a best-of-ten parses setting without time constraints. Best-of-ten parses here means the best scoring parse out of ten possible syntactic interpretations of the sentence. For analysis of the BioMed Central dataset we used the first-parse option for performance reasons. The parser makes use of a number of lexica: a general English lexicon, a lexicalized version of the UMLS specialist lexicon (2007A version) (Lindberg et al. 1993; McCray et al. 1994) and the lexicalized versions of the enzyme and metabolite thesauri discussed in chapter 3. The parser outputs a bracketed annotation of each sentence denoting the POS of the constituents and the syntax structure.
The syntax structure is then transduced and unnested into syntactic triples using the

transducer/unnester from the LCS 2.4 suite. The resulting set of syntactic triples is filtered using regular expressions to remove triples that miss one of the two terms, and triples without one of the syntactic types subject, object, predicate, attribute, modifier or preposition (6).
Wherever the term of a triple matches a term in either the enzyme or compound thesaurus, the triple is duplicated with the matching term replaced by the parent term of the thesaurus entry. In case of an enzyme the parent term is the EC number, in case of a metabolite the KEGG (Kanehisa et al. 2006) compound identifier. Contrary to CNSS, our approach requires exact string matching for enzyme and metabolite terms and results are therefore dependent on the coverage of the thesauri. In order to test the performance of the parsing approach independent of the thesaurus coverage we added the relevant enzyme and metabolite terms found in the text to the corresponding thesaurus entries if they were missing (7).
The third part of the analysis pipeline is the extraction of the actual metabolic relations from the syntactic triples. The syntactic triples are loaded into a MySQL database along with information about their origin (document ID, sentence number and position of the words within the sentence) (8). The analysis of the BioMed Central corpus (obtained 08-04-2013) was done on an identical database on the Oracle 11.2g platform for performance reasons. The metabolic relations are then extracted from the database using a set of syntactic patterns (see the Appendix Table A. 6 and Table A. 7) in combination with four word lists (Table A. 8 - Table A. 11) to filter the results (9):

a) nouns pertaining to metabolic reactions, that are found in predicative or prepositional relations with enzymes or metabolites (e.g. conversion, phosphorylation)

b) nouns, similar to a), but mainly occurring in attributive relations (e.g. enzyme, protein, metabolite)

c) verbs indicating a metabolic relationship between enzyme and compound or compound and compound (e.g. convert, catalyse, form)

d) prepositions relevant to and indicating the direction of the described reactions (e.g. by, from, into)

The lists of nouns and verbs partially overlap with the ones used by Lee et al. (Lee et al. 2012) but are specifically tailored to the sub-language of metabolic reactions as opposed to protein-protein interactions. The lists and the syntactic patterns have been constructed by evaluating a random set of Medline abstracts (NIH), and the 100 sentences used by CNSS to create their rules. The list can be found in the Appendix. The extracted relations are pairwise and either between an enzyme and a compound or between two compounds. By combining pairs, for the benchmark

only if they occur in the same sentence, a complete reaction can be recovered.

Experiments

To gain a better understanding of the performance of our analysis pipeline and the influence of the original texts on precision and recall we conducted a small number of experiments:

- **Experiment A.** Baseline, the CNSS corpus processed by means of the EP4IR grammar adapted by us.

- **Experiment B.** We replaced all Greek characters in the CNSS corpus with the written word (α -> alpha, β -> beta, etc.). In Experiment A we found that the unnester cannot handle Greek characters.

- **Experiment C.** In addition to the changes of Experiment B we removed all literature and figure references from the texts. These references confound the parser in various ways, of which we give a few examples. A long list with literature references in parentheses makes the parser fail because it no longer recognises a sentence. Due to the particular way our corpus is produced, a number reference is often glued to the preceding noun. As a result, the noun is no longer found in the lexicon and hence goes unrecognised. Moreover, when this occurs at the end of a sentence the parser sometimes regards the string as an abbreviation and the period as the end-of-abbreviation rather than the end-of-sentence.

- **Experiment D.** In the corpus obtained for Experiment C we replaced chemical reactions depicted in the text with symbols by sentence structures: "a + b → c +d" becomes "a and b are converted into c and d".

- **Experiment E.** In the corpus obtained for Experiment C, we replaced anaphoric references and short forms with the corresponding long forms. (e.g. 'This enzyme' or 'ADH' replaced by 'Alcohol dehydrogenase')

Evaluation

The performance of our system was evaluated with the same approach as CNSS. Recall in relation to the reference pathway in the EcoCyc database is determined by counting the number of reactions or binary relations in the reference pathway identified by us for each of the three pathways studied. Precision is determined by the number of identified relations that are correct, referring to the pathways. We

evaluated precision and recall both by looking at the recovered complete reactions and the recovered binary interaction pairs. CNSS also introduce variants on recall and precision that look at the number of actual relations found in the texts regardless of whether they are part of a metabolic path or not. We do not use these variants.

Results

Pathway recovery

Our analysis approach to recover metabolic reactions from scientific literature has a slightly lower recall and higher or much higher precision compared to CNSS for all three studied pathways (Table 1).

CNSS measure the ability to recover a reaction from substrate to product (ignoring compounds like ATP and phosphate) both with and without the enzyme identified. The difference in recall with our method is one (pantothenate and fatty acid β-oxidation pathway) or two (tetrahydrofolate pathway) reactions for either measurement.

Our precision, however, is higher compared to CNSS, 20-70% for reactions without the enzyme. When looking at reactions including the enzyme the increase in precision only ranges between 3% and 12%. This shows a weakness of our approach to correctly identify and link enzymes to reactions.

Disregarding enzymes for now, the higher specificity of our approach is not only evident in the figures for recall and precision but also in the total number of recovered reactions which range from one third to a quarter of the number of reactions recovered by CNSS.

	correct reactions (ignoring enzyme)		correct reactions (including enzyme)	
	CNSS	ours	CNSS	ours
pantothenate and coenzyme A pathway				
recall	78% (7/9)	66% (6/9)	56% (5/9)	44% (4/9)
precision	59% (24/41)	89% (8/9)	41% (17/41)	44% (4/9)
tetrahydrofolate pathway				
recall	90% (9/10)	70% (7/10)	70% (7/10)	50% (5/10)
precision	60% (39/65)	83% (20/24)	38% (25/65)	50% (12/24)
aerobic fatty acid beta-oxidation pathway				
recall	29% (2/7)	29% (2/7)	29% (2/7)	14% (1/7)
precision	30% (11/17)	100% (5/5)	14% (5/37)	20% (1/5)

Table 1 – Performance on three evaluation pathways
Performance of our approach for the recovery of complete reactions in all three studied pathways with (right) or without (left) enzymes compared to CNSS using the baseline settings of our experiments.

Interaction pair recovery

When looking at the results only taking into account interaction pairs (substrate – product, substrate – enzyme, enzyme – product) instead of complete reactions, a similar picture presents itself (Table 3). For the first two types recall is slightly lower for all three studied pathways and precision significantly higher with the exception of the fatty acid β-oxidation pathway substrate-enzyme pairs. Looking at the enzyme-product interaction pairs, recall is slightly lower than before again illustrating the challenge for our approach of identifying the interacting enzyme.

Experiments

The results of the experiments to measure the influence of changes to the input data on precision and recall are shown in Table 3 and the number of changes required in the test corpora in Table 4.
In Experiment B, the number of replacements for Greek characters varies greatly across the three studied pathways with the tetrahydrofolate pathway requiring seven replacements and the fatty acid β-oxidation pathway 87, 34 of them 'beta-oxidation' and a further 18 'alpha-subunit'. The effect of these replacements on precision and recall is limited. A slight increase in recall can be seen in the pantothenate and coenzyme A pathway. The precision of the analysis of the tetrahydrofolate pathway drops slightly owing to the effect of uncovering a reaction spanning two steps in the pathway. Fatty acid β-oxidation pathway results show an increase in precision in the

product-enzyme category by losing two false positives to the fact that the sentence in question now parses correctly.
The changes in precision and recall as a result of the removal of literature and figure references in Experiment C are either not there or negligible. While the number of sentences recognised by the parser decreases by two in the first two pathways and by twelve in the fatty acid β-oxidation pathway, these sentences are not relevant to metabolic reactions.
In Experiment D, the conversion of chemical reactions to syntactic constructs does have a positive effect on precision and recall. Yet the use of arrows and "+"-signs is rare, with a total of ten chemical reactions across all three pathways. They do not affect the recall of interaction pairs containing the enzyme as enzyme names are not mentioned in the reactions themselves.
In Experiment E, finally, the measured effect of replacing anaphoric references is limited again. Reasons vary. In the pantothenate pathway texts there are no anaphoric references and so no effect can be measured. The tetrahydrofolate pathway texts have 56 anaphoric replacements and 88 short forms to long form conversions, yet only 20% and 30% respectively are in sentences containing a reaction. For the fatty acid β-oxidation pathway texts the contrast is even starker with only two out of 43 anaphoric replacements occurring in sentences with a reaction.
The anaphoric and long form replacements in the tetrahydrofolate pathway lead to a drop in precision for substrate – product pairs, owing again to the uncovering of some substrate – product relations spanning more than one enzymatic reaction step. While these recovered relations still hold valid information, they are false positives in the CNSS evaluation system we also use.

BioMed Central Corpus

Analysis of the BioMed Central corpus took around 1h per 10.000 full-text papers for steps 1-7 (Figure 1). Loading of the data into the Oracle 11.2g database took approximately 9h with an additional 20h required for creation of necessary indexes and views. Extraction of the information triples from the syntactic triples took approximately 2h. An overview of the results for the Biomed Central Corpus is given in Table 2.

378014	documents
5778414	sentences
3764623	documents with at least one enzyme or metabolite
26852243	sentences with at least one enzyme or metabolite
48712	interaction triples
22674	distinct relations between entities
4532	distinct entities in at least one relation

Table 2 – Overview of results of the BioMed Central corpus

substrate-product												
	ketopantoate				tetrahydrofolate				b-oxidation			
	recall	%	prec.	%	recall	%	prec.	%	recall	%	prec.	%
CNSS	10/15	67	35/59	59	9/11	82	55/114	48	2/10	20	12/30	40
Exp. A	9/15	60	12/13	92	8/11	72	27/34	79	1/10	10	5/5	100
Exp. B	10/15	67	16/17	94	8/11	72	27/35	77	1/10	10	5/5	100
Exp. C	10/15	67	16/18	89	8/11	72	27/35	77	1/10	10	5/5	100
Exp. D	14/15	93	20/22	91	9/11	82	31/40	78	2/10	20	7/7	100
Exp. E	10/15	67	16/18	89	9/11	82	28/41	70	1/10	10	5/5	100

substrate-enzyme												
	ketopantoate				tetrahydrofolate				b-oxidation			
	recall	%	prec.	%	recall	%	prec.	%	recall	%	prec.	%
CNSS	7/12	58	13/20	65	7/11	64	28/45	62	3/8	38	8/10	80
Exp. A	6/12	50	7/8	88	6/11	54	15/15	100	1/8	13	6/8	75
Exp. B	7/13	54	9/10	90	6/11	54	15/16	93	1/8	13	6/8	75
Exp. C	7/13	54	9/10	90	6/11	54	15/16	93	1/8	13	6/8	75
Exp. D	7/13	54	9/10	90	6/11	54	15/16	93	1/8	13	6/8	75
Exp. E	7/13	54	9/10	90	7/11	64	16/18	89	1/8	13	6/8	75

product-enzyme												
	ketopantoate				tetrahydrofolate				b-oxidation			
	recall	%	prec.	%	recall	%	prec.	%	recall	%	prec.	%
CNSS	6/11	55	13/22	59	7/10	70	26/45	58	3/8	38	6/9	67
Exp. A	3/11	27	4/4	100	7/10	70	16/17	94	0/8	0	2/5	40
Exp. B	3/11	27	4/4	100	7/10	70	16/18	89	0/8	0	2/3	67
Exp. C	3/11	27	4/4	100	7/10	70	16/18	89	0/8	0	2/3	67
Exp. D	3/11	27	4/4	100	7/10	70	16/18	89	0/8	0	2/3	67
Exp. E	3/11	27	4/4	100	8/10	80	17/19	89	0/8	0	2/3	67

Table 3 - Evaluation of binary interaction pairs in the different experiments

Evaluation of the binary interaction pairs for all our experiments for the pantothenate and coenzyme A pathway (ketopantoate), tetrahydrofolate pathway (tetrahydrofolate) and aerobic fatty acid beta-oxidation pathway (b-oxidation) compared to CNSS. Highlighted in yellow is the experiment with the most change, in red the drop in precision due to an increase in recovered interactions spanning more than one reaction step.

	Exp. B	Exp. C	Exp. D	Exp. E
pantothenate and coenzyme A pathway	10	105	2	0
tetrahydrofolate pathway	7	126	3	144
fatty acid β-oxidation pathway	87	153	5	73

Table 4 – Changes to the input texts per pathway and experiment

Discussion

Evaluation of text mining systems

As others have noted before us (Ananiadou&McNaught 2006 Ch. 8; Czarnecki et al. 2012) it is difficult to find sufficiently large annotated corpora to test text mining applications and thus to come to meaningful conclusions about their performance, strength and weaknesses. We follow the solution proposed by CNSS, based on an approach by Rodrigues-Penagos (Rodriguez-Penagos et al. 2007), by investigating the capability of a text mining application to recover an already known pathway based on the papers cited as references for said pathway. This circumvents the problem of having no annotated corpus. However the significance of the individual results has to be regarded with caution as the amount of input data is still small, as is the number of reactions in the pathways to recover. While the density of relevant sentences in this corpus is suitable (e.g. pantothenate pathway 180 sentences total, 45 with reaction) all papers are known to be relevant to the investigated pathways. Thus, all conclusions are tentative.

Results compared to CNSS

In line with our initial assumption our analysis approach gains high precision results, higher than the analysis of CNSS. Also the type of false positives we recover is different and in our opinion less severe than those found with the rule-based approach. Where most of our false positives are reactions in the pathway that skip one or more steps between substrate and product, the false positives found by CNSS contain a large number of relationships that are not true, largely due to the extraction of terms that are neither enzyme nor metabolite. The occurrences that skip reaction intermediates still hold relevant information and in future attempts can be labelled explicitly as such. When processing large volumes of text, skipped reaction steps may become visible because the steps themselves are found in other passages.

The downside of a high precision approach is lower recall, which is also the case for our results. Yet the influence of lower recall of individual reactions from the text on the ability to recover the pathway in question is significantly lower than expected. Where we are only able to recover roughly half the amount of true positives compared to CNSS we manage to recover 75 – 100% of the pathway reactions also retrieved by CNSS. This is in line with the expectation that reactions of interest in

the analysed literature are mentioned multiple times. We are able to recover at least one of each and therefore are able to recover that particular step in the pathway, whereas CNSS recovers most reactions more than once.
When applying any text mining approach to very large collections of text (e.g. Medline, BioMed Central) high precision of the results, like in our approach, is preferable. It is better to have a smaller result set of a very high quality than a large set of average quality, as also noted by Lee et al. (Lee et al. 2012). High precision limits the number of results to be studied and cleared of false positives by a human user. With regards to metabolic reactions the loss of recall associated with high precision may result in very rare reactions to be missed, but when a corpus is sufficiently large reactions will be mentioned more than once and thus have a higher probability to be picked up, as is borne out by this experiment. For our approach that increase in probability is not dependent on the number of occurrences of a given reaction in the text, but dependent on the fact that the more often a reaction is mentioned the higher the probability that the reaction is described in an explicit manner. This allows us to identify rare reactions and potential uncatalogued links in pathways.

Lack of recall when including enzymes

Studying the results of CNSS and our method, it becomes apparent that it is more difficult to recover a complete reaction including the enzyme than to recover a reaction with only the participating metabolites. We identified a number of reasons for this (*see also Appendix* Table A. 12 - Table A. 14).
Firstly, as part of the named entity recognition task, recovering a complete reaction involves the recognition of at least three named entities rather than two, either enzyme or metabolites, which might not be detected in the NER task. Secondly, as a special case of the former, it is perfectly normal to describe a chemical conversion of substrate to product without actually mentioning the active enzyme in the same sentence. This does occur in a number of sentences in all three investigated pathways. In this case extraction of a complete reaction is impossible. A third cause of loss of recall is on the level of relation extraction, and occurs even when all relevant entities in a sentence have been correctly identified. Our approach has specific difficulties in identifying the relationship between enzyme and metabolite as opposed to the relationship between metabolites. The reason for this lies in the way the sentences are constructed by authors: metabolites mostly occur closely together within the sentences, often linked in simple relations like ‘conversion of metabolite A to metabolite B ’. For these constructs the extraction of the relationship is simple and requires only linkage of two syntactic triples. The distance both in words and syntactic triples between the active enzyme and the metabolites within the sentences is greater, resulting in more syntactic triples that have to be linked in order to extract the relationship. Also the variety of syntactic triples and keywords necessary to link enzyme and metabolite is greater, making it more difficult to describe generic rules for extraction. One could argue that this is specific to our approach of using syntactic

triples, but any other linguistic approach will suffer in a similar way as the linguistic degrees of freedom to describe an enzyme and its role in a reaction are much higher than the ones to describe the conversion of one metabolite into the other. A good example of this is the following sentence, which actually occurs in our corpus (with boldface highlighting relevant entities introduced by us):
"Hence, in addition to control of **CoA** synthesis on the level of **pantothenate kinase**, further modulation of flux through the pathway could occur at **phosphopantetheine adenylyltransferase** (**PPAT**), which catalyses the penultimate step in the pathway (Fig.1), the reversible adenylation of **4"-phosphopantetheine** to form **3"-dephospho-CoA** (**dPCoA**) and **pyrophosphate** (**PPi**)." (Geerlof et al. 1999)

Analysis of the experiments

The different experiments we performed to investigate the influence of the kind of input text on the performance of our approach turned out to have little effect on the results. One reason is the limited number of reactions and reaction pairs in the studied articles as well as the limited number of occurrences of each experiment's targets. For example there are only seven Greek characters in the tetrahydrofolate pathway texts and none of them is part of a sentence with an interaction.
As for Experiment E, the pantothenate pathway texts do not contain a single anaphoric reference. The tetrahydrofolate pathway corpus shows a completely different picture, where there have been a total of 144 anaphor replacements of which 56 were true anaphora and 88 short-forms to long-form conversions. Only 20% (11) of the true anaphoric relations are found in sentences describing a reaction, and indeed we find an additional nine reaction pairs. Only three of them are true positives, the others skip steps in the pathway. While these still hold relevant information, they are strictly speaking false positives. The unbalance of anaphora between sentences describing reactions and those that do not can be just indicative for this corpus, or it can be an indication that authors prefer to give the enzyme name in a reaction sentence and use anaphora like 'the enzyme' when describing other content. In the fatty-acid β-oxidation pathway, generic terms are used in an anaphoric role to describe steps in the pathway (e.g. 'fatty acyl CoA'), however these forms could not be resolved to more specific terms as the antecedents are not mentioned in the text at all.
As CNSS already note, the quality of results for the fatty acid β-oxidation pathway is much lower than for the other two pathways and is indicative of a different style of describing reactions.
The one experiment that clearly increases recall with little or no loss of precision is the conversion of chemical formula in the text to syntactic constructs that can be parsed. The results appear limited as there are only ten chemical reactions described across all three texts, yet the increase in true positives is seven for substrate – product interaction pairs, enzymes are not mentioned in the reactions.

Effects on text format on the result

The starting format of text mining corpora does influence the results of any text mining approach to some extent, and the experiments we conducted also aimed to identify effects of the input text on the performance of the parser. The collections of papers used in the performance evaluation are all only available as PDF documents, most of them predating the digital publishing period. This not only makes it difficult to extract the correct flow of text but also to recognize and convert special characters. In the conversion to flat text, as supplied by CNSS., superscript numbers indicating references had been converted to normal numbers making them indistinguishable from correctly occurring in-line numbers. The chemical reactions breaking up the text flow similar to a table or figure had been integrated within the surrounding sentences during conversion. Automated application of the changes required for the experiments was impossible without the information implicit in the original document, therefore all but the replacement of Greek characters had to be performed manually. (CNSS appear not to have performed these conversions at all.) When using the BioMed Central corpus as input data automated application of the changes was possible. In the XML-format references to cited papers and tables and figures are tagged with '<xref>' tags allowing their simple removal. As mentioned, Greek characters can be easily replaced, in this case by exchanging the XML replacement character (e.g. *α* with *alpha*). Nevertheless, many parser systems can handle the entire Unicode range of characters or at least a large subset of them. The same approach can be taken for chemical reactions by identifying the pattern created by the XML characters and replacing it with the corresponding sentence.
Anaphora are not easily replaced on a large scale, but there are software solutions available (Castano et al. 2002; Liang&Lin 2005; Gasperin et al. 2007; Torii&Vijay-Shanker 2007) that might be included as a pre-processing step. Short forms can possibly be replaced by long forms by using abbreviation identifying tools (Schwartz&Hearst 2003; Yamaguchi et al. 2012) or picking up definitions from bracketed explanations as in "ADH (alcohol dehydrogenase)". Anaphora resolution has a tradition in the NLP community that goes back several decades. It seems that full anaphora resolution has to follow syntactic analysis and more often than not spans over sentence boundaries (Mitkov 2003).

Named entity recognition (NER)

As CNSS already note, extracting metabolic pathways from literature is a step up from protein-protein interactions as it requires identification of terms from two different categories, namely enzymes and metabolites. The performance of both our approach and that of CNSS are inexorably linked with that NER performance. CNSS use two off-the-shelf NER tools, BANNER (Leaman&Gonzalez 2008) and OSCAR3 (R. Berthold et al. 2006), for the NER task. CNSS purposefully chose a loose approach for the NER task by allowing fuzzy matching and stemmed words for the initial

selection of relevant sentences with two or more entities and by their rules to filter the relevant sentences. They also point out that more stringent conditions for NER will affect performance, making it unlikely that the CNSS method in combination with a thesaurus like ours would yield results similar to those actually reported by CNSS. Let us call the original CNSS setup "CNSS-NER" and CNSS combined with our thesauri "CNSS-thesaurus". If we were to speculate, we would expect precision and recall of CNSS-thesaurus to move closer to our approach. CNSS-thesaurus compared with CNSS-NER would lose the false positives associated with clearly not relevant entities and would lose recall based on the limitations of the coverage of the thesauri. However all potential benefits of a full grammar approach, like the identification of negations and gradations of confidence in utterings, would be lost.
We chose a thesaurus-based approach with two high-quality thesauri. These thesauri perform a double function. First they allow the correct identification of multiword term boundaries. For this task, our approach relies on exact term matches. The more loose CNSS-NER approach would not work here as AGFL does not allow wildcards in its lexicon. The other function of the thesauri is the same as that performed by BANNER and OSCAR3 in CNSS, the identification of entities relevant to the analysis. For this, precision and recall depend on the coverage and accuracy of the thesauri. For text mining in rapidly progressing domains such as molecular biology, thesauri have a weakness in that their coverage is usually limited. This is mainly due to the distribution of terms used to describe entities in text, with research showing that the term mention distribution follows Zipf's law and that up to 40% of entity terms in PubMed are unique [Rebholz]. Adding resources to a thesaurus to increase coverage is not a simple task as the granularity of the sets of synonyms of the new datasets would have to be matched to that of the existing ones. Adding data sources also increases the likelihood of adding homonyms. As already evident in the results of Hettne et al,, adding data sources does not necessarily increase performance of a given thesaurus. A partial solution for the homonym problem would be the use of specific thesauri for specific species analyses. This is not very practical, however, because the creation of thesauri is very labour-intensive. For these reasons, we chose two thesauri with coverage independent of species and pathways. Regardless of coverage and problems with synonymy and homonymy, the best thesauri would be of little use if the grammar did not perform accordingly. If the sentences were not parsed accurately, the dependency triples were not recognised and our patterns of those triples were incorrect or incomplete, neither the best thesauri nor an NER approach would rescue performance.

Named entity identification (NEI)

For our system we chose a thesaurus-based approach for the identification of the enzymes and metabolites, and we require exact string matches. This approach not surprisingly results in a loss of recall as the thesauri do not cover all used terms. However the named entity identification (NEI) (Hettne et al. 2009) approach has

an advantage as it enables us to group identified terms together based on synonym information and map them to individual enzymes and metabolites . Using these grouped sets of terms (synsets) different reaction mentions can be joined independently from the actual terms used in the text, consequently coming closer to the ultimate goal of extracting complete metabolic pathways from literature as opposed to only unconnected interaction pairs. Database identifiers, in our case enzyme numbers and KEGG compound identifiers also allow a comparison of extracted reactions with existing pathway databases helping to distinguish 'old' reactions from 'new' ones.

Can the results be generalised?

A relevant question given the limited nature of the test datasets and test pathways is whether the results can be transferred to other pathways and other organisms. All three evaluated pathways are from E. coli and from the EcoCyc database. This is on first sight rather limited. We did, however, create the thesauri from data sources that are generic in their coverage regarding organisms and pathways. Therefore, there is no reason why the NER task (including the word boundary part) would work differently for other species. The unit of text of interest for our approach is the description of a chemical conversion of metabolites catalysed by enzymes. As there are only so many ways to describe such a metabolic reaction it is not too farfetched that this way of describing the reaction is relatively independent of the studied species or pathway. We expect performance for individual reactions and pathways to vary based on the way the entities in such a reaction are named as can be already seen in the variation of performance in the three studied pathways. For example, in the fatty acid pathway texts the exact chain length of fatty acid molecules is usually not given and generic terms like "medium chain fatty acid" and "long chain fatty acid" are used instead. This makes identification of the correct reaction constituents difficult, also for a human reader. In biological research and in the resulting databases there also remains the bias towards reference species and highly studied pathways. The bias makes it more likely that highly studied reactions in well studied species are more easily and precisely identified than reactions in pathways studied less well. Overall we expect that our method transfers well to other species and pathways. The number of different reactions extracted from the BioMed Central dataset indicate that this assumption is correct.

Advantages of a full grammar approach

We see a future for relationship extraction with a full-grammar approach. The possibilities for extracting or ignoring relationships containing certain types of modifiers (e.g. 'might convert' or 'seems to be converted') and of negations (e.g. 'does not convert'), the distinction between agents within sentences (e.g. 'enzyme 1 does not convert metabolite a, but enzyme 2 does') and the conversion of passive to active sentences, in combination promise analysis results that are far superior to those of alternatives including rule-based systems. Despite the fact that the grammar

of our system has been optimised for biomedical text, it derives from a grammar for general English and can be applied to a whole range of texts.
Text mining is not an exact science and semantic and syntactic ambiguities will always pose a problem for precision and recall. These ambiguities result usually in multiple parsing options for a given sentence and it is next to impossible for any computational approach to always choose the correct semantic interpretation. The results of text mining applications like the one presented here should therefore be treated like any other *in silico* effort and their veracity independently verified.
While the high precision and good coverage of our results already make the application useful, a number of improvements can be considered to further increase the performance of our parsing approach. Other future parsing solutions should be aware of these challenges.
On the level of the grammar the fine-tuning for syntactic ambiguities of parts of speech (e.g. verb: to form, noun: the form) is not complete and certain types of syntactic constructs result in a preference of the noun form over the verb form. This is indicative for the need to further adapt the parser to the requirements of this specific sub-language describing metabolic reactions.
When sentences contain a complex array of subordinate clauses the parser tends to run out of options and therefore is not able to completely parse the sentence. When a sentence contains an anaphoric subordinate clause (such as '.., which' or '…that'), the subordinate clause is not linked to the parent agent, instead the subject is labelled as 'it'. This complicates the identification of the parent agent in the triple form of the output.
On the level of extracting triples from the parse tree, it would be worthwhile to attach a subordinate clause not only to the immediate parent, but also the parent's parent to simplify the number of triples to traverse in order to extract the metabolic relation:

'ADH catalyses the final step in the pathway, the conversion of ethanol to aldehyde.'

[N:conversion, of ethanol]	[N:conversion, of ethanol]
[N:conversion, to aldehyde]	[N:conversion, to aldehyde]
[N:step, ATTR A:final]	[N:step, ATTR A:final]
[N:step, in pathway]	[N:step, in pathway]
[V:catalyse, OBJ N:step]	[V:catalyse, OBJ N:step]
	[V:catalyse, OBJ N:conversion]
[adh, SUBJ V:catalyse]	[adh, SUBJ V:catalyse]
[ethanol, to aldehyde]	[ethanol, to aldehyde]
[pathway, PRED N:conversion]	[pathway, PRED N:conversion]

Lastly, on the level of connecting syntactic triples to extract metabolic relations, recall can be increased by increasing the number of patterns to look for and adding reaction-specific keywords. Neither added patterns nor keywords should result in a loss of precision.

Conclusions

We conclude that despite its current recall our full parsing approach to metabolic reaction extraction has high precision and potential to be used to (re-)construct metabolic pathways in an automated setting. With future improvements to the grammar and relation extraction rules, reactions can be extracted with even higher specificity.

Software

The PHASAR software, which contains AGFL 2.8 and the EP4IR grammar as well as LCS 2.4 can be downloaded from www.bioinformatics.nl/thesauri. On the same site we also supply the modified EP4IR grammar and lexicon and the adapted scripts used in this approach.

Acknowledgements

The authors would like to thank Jan Czarnecki for providing the pathway corpora. We would like to thank the late Cornelis Koster and Olaf Seibers for providing the PHASAR source code, which while published under GPL is not downloadable.
All text mining analysis tasks were performed on the SARA HPC cluster.
This research has been funded by the Netherlands Bioinformatics Centre as part of the BioRange Project SP 4.1.1.
The authors are also indebted to the late Jack Leunissen who sadly passed away before the manuscript was finished.

References

Abi-Haidar A, Kaur J, Maguitman A, Radivojac P, Retchsteiner A et al. 2008. Uncovering protein interaction in abstracts and text using a novel linear model and word proximity networks. *Genome Biology* **9**(2): S11.

Alex B, Grover C, Haddow B, Kabadjov M, Klein E et al. 2008. Automating curation using a natural language processing pipeline. *Genome Biology* **9**(Suppl 2): S10.

Ananiadou S, McNaught J. 2006. *Text mining for biology and biomedicine*. Artech House Boston, London.

Baeza-Yates R, Ribeiro-Neto B. 2011. Modern information retrieval. In *Modern information retrieval*, Vol 463. ACM press New York.

Bundschus M, Dejori M, Stetter M, Tresp V, Kriegel H-P. 2008. Extraction of

semantic biomedical relations from text using conditional random fields. *BMC Bioinformatics* **9**(1): 207.

Castano J, Zhang J, Pustejovsky J. 2002. Anaphora resolution in biomedical literature. In *Proceedings of International Symposium on Reference Resolution for NLP*, Alicante, Spain.

Cimiano P, Reyle U, Šarić J. 2005. Ontology-driven discourse analysis for information extraction. *Data & Knowledge Engineering* **55**(1): 59-83.

Croft D, O'Kelly G, Wu G, Haw R, Gillespie M et al. 2011. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Research* **39**(suppl 1): D691-D697.

Czarnecki J, Nobeli I, Smith A, Shepherd A. 2012. A text-mining system for extracting metabolic reactions from full-text articles. *BMC Bioinformatics* **13**(1): 172.

Ferreira JD, Couto FM. 2010. Semantic similarity for automatic classification of chemical compounds. *PLoS Computational Biology* **6**(9): e1000937.

Fiehn O. 2002. Metabolomics - The link between genotypes and phenotypes. *Plant Molecular Biology* **48**(1-2): 155-171.

Friedman C, Kra P, Yu H, Krauthammer M, Rzhetsky A. 2001. GENIES: A natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics* **17**(SUPPL. 1).

Gasperin C, Karamanis N, Seal R. 2007. Annotation of anaphoric relations in biomedical full-text articles using a domain-relevant scheme. In *Proceedings of DAARC*, Vol 2007. Citeseer.

Geerlof A, Lewendon A, Shaw WV. 1999. Purification and characterization of phosphopantetheine adenylyltransferase from Escherichia coli. *J Biol Chem* **274**(38): 27105-27111.

Hall RD. 2006. Plant metabolomics: From holistic hope, to hype, to hot topic. *New Phytologist* **169**(3): 453-468.

Hearst MA. 1999. Untangling text data mining. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pp. 3-10. Association for Computational Linguistics.

Hettne KM, Stierum RH, Schuemie MJ, Hendriksen PJM, Schijvenaars BJA et al. 2009. A dictionary to identify small molecules and drugs in free text. *Bioinformatics* **25**(22): 2983-2991.

Hollywood K, Brison DR, Goodacre R. 2006. Metabolomics: Current technologies and future trends. *Proteomics* **6**(17): 4716-4723.

Jenssen T-K, Laegreid A, Komorowski J, Hovig E. 2001. A literature network of human genes for high-throughput analysis of gene expression. *Nat Genet* **28**(1): 21-28.

Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M et al. 2006. From genomics to chemical genomics: new developments in KEGG. *Nucl Acids Res* **34**(suppl_1): D354-357.

Keseler IM, Collado-Vides J, Santos-Zavaleta A, Peralta-Gil M, Gama-Castro S et

al. 2011. EcoCyc: a comprehensive database of Escherichia coli biology. *Nucleic Acids Research* **39**(suppl 1): D583-D590.

Kim J-D, Ohta T, Tateisi Y, Tsujii Ji. 2003. GENIA corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics* **19**(suppl 1): i180-i182.

Kim J-j, Park JC. 2004. BioIE: retargetable information extraction and ontological annotation of biological interactions from the literature. *Journal of bioinformatics and computational biology* **2**(03): 551-568.

Klinger R, Kolářik C, Fluck J, Hofmann-Apitius M, Friedrich CM. 2008. Detection of IUPAC and IUPAC-like chemical names. *Bioinformatics* **24**(13): i268-i276.

Koster C, Seibert O, Seutter M. 2006. The PHASAR Search Engine. In *Natural Language Processing and Information Systems*, Vol 3999, pp. 141-152-152. Springer Berlin / Heidelberg.

Koster C, Seutter M, Seibert O. 2007. Parsing the medline corpus. In *Proceedings RANLP*, Vol 2007, pp. 325-329.

Koster CA, Seutter M. 2003. Taming Wild Phrases. In *Advances in Information Retrieval*, Vol 2633 (ed. F Sebastiani), pp. 161-176. Springer Berlin Heidelberg.

Koster CHA, Verbruggen E. 2002. The AGFL Grammar Work Lab. In *Proceedings of the FREENIX Track: 2002 USENIX Annual Technical Conference*. USENIX Association.

Krallinger M, Leitner F, Rodriguez-Penagos C, Valencia A. 2008. Overview of the protein-protein interaction annotation extraction task of BioCreative II. *Genome Biology* **9**(Suppl 2): S4.

Leaman R, Gonzalez G. 2008. BANNER: an executable survey of advances in biomedical named entity recognition. In *Pacific Symposium on Biocomputing*, Vol 13, pp. 652-663.

Lee J, Kim S, Lee S, Lee K, Kang J. 2012. High precision rule based PPI extraction and per-pair basis performance evaluation. In *Proceedings of the ACM sixth international workshop on Data and text mining in biomedical informatics*, pp. 69-76. ACM, Maui, Hawaii, USA.

Liang T, Lin YH. 2005. Anaphora resolution for biomedical literature by exploiting multiple resources. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol 3651 LNAI, pp. 742-753.

Lindberg DA, Humphreys BL, McCray AT. 1993. The Unified Medical Language System. *Methods Inf Med* **32**(4): 281-291.

Ma H, Sorokin A, Mazein A, Selkov A, Selkov E et al. 2007. The Edinburgh human metabolic network reconstruction and its functional analysis. *Mol Syst Biol* **3**.

McCray AT, Srinivasan S, Browne AC. 1994. Lexical methods for managing variation in biomedical terminologies. *Proc Annu Symp Comput Appl Med Care*: 235-239.

Mitkov R. 2003. Anaphora resolution. In *The Oxford Handbook of Computational Linguistics*, (ed. R Mitkov). Oxford University Press.

Miyao Y, Sagae K, Sætre R, Matsuzaki T, Tsujii Ji. 2009. Evaluating contributions of natural language parsers to protein–protein interaction extraction. *Bioinformatics* **25**(3): 394-400.

Medline [http://www.nlm.nih.gov/pubs/factsheets/medline.html]

Nobata C, Dobson P, Iqbal S, Mendes P, Tsujii Ji et al. 2011. Mining metabolites: extracting the yeast metabolome from the literature. *Metabolomics* **7**(1): 94-101-101.

Novichkova S, Egorov S, Daraselia N. 2003. MedScan, a natural language processing engine for MEDLINE abstracts. *Bioinformatics* **19**(13): 1699-1706.

Pico AR, Kelder T, van Iersel MP, Hanspers K, Conklin BR et al. 2008. WikiPathways: pathway editing for the people. *PLoS biology* **6**(7): e184.

Pyysalo S, Ginter F, Pahikkala T, Boberg J, J\ J et al. 2004. Analysis of link grammar on biomedical dependency corpus targeted at protein-protein interactions. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pp. 15-21. Association for Computational Linguistics, Geneva, Switzerland.

R. Berthold M, Glen R, Fischer I, Corbett P, Murray-Rust P. 2006. High-Throughput Identification of Chemistry in Life Science Texts. In *Computational Life Sciences II*, Vol 4216, pp. 107-118-118. Springer Berlin / Heidelberg.

Rinaldi F, Schneider G, Kaljurand K, Hess M, Andronis C et al. 2007. Mining of relations between proteins over biomedical scientific literature using a deep-linguistic approach. *Artificial Intelligence in Medicine* **39**(2): 127-136.

Rindflesch TC, Tanabe L, Weinstein JN, Hunter L. 2000. EDGAR: extraction of drugs, genes and relations from the biomedical literature. In *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing*, p. 517. NIH Public Access.

Rodriguez-Penagos C, Salgado H, Martinez-Flores I, Collado-Vides J. 2007. Automatic reconstruction of a bacterial regulatory network using Natural Language Processing. *BMC Bioinformatics* **8**(1): 293.

Sayle R, Xie PH, Muresan S. 2011. Improved chemical text mining of patents with infinite dictionaries and automatic spelling correction. *Journal of chemical information and modeling* **52**(1): 51-62.

Schwartz AS, Hearst MA. 2003. A simple algorithm for identifying abbreviation definitions in biomedical text. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing*: 451-462.

Settles B. 2005. ABNER: An open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics* **21**(14): 3191-3192.

Sharma A, Swaminathan R, Yang H. 2010. A verb-centric approach for relationship extraction in biomedical text. In *Semantic Computing (ICSC), 2010 IEEE Fourth International Conference on*, pp. 377-385. IEEE.

Stobbe M, Houten S, Jansen G, van Kampen A, Moerland P. 2011. Critical

assessment of human metabolic pathway databases: a stepping stone for future integration. *BMC Systems Biology* **5**(1): 165.

Thiele I, Swainston N, Fleming RMT, Hoppe A, Sahoo S et al. 2013. A community-driven global reconstruction of human metabolism. *Nat Biotech* **31**(5): 419-425.

Torii M, Vijay-Shanker K. 2007. Sortal Anaphora Resolution in Medline Abstracts. *Computational Intelligence* **23**(1): 15-27.

Van Landeghem S, Hakala K, Rönnqvist S, Salakoski T, Van de Peer Y et al. 2012. Exploring Biomolecular Literature with EVEX: Connecting Genes through Events, Homology, and Indirect Associations. *Advances in Bioinformatics* **2012**: 12.

Wang R, Siu SW, Bockmann R. 2009. Fine-Grained Protein Mutation Extraction from Biological Literature. In *Electronic Computer Technology, 2009 International Conference on*, pp. 401-405. IEEE.

Yamaguchi A, Yamamoto Y, Kim J-D, Takagi T, Yonezawa A. 2012. Discriminative application of string similarity methods to chemical and non-chemical names for biomedical abbreviation clustering. *BMC Genomics* **13**(Suppl 3): S8.

Zweigenbaum P, Demner-Fushman D, Yu H, Cohen KB. 2007. Frontiers of biomedical text mining: current progress. *Briefings in Bioinformatics* **8**(5): 358-375.

Appendix

Syntax relations for dependency triples

SUBJ	PRED	IT	INVSUBJ
OBJ	PREP	YOU	INVOBJ
VERB	MOD	DET	INVPRED
ATTR	THAT	INTER	INVof
	SUCHTHAT	QUANT	INVto
	BE	QUESTION	INVOBJ

Table A. 1 – Complete List of syntactic relations for dependency triples. INV relations are normalised in passive – active conversion.

Grammar Adaptations

line number	statement changed or added
49	Wsub, statement, Wbus / statement ## changed "Wsub, statement, Wclose" to "Wsub, statement, Wbus" as Wbus is the correct corresponding closing tag.
58	$PENALTY(4), NP / "[", NP, "]" ## added to accommodate parsing of NP only.
316	noun phrase(plur,PERS,CASE): noun(sing), Wand, noun(sing),BONUS,BONUS / "[",noun.1,"|",noun.2,"]" ; LEX_PAIRDET(TEXT), noun phrase(NUMB, PERS, CASE), [pairdet and noun phrase(PERS, CASE)] / "[", noun phrase.1, "|", [], "]" ## changed to accommodate phrases of the form 'enzyme 1 and enzyme 2 convert'
334	pairdet and noun phrase(PERS, CASE): LEX_PAIRDET(TEXT), noun phrase(NUMB1, PERS, CASE), BONUS,BONUS / noun phrase . ## generic noun phrase placed before plural to get coordinators right

428	LEX_NOUNX(NUMB,metabol),PP(of|from|by|to|in|into|with|between), BONUS, BONUS, BONUS / "[", LEX_NOUNX, ",", PP, "]" ## LEX_NOUNX(NUMB,metabol) added to account for constructions like "synthesis of a and b".
512	Wcomma, predicate, Wperiod / predicate ## added to account for parenthesis at the end of a sentence.
807	OC phrase(trav,none,PREP): object, pref PP(XPREP), Wto, inf phrase(PREP), BONUS, BONUS, BONUS / object, pref PP, "]|SUBJ [", inf phrase ## Added to include sentences containing 'to form', 'to produce', 'to yield': ## The dihydrofolate synthetase adds L-glutamate to dihydropteroate to form dihydrofolate.

Table A. 2 – Exact adaptions to sentence.gra in EP4IR grammar as supplied in PHASAR

line number	statement changed or added
68	; statement ## Added statement to segment options as otherwise a segment cannot consist of partial phrases more complex than NP.

Table A. 3 – Exact adaptions to npx.gra in EP4IR grammar as supplied in PHASAR

line number	statement changed or added
105	NOUNC(NUMB), BONUS ; NOUNT(TRAD, NUMB) , BONUS / TRAD ; NOUN(NUMB) ; ## changed scoring to prefer compound noun over constituent

Table A. 4 – Exact adaptions to interface.gra in EP4IR grammar as supplied in PHASAR

line number	statement changed or added
17	CPLX :: cplx | fincomp | ascomp | infcomp | metabol . # | whinfcomp . ## added to accommodate metabol form of complex type (see sentence.gra line 428)

Table A. 5 – Exact adaptions to meta.gra in EP4IR grammar as supplied in PHASAR

Syntax patterns

enzyme stands for EC number, 'enzyme' represents the word enzyme
compound stands for compound identifier
any stands for any possible term
verb is any verb from the verb list that is not indicative of a negative relation
attr is any term from the attribute list
subst is any term from the substantiations list
syntactic relations are in all caps, ANY stands for any of the accepted syntactic relations
identical identifiers in two or more triples resemble identical terms

enzyme,SUBJ,verb verb,OBJ,compound	adh converts ethanol
enzyme,SUBJ,verb verb,PREP,compound	adh produces [aldehyde] from ethanol
enzyme,ANY,any any,ANY,subst subst,PREP,compound	adh catalyses the conversion of ethanol
enzyme,SUBJ,verb.1 verb.1,PRED,verb.2 verb.2,OBJ,compound enzyme,SUBJ,any any,ANY,subst/attr subst/attr, ATTR,compound	PPC-DC catalyses [the decarboxylation of PPanCys] to form 4'-phosphopantetheine. Catalase inhibited hydrogen peroxide production
enzyme,SUBJ,verb.1 verb.1,PRED,verb.2 verb.2,PREP,compound	5-LOX catalyzes [the oxygenation of arachidonic acid] to convert [5-HPTE] to 5-HETE
enzyme,SUBJ,any any,PREP,subst subst, PREP,compound	Acetohydroxy acid isomeroreductase requires Mg2+ for the reduction of ketopantoate.
subst,PREP,by.enzyme subst.PREP,of.compound	conversion of ethanol by adh
'enzyme',PRED,enzyme any of the above	the first enzyme, called adh, …..
'*ase',PRED,enzyme any of the above	a dehydrogenase, here adh, …..

Table A. 6 – Enzyme – Compound relations
In square brackets the parts of the sentence that are not relevant to the query.

compound,SUBJ,verb verb,OBJ,compound	[the reversible adenylation of] 4'-phosphopantetheine yielding pyrophosphate.
verb,OBJ,compound verb,PREP,compound	combine ethanol with NAD+
subst,PREP,of.compound subst,PREP,to/with.compound	conversion of ethanol to aldehyde
verb,OBJ,attr verb,PREP,from.compound attr,ATTR,compound	[Both possess a] 4'-phosphopantetheine moiety [that] is derived from pantothenate.
subst,SUBJ,verb subst,PREP,of.compound verb,OBJ,compound	Pantothenate is biosynthesised by the condensation of D-pantoate and b-alanine.
compound,SUBJ,verb verb,PREP,attr attr,ATTR,compound	[Metabolizing] glycerol [to pyruvate] generates[two ATP per six carbon atoms in contrast to one ATP gained using glucose] as carbon source.
compound,SUBJ,verb verb,OBJ,subst subst,PREP,compound	…histamine derived from conversion of histamine
compound,PRED,'precursor' 'precursor',PREP,of.compound	aspartate is a precursor of B-alanine

Table A. 7 – Compound – Compound relations

verb	category	verb	category	verb	category
catabolise	1	accumulate	2	generate	2
catabolize	1	act_on	2	incorporate	2
catalyse	1	activate	2	increase	2
catalyze	1	alter	2	interconvert	2
cleave	1	biosynthesise	2	ionize	2
cometabolise	1	chain	2	link	2
cometabolize	1	change	2	make	2
complex	1	cleave	2	phosphorylate	2
convert	1	combine	2	require	2
dehydrate	1	condense	2	supply	2
deprotonate	1	construct	2	yield	2
form	1	consume	2	decrease	3
hydrolise	1	co-transmit	2	inactivate	3
hydrolize	1	couple	2	inhibit	3
metabolise	1	dearomatize	2	lessen	3
metabolize	1	degrade	2	bind	4
oxidize	1	dehydroxylate	2	react	4
produce	1	derive	2	show	4
protonate	1	dimerize	2	use	4
synthesise	1	divide	2	interconvert	6
synthesize	1	free	2		

Table A. 8 – Verbs

List of verbs indicative of metabolic interactions: category 1 are often occurring with a metabolic interaction, category 2 are more rarely, category 3 verbs indicate negative relation, category 4 verbs are only allowed in certain enzyme-compound relations and category 6 is added in chemical equilibrium reactions when converted from symbols to sentences.

noun	category	noun	category	noun	category
activation	1	de_novo_synthesis	1	oxidocyclization	1
addition	1	decarboxylation	1	oxygenation	1
adenylation	1	epoxidation	1	phosphorylation	1
biosynthesis	1	fission	1	reaction	1
cleavage	1	formation	1	reduction	1
condensation	1	hydrolysis	1	synthesis	1
conversion	1	interconversion	1	transformation	1
cyclization	1	oxidation	1	inhibition	2

Table A. 9– Nouns

List of nouns indicative of metabolic interactions mainly found in predicative or prepositional syntactic relation: category 1 are positive, category 2 indicates a negative relation.

attribute	category	attribute	category	attribute	category
accumulation	1	intermediate	1	protein	1
biosynthesis	1	level	1	residue	1
donor	1	moiety	1	source	1
enzyme	1	precursor	1	substrate	1
equivalent	1	product	1	unit	1
formation	1	production	1		

Table A. 10– Attributes

List of nouns indicative of metabolic interactions in an attributive syntactic relationship.

preposition	category	preposition	category	preposition	category
into	1	from	3	as	6
to	1	between	4	with	7
by	2	for	5	of	8

Table A. 11 – Prepositions

List of prepositions indicative of metabolic interactions: Category numbers are to allow for source and target differentiation, category 2 is only allowed for enzyme – compound relations, category 5 is not currently in use and category 6 is only used once in a compound – compound relation.

Overview of metabolic relations per sentence

Table A. 12 - Detailed per sentence view of performance and evaluation of identification of metabolic reaction references in the pantothenate and coenzyme A pathway corpus

DOCID	SENTNR	ENZYME	SOURCE	TARGET	enzyme identified	source identified	target identified	enzyme-source	enzyme-target	source-target	annotation	found and correct	not in thes	too generic	syntax error	syntax not queried	should be in result	chemical formula
10480925	5	2.7.7.3	c01134		y	y		n			connected via step...					1		
10480925	5	2.7.7.3		c00882			y		n	y	connected via step...					1		
10480925	5	2.7.7.3		c00013			y		n	y	connected via step...					1		
10480925	21	2.7.1.33	c01134		y	y		n			connected via step...				1	1		
10480925	21	2.7.1.33		c00882			y		n	n	connected via step...				1	1		
10480925	21	2.7.1.33		c00013			y		n	n	and goes wrong				1	1		
10480925	22	3.1.4.1	c00010		y	y		y				1						
10480925	22	3.1.4.11	c00010		y							1						
10480925	22	3.1.4.35	c00010		y							1						
10480925	22	3.1.4.52	c00010		y							1						
10480925	22	3.1.13.3	c00010		y							1						
10480925	25	2.7.1.24	c00882		y	y		n					1		1	1		
10480925	34	2.7.7.3		c01134	y		y		n		'catalyze formation'missing					1		
10480925	34	2.7.7.3		c00002			y		n		'catalyze formation'missing					1		
6796563	5		c00864	c01134		y	y			y		1						
6796563	16		ACP thioester	c00162		n	y			n	thioester of..		1			1		
6796563	16		CoA thioester	c00162		n				n	thioester of..		1			1		
6796563	19	2.7.1.33	c00864		y	y		y		y		1						
6796563	19	2.7.1.33		c03492			y		y			1						
6796563	20		c01134	c03492		y	y			n						1		
6796563	21		c00002	c00020		y	y			n	to form					1		
6796563	21		c00020	c00882		y	y			n						1		
6796563	21		c00882	c00010		y	y			n						1		
6796563	22		c00010	c01134		y	y			n	serve not in verb list					1		
6796563	23		c00522	c00864		y	y			n	via nor in preplist			1				
6796563	23		c00099	c00864		y				n					1	1		
6796563	24	4.1.1.11		c00099	y		y		n						1	1		
6796563	24	hydroxymethyltransferase	a-ketoisovaleric acid		n	y		n		n			1		1	1		

DOCID	SENTNR	ENZYME	SOURCE	TARGET	enzyme identified	source identified	target identified	enzyme-source	enzyme-target	source-target	annotation	found and correct	not in thes	too generic	syntax error	syntax not queried	should be in result	chemical formula
11292795	5	dephosphocoenzyme A kinase	c00121		n	y		n			enzyme name split				1			
11292795	16		c01134	c00882		y	y			n						1		
11292795	16		c00002	c00882		y				n						1		
11292795	16		c00882	c00010		y	y			n					1			
11278255	5	6.3.2.5	c03492		y	y		y		y		1						
11278255	5	6.3.2.5	c00097			y		y		y		1						
11278255	5	6.3.2.5		c04352			y		y			1						
11278255	6	4.1.1.36	c04352		y	y		n		6						1		
11278255	6	4.1.1.36		c01134			y		n							1		
11278255	9	6.3.2.5	c00063		y	y		n								1		
11278255	17	4.1.1.36	c04352		y	y		y		y		1						
11278255	17	4.1.1.36		c01134			y		y			1						
11278255	23	EpiD	peptidylcysteine		n	n		n					1		1	1		
11278255	24	protein encoded by dfp	c03492		n	y		n		n						1		
11278255	24	protein encoded by dfp	c00097			y		n		n						1		
11278255	24	protein encoded by dfp		PPanCys			n		n				1			1		
11278255	24	protein encoded by dfp	c00063			y		n			using nog in preplist					1		
776976	6		methylenetetrahydrofolate	tetrahydrofolate		n	n			n								1
776976	6		methylenetetrahydrofolate	c00966			n			n								1
776976	6		c00141	tetrahydrofolate		n				n								1
776976	6		c00141	c00966						n								1
776976	20	2.1.2.11	5,10-Methylenetetrahydrofolate		y	n		n										1
776976	20	2.1.2.11	HC(CH3)2COCOO-			n		n					1					1
776976	20	2.1.2.11		c00101			y		n									1
776976	20	2.1.2.11		HOCH2C(CH3)2COCOO-			n		n				1					1
16990935	5	6.3.2.1	c00522		y	y		y		y		1						
16990935	5	6.3.2.1	c00099			y		y		y		1						
16990935	10		c00864	c01134		y	y			n	syntax not queried					1		
16990935	14	6.3.2.1		c00864	y		y			n						1		
16990935	15		c00522	c00864		y	y			y		1						

DOCID	SENTNR	ENZYME	SOURCE	TARGET	enzyme identified	source identified	target identified	enzyme-source	enzyme-target	source-target	annotation	found and correct	not in thes	too generic	syntax error	syntax not queried	should be in result	chemical formula
16990935	15		c00099	c00864		y				y		1						
16990935	16		c00002	c00020		y	y			n					1	1		
16990935	16		c00002	c00013			y			n					1	1		
381298	3	4.1.1.11		c00099	y		y		n							1		
381298	5	l-aspartate-a-decarboxylase		c00099	n		y		n				1					
381298	15		c00099	c00864		y	y			n	"use as substrate"				1	1		
381298	16		c00864	c00010		y	y			n					1	1		
381298	19		c00163	c00222		y	y			y		1						
381298	20		c00099	c00864		y	y			n						1		
381298	20		c00099	c00010			y			n	use not in verbs					1		
381298	20		c00049	c00099		y	y			n	o-decarboxylation not in subst.					1		
381298	21		c00049	c00099		y	y			y		1						
381298	23		c00049	c00099		y	y			n	and goes wrong				1			
391298	27		c00099	c00864		y	y			n						1		
391298	27		c00099	c00010			y			n	need not in verbs					1		
10736170	5	1.1.1.169	β-ketopantoate		y	n		n		n	greek				1			
10736170	5	1.1.1.169		D-(-)-pantoate			n		n		not in thes?		1		1	1		
10736170	17	1.1.1.169	β-ketopantoate		y	n		n		n	greek		1					
10736170	17	1.1.1.169		c00522			y		n		syntax, 2 sentences		1			1		
10736170	17		c00522	c00864		y	y			n	to form				1			
10736170	17		β-alanine	c00864		n				n	greek				1			
10736170	18		c00864	c00010		y	y			n					1	1		
10736170	19	1.1.1.86	β-ketopantoate		y	n		n		n	greek		1		1	1		
10736170	19	1.1.1.86	β-aceto-β-hydroxybutyrate			n		n		n	greek		1		1	1		
10736170	19	1.1.1.86	β-acetolactate			n		n		n	greek		1		1	1		
10736170	19	1.1.1.86		β,β-dihydroxy-β-methylvalerate			n		n		greek		1		1	1		
10736170	19	1.1.1.86		β,β-dihydroxyisovalerate			n		n		greek		1		1	1		
10736170	20		c00966	c00522		y	y			n	semantics				1			
10736170	21	1.1.1.86	c00305		y	y		y				1						
10736170	21	1.1.1.86	c00966			y		y				1						

DOCID	SENTNR	ENZYME	SOURCE	TARGET	enzyme identified	source identified	target identified	enzyme-source	enzyme-target	source-target	annotation	found and correct	not in thes	too generic	syntax error	syntax not queried	should be in result	chemical formula
10736170	22	1.1.1.86	β-ketopantoate	c00522		n	y			n	greek				1	1		
10736170	22		β-ketopantoate	NADPH			n			n	greek				1	1		

Table A. 12 - Detailed per sentence view of performance and evaluation of identification of metabolic reaction references in the tetrahydrofolate pathway corpus

DOCID	SENTNR	ENZYME	SOURCE	TARGET	enzyme identified	source identified	target identified	enzyme-source	enzyme-target	source-target	annotation	found and correct	not in thes	too generic	syntax error	syntax not queried	should be in result	chemical formula
1522070	13	EC.2.5.1.15:	compound_c00568		y	y		n		n	term not in thes, messes up syntax		1		1			
1522070	13	EC.2.5.1.15:	7,8-dihydro-6-hydroxymethylpterin-pyrophosphate			n		n		n			1		1			
1522070	13	EC.2.5.1.15:		7,8-dihydropteroate			n		n				1		1			
1522070	15	EC.2.5.1.15:	sulfonamide		y	n		n						1				
1522070	15	EC.2.5.1.15:	pAB			n		n					1		1			
17698004	6	Nudix NTP hydrolase	dihydroneopterin triphosphate			n		n		n			1			1		
17698004	6	Nudix NTP hydrolase		compound_c00013	n		y		n				1			1		
17698004	20		dihydroneopterin triphosphate	dihydroneopterin monophosphate		n	n			n			1			1		
17698004	20			compound_c00013			y			n			1			1		
17698004	27		compound_c00131	compound_c00360		y	y										1	
17698004	27			compound_c00013			y										1	
14617668	13	EC.6.3.2.12	H2-folate	H4-folate	y	n	n	n	n	n			1			1		
14617668	13	EC.1.5.1.3	H2-folate	H4-folate	y			n	n				1			1		
14617668	14	EC.6.3.2.12		H4-folate	y								1					
14617668	14	EC.1.5.1.3		H4-folate	y								1			1		
14617668	14	EC.6.3.2.12		H2-folate									1			1		
14617668	14	EC.1.5.1.3		H2-folate									1			1		
14617668	14	EC.2.1.1.45	methylene-H4-folate		y	n		n					1			1		
14617668	14	EC.2.1.1.46	H4-folate			n		n					1			1		
14617668	14	EC.2.1.1.46		compound_c00365			y		n				1			1		
14617668	19	EC.1.5.1.34	H2-folate		y	n		n					1		1	1		
14617668	19	EC.1.5.1.34		H4-folate			n		n				1		1	1		
14617668	23	EC.1.5.1.33	folates		y	n		n					1					
14617668	23	EC.1.5.1.33		H4-folate			n		n				1					
14617668	24	EC.1.5.1.33	compound_c00268		y	y		n								1		
14617668	24	EC.1.5.1.33	compound_c00005			y		n								1		
2071583	5		compound_c00251	compound_c00568		y	y			n					1	1		
2071583	7		compound_c00251	compound_c11355		y	y			n	to form				1	1		

DOCID	SENTNR	ENZYME	SOURCE	TARGET	enzyme identified	source identified	target identified	enzyme-source	enzyme-target	source-target	annotation	found and correct	not in thes	too generic	syntax error	syntax not queried	should be in result	chemical formula
2071583	27	EC.4.1.3.27		compound_c00108	y		y		n						1	1		
2071583	30		chorismate	compound_c00568		n	y			n					1	1		
2071583	32		compound_c00251	compound_c00568		y	y			n							1	
2071583	33	EC.2.6.1.85	compound_c00252		y	y		y		y		1						
2071583	33	EC.2.6.1.85	compound_c00064			y		y		y		1			1			
2071583	33	EC.2.6.1.85		compound_c00568			y		y			1						
2071583	34		compound_c00251	compound_c00108		y	y			n							1	
2071583	37		compound_c00251	compound_c00568		y	y			n					1			
2071583	37			pyruvate			y			n								
2071583	37		compound_c00064	compound_c00568		y				n					1			
2071583	37			pyruvate						n	and?				1			
1644759	5		compound_c00251	compound_c00568		y	y			n							1	
1644759	5		compound_c00064			y				n							1	
1644759	6	EC.2.6.1.82	compound_c00251		y	y		n		n							1	
1644759	6	EC.2.6.1.82	compound_c00064			y		n		n							1	
1644759	6	EC.2.6.1.82		compound_c11355			y		n								1	
1644759	6	EC.2.6.1.82		compound_c00025			y		n								1	
1644759	7	EC.4.1.3.38	compound_c11355		y	y		y			syntax?				1			
1644759	7	EC.4.1.3.38		compound_c00568			y		n	n					1			
1644759	7	EC.4.1.3.48		compound_c00022			y		y	n		1						
1644759	21	EC.2.6.1.85	compound_c00251		y	y		n			commit not in verbs				1			
1644759	21	EC.2.6.1.85		compound_c00108			y		n	n					1			
1644759	21	EC.2.6.1.85		compound_c11355			y		n	n					1			
1644759	21	EC.4.1.3.27	compound_c00251		y	y		n							1			
1644759	21	EC.4.1.3.27		compound_c00108			y		n	n					1			
1644759	21	EC.4.1.3.27		compound_c11355			y		n	n					1			
1644759	24	EC.5.4.4.2	compound_c00251		y	y		y				1						
1644759	24	EC.5.4.4.2		compound_c00885			y		y			1						
1644759	26		compound_c00251	compound_c00568		y	y			n							1	
1644759	26		compound_c00064			y				n							1	

DOCID	SENTNR	ENZYME	SOURCE	TARGET	enzyme identified	source identified	target identified	enzyme-source	enzyme-target	source-target	annotation	found and correct	not in thes	too generic	syntax error	syntax not queried	should be in result	chemical formula
1644759	27	EC.2.6.1.85	compound_c00251		y	y		n		n					1		1	
1644759	27	EC.2.6.1.85	compound_c00064			y		n		n					1		1	
1644759	27	EC.2.6.1.85		compound_c11355			y		n						1		1	
1644759	27	pabA	compound_c00251		n			n			should be in now				1		1	
1644759	27	pabA	compound_c00064					n							1		1	
1644759	27	pabA		compound_c11355					n						1		1	
1644759	28	EC.4.1.3.38	compound_c11355		y	y		y							1			
1644759	28	EC.4.1.3.38		compound_c00568			y		n						1			
1644759	28	EC.4.1.3.38		compound_c00022			y		y						1			
9651328	5	EC.4.1.2.25	7,8-dihydroneopterin		y	n		n					1			1		
9651328	5	EC.4.1.2.25		6-hydroxymethyl-7,8-dihydropterin		n			n				1			1		
9651328	10	EC.4.1.2.13	L-threo-dihydroneopterin		y	n		n					1					
9651328	10	EC.4.1.2.13	D-erythro-dihydroneopterin			n		n					1					
9651328	10	EC.4.1.2.13		6-hydroxymethyldihydropterin			n		n				1			1		
9651328	11	epimerase	dihydromonapterin		n	n		n					1			1		
9651328	11	epimerase	compound_c04874			y		n					1			1		
9651328	28	phosphatase		compound_c04874	n	y			n					1				
9651328	28	EC.3.6.1.1		compound_c04874	y				y			1						
9651328	30	EC.4.1.2.25		6-hydroxymethyl-7,8-dihydropterin	y		n		n				1					
9651328	31											1						
9651328	32		dihydroneopterin_triphosphate	compound_c00504		n	y			n			1		1	1		
1939056	17	EC.6.3.2.12	compound_c00025		y	y		n			to form				1	1		
1939056	17	EC.6.3.2.12	compound_c00921			y		n							1	1		
1939056	17	EC.6.3.2.12		compound_c00415			y		n						1	1		
1939056	18	EC.1.5.1.3	compound_c00415		y	y		y				1						
1939056	18	EC.1.5.1.3		compound_c00101			y		y			1						
1939056	18	EC.6.3.2.17	compound_c00025		y	y		n					1		1	1		
1939056	18	EC.6.3.2.17		polyglutamates			n		n					1				
1939056	19		glutamates	triglutamate		n	n			n				1				
1939056	20	EC.6.3.2.17	compound_c00101		y	y		n									1	

DOCID	SENTNR	ENZYME	SOURCE	TARGET	enzyme identified	source identified	target identified	enzyme-source	enzyme-target	source-target	annotation	found and correct	not in thes	too generic	syntax error	syntax not queried	should be in result	chemical formula
1939056	20	EC.6.3.2.17	10-formyltetrahydrofolate monoglutamate			y		n					1		1			
1939056	20	EC.6.3.2.17	5,10-methyleneTetrahydrofolate diglutamate			n		n			typo		1		1			
1939056	21	EC.6.3.2.17		polyglutamate	y		n		n					1				
15705579	5	EC.6.3.2.12	compound_c00025		y	y		n							1	1		
15705579	5	EC.6.3.2.12	compound_c00921			y		n							1	1		
15705579	5	EC.6.3.2.17	compound_c00025		y	y		n							1	1		
15705579	5	EC.6.3.2.17	compound_c00101			y		n							1	1		
15705579	5	folc	compound_c00025		n			n					1		1	1		
15705579	5	folc	compound_c00921					n					1		1	1		
15705579	5	folc	compound_c00101					n					1		1	1		
15705579	17	EC.6.3.2.12	compound_c00025		y	y		n			to form				1	1		
15705579	17	EC.6.3.2.12	compound_c00921			y		n							1	1		
15705579	17	EC.6.3.2.12		compound_c00415			y		n		to form				1	1		
15705579	18	EC.1.5.1.3	compound_c00415		y	y		y				1						
15705579	18	EC.1.5.1.3		compound_c00101			y		y			1						
15705579	18	EC.6.3.2.17	compound_c00025		y	y		n			wrong compounds assigned to reaction				1	1		
15705579	18	EC.6.3.2.17		polyglutamates			n		n				1		1	1		
15705579	40	EC.6.3.2.9	compound_c00217		y	y		n			syntax not really queried						1	
15705579	40	EC.6.3.2.9	compound_c01212			y		n			identified as target						1	
15705579	40	EC.6.3.2.9	compound_c00002			y		n						1				
4304228	6	EC.2.7.6.3	H2-pteridine-CH2OH		y	n		n					1		1	1		
4304228	6	EC.2.7.6.3	compound_c00002			y		n							1	1		
4304228	6	EC.2.7.6.3		pyrophosphate ester of H2-pteridine-CH2OH			n		n				1		1	1		
4304228	10	H2-pteroate synthetase	compound_c00568		n	y		n			semantics		1					
4304228	10	H2-pteroate synthetase	pyrophosphate ester of H2-pteridine-CH2OH			n		n					1			1		
4304228	10	H2-pteroate synthetase		H2-pteroate			n		n				1					
4304228	14		H2-pteridine-CH2O-PP	H2-pteroate		n	n			n			1		1			
4304228	19		H2-pteridine-CH2OH	H2-pteroate		n	n			n			1					
4304228	19		AB			n				n			1					

DOCID	SENTNR	ENZYME	SOURCE	TARGET	enzyme identified	source identified	target identified	enzyme-source	enzyme-target	source-target	annotation	found and correct	not in thes	too generic	syntax error	syntax not queried	should be in result	chemical formula
1459137	5	EC.3.5.4.16		dihydroneopterin_triphosphate	y								1		1			
1459137	24	EC.3.5.4.16	compound_c00044		y	y		y				1						
1459137	24	EC.3.5.4.16		dihydroneopterin_triphosphate			n		n				1					
1459137	34	EC.3.5.4.16		dihydroneopterin_triphosphate	y		n		n				1			1		
4362677	5		2-amino-4-hydroxy-6- (D-erythro-1',2',3'-trihydroxypropyl)-7,8- dihydropteridine	compound_c00013		n	y			n			1		1			
4362677	9		H2-pterin-CH2OH	H2-pteroic acid		n	n			n			1		1	1		
4362677	9			H2-folic acid			n			n			1		1	1		
4362677	9		compound_c00044	H2-pterin-CH2OH		y				n			1		1	1		
4362677	10		gtp	H2-neopterin-PPP		n	n			n								1
4362677	10			compound_c00058			y			n								1
4362677	11		H2-neopterin-PPP	H2-neopterin						n								1
4362677	12		H2-neopterin	compound_c00266		n	y			n								1
4362677	12		H2-pterin-CH2OH				n			n								1
4362677	15	EC.4.1.2.13	H2-neopterin		y	n		n					1					
4362677	15	EC.4.1.2.13	H2-neopterin-PPP			n		n					1					
4362677	15	EC.4.1.2.13	H2-neopterin-P			n		n					1					
4362677	16	phosphatase	H2-neopterin-PPP		n	n		n					1		1	1		
4362677	16	phosphatase		H2-pteroic acid			n		n				1		1	1		
4362677	16	phosphatase		H2-folic acid			n		n				1		1	1		
4362677	20	H2-neopterin-PPP pyrophosphohydrolase	H2-neopterin-PPP		n	n		n					1		1			
4362677	20	H2-neopterin-PPP pyrophosphohydrolase		compound_c00013			y		n				1		1			
4362677	20	H2-neopterin-PPP pyrophosphohydrolase		H2-neopterin-P			n						1		1			
4362677	22	phosphatase	H2-neopterin-P		n	n		n			very complex, syntax not queried			1				
4362677	22	phosphatase		H2-neopterin			n		n					1				
1325970	21	EC.2.5.1.15	compound_c00568		y	y		y				1						
1325970	21	EC.2.5.1.15	7,8-dihydro-6-hydroxymethylpterin-pyrophosphate			n		n			and		1		1			

DOCID	SENTNR	ENZYME	SOURCE	TARGET	enzyme identified	source identified	target identified	enzyme-source	enzyme-target	source-target	annotation	found and correct	not in thes	too generic	syntax error	syntax not queried	should be in result	chemical formula
1325970	21	EC.2.5.1.15		compound_c00921			y		n		to form		1		1			
1325970	23	EC.2.7.6.3		7,8-dihydro-6-hydroxymethylpterin-pyrophosphate	y		n		n				1					
2251281	6	EC.2.6.1.85	compound_c00014	compound_c00568	y	y	y	n	n	6	syntax too complex					1		
2251281	6	EC.2.6.1.85	compound_c00251			y		n		6						1		
2251281	6	EC.4.1.3.38	compound_c00014	compound_c00568	y	y	y	n	n							1		
2251281	6	EC.4.1.3.38	compound_c00251			y		n								1		
2251281	7	EC.2.6.1.85	compound_c00014		y	y		n								1		
2251281	7	EC.2.6.1.85	compound_c00251			y		n								1		
2251281	7	EC.2.6.1.85		compound_c11355			y		n							1		
2251281	8	pabB	compound_c00251		n	y		n			pabb-catalyzed, no thes mapping		1		1	1		
2251281	8			aminodeoxychorismate			n		n				1		1	1		
2251281	10	EC.4.1.3.38	aminodeoxychoris-mate		y	n		n					1					
2251281	10	EC.4.1.3.38		compound_c00568			y		y	n		1						
2251281	10	EC.4.1.3.38		compound_c00022			y		y	n		1						
2251281	13		compound_c00568	compound_c00251		y	y			n						1		
2251281	18	EC.4.1.3.27	compound_c00251		y	y		n								1		
2251281	18	EC.4.1.3.27		compound_c18054			y		n							1		
2251281	18	EC.4.1.3.27		compound_c00022			y		n							1		
2251281	19	EC.2.6.1.85	compound_c11355		y	y		n								1		
2251281	19		compound_c11355	compound_c00568			y			n						1		
2251281	20	EC.5.4.4.2	compound_c00251		y	y		n			interconvert not in verbs					1		
2251281	20	EC.5.4.4.2		compound_c00885			y		n		requires different prep					1		
2251281	23	EC.2.6.1.85	compound_c00251		y	y		n								1		
2251281	23	EC.2.6.1.86	compound_c00014			y		n			??				1			

Table A. 14 - Detailed per sentence view of performance and evaluation of identification of metabolic reaction references in the aerobic fatty acid beta-oxidation pathway corpus

DOCID	SENTNR	ENZYME	SOURCE	TARGET	enzyme identified	source identified	target identified	enzyme-source	enzyme-target	source-target	annotation	found and correct	not in thes	too generic	syntax error	syntax not queried	should be in result	chemical formula
1460045	5	acyl coenzyme A synthetase	compound_c00638		n	y		n			: in sentence				1			
1460045	5	acyl coenzyme A synthetase		CoA thioester			n		n						1			
1460045	23		compound_c00162	CoA thioester		y	n			n			1	1				
1460045	26	Acyl-CoA synthetase	fatty acyl-CoA		n	n		n					1		1			
1460045	26	Acyl-CoA synthetase	compound_c00002			y		n								1		
1460045	35		CoA thioester	compound_c00865		n	y			n				1	1			
9030548	5-6	Fatty acyl-CoA synthetase	compound_c00013		n	y		n			; in sentence				1			
9030548	5-6	Fatty acyl-CoA synthetase		fatty acyl-CoA		n			n						1			
9030548	20-21	Fatty acyl-CoA synthetase	compound_c00162		n	y		n			; in sentence				1			
9030548	20-21	Fatty acyl-CoA synthetase	compound_c00002			y		n							1			
9030548	20-21	Fatty acyl-CoA synthetase	compound_c00010			y		n							1			
9030548	20-21	Fatty acyl-CoA synthetase		fatty acyl-CoA			n		n				1	1	1			
9030548	22		Fatty acyl-CoA	phospholipid		n	n			n				1	1	1		
9030548	26		fatty acyl-adenylate	fatty acyl-CoA		n	n			n			1	1				
9030548	26			compound_c00020			y			n						1		
12057976	10		compound_c00040	compound_c00658		y	y			y		1						
12535077	22		compound_c00638			y				y		1						
12535077	22			compound_c00024			y											
12535077	26	EC.1.3.99.3	compound_c00040		y	y		n			via not in prepositions					1		
12535077	26	EC.1.3.99.3		enoyl-CoA			n		n				1			1		
12535077	30		acyl-CoA	acetyl-CoA		n	n			n	greek characters				1	1		
12535077	33		compound_c00638	long-chain acyl-CoA		y	n			n					1			
6271734	26	EC.2.3.1.16	compound_c00162		y	y		n							?			
6271734	26	fadB	compound_c00162		n			n			fadA,B,C				?			
6271734	26	fadC	compound_c00162		n			n			fadA,B,C				?			
6271734	26	EC.1.3.99.3	compound_c00162		y			n			internal anaphora				?			
6271734	26		compound_c00162	compound_c00040		y	y			y		1						
9748275	6		compound_c05332	compound_c00163		y	y			y		1						

DOCID	SENTNR	ENZYME	SOURCE	TARGET	enzyme identified	source identified	target identified	enzyme-source	enzyme-target	source-target	annotation	found and correct	not in thes	too generic	syntax error	syntax not queried	should be in result	chemical formula
9748275	7	EC.6.2.1.30	compound_c00163		y	y	y	y				1						
9748275	7	EC.6.2.1.30		compound_c00582			y		y			1						
9748275	23	EC.6.2.1.30	compound_c00163		y	y		n		y		1				1		
9748275	23	EC.6.2.1.30		compound_c00582			y		n							1		
8755745	17	EC.1.1.1.35	l-3-hydroxyacyl-CoA		y	n		n										1
8755745	17	EC.1.1.1.35	NAD			n		n										1
8755745	17	EC.1.1.1.35	3-ketoacyl-CoA			n		n										1
8755745	17	EC.1.1.1.35	NADH			n		n										1
8755745	17	EC.1.1.1.35	H+			n		n										1
8755745	27	EC.4.2.1.17	l-3-hydroxyacyl-CoA		y	n		n			d- and l-hydroxyacyl-CoA				?			
8755745	27	EC.4.2.1.18	d-3-hydroxyacyl-CoA			n		n							?			
9283097	12	EC.4.2.1.17	compound_c00658		y	y		n							1	1		
9283097	22	EC.4.2.1.17	compound_c00001		y	y		n			no link between sentence parts				1			
9283097	22	EC.4.2.1.17		α,α-unsaturated fatty acyl-CoA thioester		n			n		greek characters						1	
9283097	34	enoyl-CoA hydratase	compound_c00658		n	y		n							1	1		
8993342	23	EC.4.2.1.74	l-3-hydroxyacyl-CoA		y	n		n			d- and l-hydroxyacyl-CoA				?			
8993342	23	EC.4.2.1.74	d-3-hydroxyacyl-CoA			n		n							?			
8993342	23	EC.4.2.1.74		compound_c00658			y		n						1			
15213221	11	EC.2.7.10.2	compound_c00162		y	y		n								1		
15213221	12	EC.2.7.10.2		acyl-AMP			n		n		to form		1			1		
15213221	12	EC.2.7.10.2		compound_c00040			y		n							1		
15213221	16	Acyl-CoA synthetase	compound_c00162		n	y		n							1			
15213221	16	Acyl-CoA synthetase		CoA thioester			n		n					1	1			
15213221	17		fatty acid	fatty acyl-AMP		n	n			n								1
15213221	17		ATP			n				n								1
15213221	17		fatty acyl-AMP	fatty acyl-CoA		n	n			n								1
15213221	17		fatty acyl-AMP	AMP			n			n								1
15213221	17		CoA			n				n								1
15213221	21		compound_c00162	acyl-CoA thioester		y	n			n				1				
15213221	45	EC.2.7.10.2	compound_c00162		y	y		n							1	1		

Chapter 5

General Discussion

The goals of this research are to reconstruct existing metabolic pathways, construct new ones and gather evidence from literature for the metabolic reactions that constitute those pathways.
In this thesis we describe a core text mining system built around a linguistic parser for the extraction of metabolic reactions from scientific literature (Chapter 4). As part of the text mining system we created two thesauri, one for enzymes and one for metabolites (Chapter 3). To be able to make an informed choice about the database system in which to store the literature and to facilitate access to the literature and resulting large datasets we evaluated a variety of XML-type (W3C) database solutions (Chapter 2).This evaluation was combined with a study of the text mining capabilities of these systems such as co-occurrence techniques and proximity search. In this chapter I discuss the reasoning behind the choice for a complex text mining approach. Potential alternatives and possible improvements and extensions to the core text mining system that are beyond the scope of the individual chapters are also discussed.

Necessity of literature analysis

Knowledge on metabolites and metabolic pathways is central to our understanding of living organisms as they form an important link between genotype and phenotype (Fiehn 2002). Information on metabolic reactions and pathways stored in databases is still incomplete as has been shown by the recent efforts of the RECON2 project (Thiele et al. 2013). The research community could greatly benefit from augmenting that information with the information contained in the ever increasing number of scientific publications. Traditionally pathway databases are filled by manual curation of scientific literature or active submission of reactions to the database provider. With the number of scientific papers published every day it is impossible for manual curators to keep up. To analyse all papers in PubMed would take around 5 million person-days (estimated from (Ceol et al. 2008)). Unlike the requirements for e.g. genes (GenBank (Benson et al. 2013) or proteins (UniProt (UniProt 2013)) there is no requirement for providing database accessions for metabolites, enzymes or metabolic reactions in scientific publications. Efforts have been made to formalize information held in scientific papers beyond keywords, MeSH terms and gene/protein identifiers to make the content accessible to automated data mining mainly in the form of structured digital abstracts as proposed by Gerstein (Gerstein et al. 2007) and Seringhaus (Seringhaus&Gerstein 2008). Mons et al. (Mons&Velterop 2009) propose to dispense with classical scientific publications altogether in exchange for nano-publications in the form of rich RDF-triples (Resource Description Framework (W3C)). These triples consist of entities and relations with unique identifiers (so-called Uniform Resource Identifiers), defined in a way similar to XML-schemas. RDF triples are by their nature machine readable. Apart from the discussion about usefulness and incentives for such schemes (Hahn et al. 2007), papers published up to the introduction of such a scheme do not hold this formalised information. Scientists

also struggle with the correct application of controlled vocabularies required for consistent annotation. This results in local variations in annotation quality, whereas text mining approaches have a uniform annotation quality, regardless of their overall performance. This leaves us for now with only the unstructured free text as source for automated analysis and text mining approaches as the only option to unlock the potential of scientific literature on a large scale.

Text mining

Natural language has many degrees of freedom for authors to describe their work, both on the level of terminology and on the semantic and syntactic level. Often the reader has to make assumptions about the intended meaning of the author. In text mining we want a machine to interpret human language. Success of a text mining system is customarily measured in terms of precision and recall, familiar from the Information Retrieval community (Baeza-Yates&Ribeiro-Neto 2011). Briefly, recall measures the number of recovered hits as proportion in relation to the number of hits in the entire source; precision measures the number of recovered hits as proportion in relation to the number of recovered items. It is an interesting question what the best is we can do. To put the current (and future) accomplishments of text mining by machines in context, we can inquire how well humans are at the task. One measure is the so-called inter-annotator agreement. When two humans are asked to evaluate a text, the inter-annotator agreement measures how well two informed readers agree about what the text is supposed to say. Inter-annotator agreement ranges usually between 85% and 95% for information extraction tasks (Véronis 1998; Brants 2000; Nobata et al. 2011).

Another hurdle for automated text mining is the fact that different types of written text (e.g. newspaper articles, scientific paper, and patent) and even different sections of scientific papers (e.g. introduction, materials and methods, discussion) use different styles and conventions.

These challenges posed by natural language mean that up to now there is no generic 'one size fits all' text mining system capable of analysing different types of texts or answering different types of questions with sufficient quality (Ananiadou&McNaught 2006). Given the great variability in writing conventions used in the different scientific disciplines and the variety of relations described it is very unlikely that such a system will exist in the future. The focus instead has been on specialist applications targeting specific entities and relations. Most published text mining applications in the biomedical world focus on proteins and genes and their interrelationships (Rzhetsky et al. 2004; Zhu et al. 2012). Only a few target the challenges posed by metabolic reactions, i.e. the identification of two different types of entities, enzymes and metabolites, and the relations that span a reaction of at least two metabolites and one enzyme.

The most straightforward approach to text mining would be to treat text as structured data in data mining by leaving it in its unstructured form, ignoring syntax and semantics, and retrieve information (e.g. named entities or co-occurring terms) from the text as is. This approach in theory provides maximum flexibility as to what kind of information can be retrieved from the text as there is no text analysis or pre-processing step limiting what can be accessed from the text. We briefly explored this naïve approach to text mining in Chapter 2 of this thesis in parallel to the main focus of performance analysis of off-the-shelf database solutions. For the analysis we focussed on XML-type data as this is one of the most used formats in bioinformatics (Medline, MIAME, PSI-MI). As part of the study we also evaluated the capabilities of the database systems for full-text searches and the simplest relation extraction approach, namely co-occurrence, to extract metabolic relations. Apart from the limitations of such a co-occurrence approach for identifying the type and direction of a relationship there were technical challenges to overcome as well. Especially the XPath (W3C) and XQuery (W3C) standards do not allow proximity searches (searching for one word occurring within a certain word distance from another). The more recent XPath full-text standard (W3C 2010b) (not to be confused with the XPath standard) does allow such searches.

Led by the results of Chapter 2, we chose to convert the unstructured text to structured information. This allows us to use a more complex text mining approach to extract more specific information, like direction, type, and confidence, about identified interactions and limit the number of falsely identified reactions (as opposed to co-occurrence). Against this, complex approaches tend to be tailored to specific domains, which somewhat limits the type of questions that can be answered by means of the extracted information. The system we created is a core text mining system, using a deep parsing approach, capable of extracting metabolic reactions from scientific publications with high precision. We built the system around EP4IR, a generic English grammar, and the AGFL system (Koster&Verbruggen 2002), a parser generator. We created two thesauri by semi-automated means from existing resources (KEGG (Kanehisa et al. 2006), ChEBI (Hastings et al. 2013) and BRENDA (Barthelmes et al. 2007)) to be able to identify enzymes and metabolites in the texts (Chapter 2). Our thesaurus-based approach allows us to link named entities from the text with those already stored in metabolic pathway databases like KEGG, allowing us to link evidence of reactions to the databases and to discriminate between described and un-described reactions. We adapted the EP4IR grammar to the specific needs of biomedical text in a bid to increase performance of the grammar. For the resulting extracted data we built a MySQL database for storage and querying. The creation and evaluation of the core text mining system is discussed in Chapter 4. While the results are promising, critical inspection of the results of the evaluation show that there is room for improvements and a number of big challenges remain. These can be roughly split into two topics, (I) improvements to the core system itself and (II) extensions to the system.

Improvements of the core system

While the precision of our results is very high (>95%, Chapter 4), the recall of about 20% leaves room for improvements. High precision is very desirable when studying large collections of texts as it limits the number of false positives to filter by manual work (Lee et al. 2012). The low recall however means that only a relatively small amount of the potentially available information is actually extracted from the text. Improvements to the precision are likely to lower recall even further at this stage. Therefore the emphasis for improvements has to be on increasing recall without sacrificing precision. These improvements can be introduced on various levels of the application, as discussed below.

Named Entity Identification (NEI)

One of the limiting factors in recognising metabolic reactions in the text is the ability to recognise metabolites and enzymes. The more terms are recognised, the higher the potential recall, making it an obvious target for improvements. The identification of terms falls into the field of named entity recognition (NER), where terms are recognised to belong to a given category, in the case of this thesis enzymes or metabolites. See the more elaborate discussion of Chapter 3. Various techniques can be applied to NER ranging from rule-based (Tanabe&Wilbur 2002) to machine-learning (Settles 2005; Leaman&Gonzalez 2008) or dictionary -based methods (Ono et al. 2001; Hettne et al. 2009), the latter is applied by us. Challenges for NER in general and dictionary-based methods in particular are synonyms, homonyms and spelling variations (Ananiadou&McNaught 2006). Synonyms are different terms describing the same entity. The occurrence of synonyms arises from a combination of sources like separate official and trivial names for entities, historic names or abbreviations. Spelling variations de facto give rise to synonyms as well. This large variation in terminology makes it difficult to capture all forms of a term in a thesaurus resulting in lack of coverage. Homonyms are exactly the opposite of synonyms, here a single term describes multiple entities. Reasons for the occurrence of homonyms are ambiguous abbreviations, generic terms or, in case of enzymes, multi-function enzymes that have to be assigned to more than one category. Synonyms lead to false negatives, lowering recall, while homonyms lead to false positives, lowering precision.

NER only allows the classification of terms, not the mapping of terms to class names. Yet for the identification of metabolic pathways this mapping to class names, so-called named entity identification (NEI), is paramount as pathways consist of specific enzymes and metabolites joined into networks. Currently only a thesaurus-based approach as applied by us allows the mapping of many synonyms to unique classes and thus to single entities with high confidence. However such a thesaurus-based approach has its own particular shortcomings. We will now discuss some ways to improve the contribution of the thesaurus to the entire system.

Thesaurus adaptations

Our parsing system needs a lexicon that specifies the basic building blocks of sentences complete with their part of speech (POS) such as "noun", "verb", and so on; in fact, it can handle multiple lexica that we will consider as one here for simplicity. One of the strengths of this approach is that it enables easy identification of compound nouns, nouns consisting of multiple words separated by word delimiters like spaces. Compound nouns occur frequently in our texts, for example, 'methyl branched-chain enoyl coenzyme a reductase'. We converted the entries of the thesauri of Chapter 3 into lexicon entries, ensuring that each term can later be related to its class name. This enables integration of the text mining results with existing databases by way of shared database identifiers.

The weakness of our approach lies in the static nature of the thesauri. Terms can only be identified and classified if they are included in the thesaurus. New terms have to be added in a separate step. This may result in a lack of coverage, which indeed became obvious in the evaluation described in Chapter 4. A number of potential true positives were lost due to the fact that the entities in question were not present in the thesaurus in the form used in the text. These missing entities were very often abbreviations and shortened forms (e.g. H2-folate for dihydrofolate) 'invented' by the authors.

There are solutions to increase the coverage of the thesauri (Schuemie et al. 2007). Still, it turns out that not every measure results in increased coverage or, even if the coverage is increased, in increased recall of the entire system. We mention a number of potential improvements with the proviso that their efficacy has to be determined in future research.

One seemingly obvious measure is the inclusion of more specialist or species-specific data sources like EcoCyc (Keseler et al. 2011) or AraCyc (Mueller et al. 2003) to increase the number of synonyms with those from specific data sources. Conversely this might also increase the number of homonyms in the thesaurus, thereby lowering synset purity (see Chapter 3). This negative effect may be mediated by applying specialist thesauri only to relevant corpora.

To maximise the potential of already collected synonyms one could apply a rule-based approach to create more spelling variants, as described by Schuemie et al. (Schuemie et al. 2007), despite the lack of influence most of these appear to have on recall as stated in their paper. As long as the adaptations do not negatively affect precision the inclusion of spelling variants is warranted as even small percentage increases of recall represents a significant number of interactions extracted from a large corpus. For the compound thesaurus specifically one could apply the approach of Engelken et al. (Engelken et al. 2009) to increase the number of synonyms by applying the IUPAC/IUBMB (McNaught&Wilkinson 1997) rewrite rules from their system to our thesaurus terms. It is also possible to scour the literature for potential new candidate synonyms in an approach similar to that of Tsuruoka and Tsujii (Tsuruoka&Tsujii 2004). They use a probabilistic variant generator to

expand the synset of an existing thesaurus with potential spelling variants, trying subsequently to identify actual use of these. Additional terms could be recovered by employing abbreviation identification systems (Schwartz&Hearst 2003), particularly addressing the problem of ad hoc invention of short forms.

To put the contribution of adding extra terms to the thesaurus in perspective, Scherf et al. (Scherf et al. 2005) note that all text mining approaches come up against Zipf's law (Li 1992). They determined experimentally that of all terms occurring in Medline, 40% occur only once. Given this scarcity of occurrences of single terms in literature, increase of coverage will require a lot of effort for measurable improvements. To make matters worse, biomedical terminology is also a moving target. Scientists will continue to invent new terms, ensuring that a thesaurus will always lag behind the facts. While some of the proposed solutions can be employed on the fly, our current parsing implementation, which uses the lexicon as POS tagger, requires all terms to be known beforehand. Also named entity identification on the fly, in which new terms are automatically assigned to identifiers, is likely to result in more false positives.

Instead of trying to increase the coverage of the thesaurus/lexicon beforehand, we can move from the exact string matching for terms we currently use to more fuzzy matching. One option is the application of rules to calculate Levenshtein distances (Levenshtein 1965) between unknown words in the text and thesaurus terms to determine the most likely matching thesaurus term. This would still allow NEI with a reasonable level of confidence. Great care has to be taken in setting up the penalty matrix as single character replacements can already result in very different compounds: compare "alkane" and "alkene". This can be compared with the penalty systems used in software such as BLAST. Krauthammer et al. (Krauthammer et al. 2000) use a sequence alignment approach for NEI in which they convert sequences of text characters into nucleotide sequences using a character to sequence conversion table. Both corpus text and thesaurus terms are converted in this way and aligned using the BLAST tool (Altschul et al. 1990) to determine the thesaurus entities most likely mentioned in the corpus. Here again the difficulty lies in the fact that a single mismatch can result in a different compound and a different compound normally means a false positive. Any increase in false positives will result in lower precision and this we set out to avoid. Sayle et al. (Sayle et al. 2011) use a finite state machine model to create a dictionary for IUPAC chemical names from an existing dictionary of chemical terms allowing the identification of all possible combinations of IUPAC-like terms from text. This also includes chemically false combinations resulting in false positives. This approach is not suitable for our method of mapping the identified chemicals to database identifiers as the created terms allow NER, but not NEI. The approach of Sayle et al. to use term to structure conversion to remove the false positives and to map the identified terms to actual compounds could be used instead. Another solution would be to attempt NEI as high quality result using the known thesaurus terms and use NER for all other terms and synonyms identified with above discussed methods.

In future work some ideas of the approaches to NER we discussed just now can be tested for their efficacy in increasing the coverage of both the enzyme and the metabolite thesauri. Despite all their shortcomings, good quality thesauri still are valuable resources in NEI for relation extraction.

Grammar adaptations

A second factor influencing precision and recall is the ability of the grammar to correctly parse the sentences. Despite our best efforts to adapt the rules and penalties of the grammar to the peculiarities of scientific biomedical text the results are far from perfect. We identified a number of syntactic constructs for which the parser consistently fails to produce the correct parse tree or for which the semantically correct parse tree is so heavily penalized that it does not gain preference above others. In some cases the failure to parse results from failure to identify the correct boundaries of a chemical term, the problem stemming from punctuation signs in the remainder of the term confusing the grammar. As discussed above increase of thesaurus coverage will alleviate more and more of these errors. In other cases very specific syntactic constructs fail to parse correctly. One prime example already discussed in Chapter 4 is the construct 'enzyme E converts metabolite A to form/to yield/to produce metabolite B'. One linguistic challenge lies in the ambiguity of the part of speech of each constituent. A second challenge is the identification of the correct agent: enzyme E or metabolite A. In this syntactically very simple sentence the correct parse is easily identified, yet it is difficult when subordinate clauses come into play which increase the distance between agent (the part that acts) and its target (the part that is acted upon). To adapt the grammar to favour the semantic interpretation describing a metabolic interaction with the enzyme E (the agent) and metabolites A and B (the patients) is not a simple task and requires careful monitoring of the parsing results for any unintended consequences.

A large concern is the inability of the parser to handle sentences with a large number of subordinate and/or relative clauses. Here the parser will often run out of options resulting in a partially parsed sentence. This limitation in the grammar is a trade-off between the number of possible parses and the maximum parsing depth. To allow complete parsing of such complex sentences the number of possible combinations of clauses has to be increased while avoiding circular reasoning in the grammar. An increase in possible parses will also result in an increase in computing time. Even marginal increases in parsing time have large consequences for overall computing time or resource requirements when large corpora are involved, even though parsers in practice are found to scale polynomially as n^3, where n is a measure of the length of the input. For our system n can be determined as the number of times a lexicon entry occurs in the text (where each compound noun specified by the lexicon counts as one) plus the number of unrecognised strings.

Rewriting grammar rules (including setting other values for penalties or bonuses) is very time consuming and requires in depth linguistic knowledge. Great care

has to be taken to evaluate the change in parsing performance for any unintended consequences of grammar changes effecting precision and recall.
Complete parsing is not always necessary for successful relation extraction as GenIE shows a strong partial parsing approach and GENIES reverts to partial parses when full parses are not possible. AGFL/EP4IR is also capable of returning parses of syntactically incomplete sentences like titles making it very robust. Yet the current penalty-based parsing approach favours left-sided incomplete parses over phrase parses in long sentences. This results in the loss of, usually, metabolite mentions towards the end of sentences and the link of those with enzyme mentions more towards the beginning of the sentence. Phrase-based or chunked parsing also results in the loss of syntactic dependencies between prepositional and relative clauses and their antecedent, limiting the potential to extract more complex relations. A grammar adaptation with the aim of extracting complex relations and semantic nuances should therefore focus on complete parsing, and only when that is impossible revert to partial parsing similarly to the panic mode of Pyysalo et al. (Pyysalo et al. 2004), in which full parsing is abandoned after a set time period allowed for the parse in favour of partial parsing. Setting a time period makes the approach deliver different results on systems with different hardware. Setting a limit on the number of backtracks will not suffer from this drawback but then it can no longer be predicted how long a parse will take.

Anaphora resolution

To even further increase recall we have to maximise the amount of useful information retrieved from the text. A large amount of that information is obscured in the form of anaphoric references. Anaphora are a way to refer to subjects in text by indirectly mentioning them with, for example, pronomials like 'it', demonstratives like 'this' or 'that', or sortals such as 'this enzyme'. The most occurring anaphoric construction in biomedical text such Medline is sortals such as 'this enzyme' and 'both products' (Torii&Vijay-Shanker 2007). Anaphora can occur within sentences (as in "the enzyme and its cofactor") or as discourse anaphora that span across sentence boundaries (as in "We studied alcohol dehydrogenase. This enzyme occurs in many forms"). Castano et al. (Castano et al. 2002) inspected 100 discourse anaphora collected from Medline abstracts and found that around 60 of them are sortal anaphora. This would make sortal anaphora a prime target for increase of recall.
The challenge for the resolution of anaphora lays in the identification of the correct antecedent, the term or noun phrase referenced by the anaphor. While the choice of antecedent within a sentence is naturally limited, it is still difficult to identify as this relies heavily in semantic clues as well as syntactic constraints. Even more complex is identification of the correct antecedent across sentence boundaries which relies even more on semantic clues and identification of the correct context-bearing keywords (Thiele et al. 2013).

Within-sentence anaphora fall within the scope of our sentence-based parser and indeed some of the simpler forms are solved in the syntax analysis:

a) The second enzyme, which has been named H2-pteroate synthetase, catalyses the synthesis of dihydropteroic acid from p-aminobenzoic acid.

b) stat:[[N:second@4-11 ,INVof [V:catalyzes@64-74 ,OBJ [[N:synthesis@78-88 ,|of [acid@106-111 ,ATTR A:R:-dihydropteroic@91-106]o|from [N:p-aminobenzoic acid@116-136]]]|INVSUBJ [enzyme@11-17 ,INVOBJ [V:named@34-40 ,INVSUBJ P:it@62-62]|PRED [N:h2-pteroate synthetase@40-62]]]]]

c) enzyme,SUBJ,catalyse

catalyse,OBJ,synthesis
enzyme,PRED, H2-pteroate synthetase

Currently the syntax analysis does not identify the antecedent by linking it directly, instead using the impersonal pronoun 'it' in the syntax tree and a predicative relation between antecedent and anaphor in the extracted syntactic triples. In this particular example the reference precedes what is referred to. This is known as a cataphor. Lacking more intricate experimental research into these phenomena we cannot say at this point whether cataphors are more easily resolved than other within-sentence anaphors. However, it seems not far-fetched to assume that left-corner parsers such as AGFL indeed handle cataphors better than other anaphors.

The correct antecedent is not always found, especially in the cases where the sentence contains more than one anaphor and subordinate clauses (*see* Grammar Adaptations and Chapter 4).

Discourse anaphora are beyond the scope of the parser and would have to be solved by different means in a post-processing step. There are different published approaches to anaphora resolution (Castano et al. 2002; Torii&Vijay-Shanker 2007) focussing on biomedical text and sortal anaphora that could be integrated with our system. It should be noted that inclusion of such an additional step will increase false positives. Torii report 71% precision with 77% recall overall and Castano between 68% and 80% precision and 61% and 75% recall depending on applied technique and type of anaphor for their respective systems.

We may increase anaphora resolution in our system, but in our anaphora experiment reported in Chapter 4 we found that although anaphora are very frequent, anaphora resolution does not make a large difference. It turned out that anaphora are not commonly used in sentences describing metabolic reactions. Whether this is a peculiarity of the corpus we used or occurs more widely remains to be seen.

Extensions to the core text mining system

So far we discussed adaptation of the core system to increase recall. In this section we want to illustrate and discuss the potential of our text mining approach and how it could function in a larger application that may both improve performance and increase accessibility.

Database integration

A first step to increase the potential of the reactions uncovered from literature would be to integrate the results with other databases. Immediately obvious is the link to other pathway databases like KEGG and Reactome or WikiPathways (Pico et al. 2008) by the compound and enzyme identifiers used in the thesauri. Most of our compounds carry ChEBI identifiers in addition to the KEGG compound ID and databases referencing either identifier can be linked. We use EC numbers as enzyme identifiers, and results can be linked to any database employing the same classification. When the goal is knowledge discovery, this mapping can function as filter to distinguish between already known and potentially novel reactions. From an information retrieval perspective our results add literature evidence to already known reactions. Both classes of results can be used to assist manual curation efforts (e.g. RECON2 (Thiele et al. 2013)) in which the results can be used as link to literature for both already known and potentially novel reactions. The aid would be on a document classification level, limiting the number of publications to be analysed manually. It can also be used to power a visual aid that highlights the identified reactions in the relevant documents.
It would not only be interesting to link the recovered entities and reactions to pathway databases but also to other biological databases like UniProt (UniProt 2013) . Once the enzymes of the reactions are linked to the proteins stored in UniProt, the reaction evidence is linked to all cross-referenced databases, thus connecting the information to protein-protein interactions (MiNT (Chatr-aryamontri et al. 2007)), 3D structures (SMR (Kopp&Schwede 2006) , PDB (Berman et al. 2000)) and protein family databases (PROSITE (Sigrist et al. 2013)1V111XE z,Zy, Pfam (Punta et al. 2012)), and more, integrating it into a vast universe of bioinformatics resources. An obstacle for integration with other databases is the potentially different levels of granularity between the entities that have to be mapped. In the example of UniProt, we recover enzymes grouped by function as defined in the enzyme nomenclature (IUPAC (Enzyme Nomenclature)) whereas UniProt is a database of species-specific proteins. The resulting pathways our system recovers are what one could call reference pathways, where reactions are recovered and linked regardless of the originating species. In principle the results of a text mining effort can be linked to a species but that takes extra measures. We can thus either keep our broader granularity and loose the specificity provided by UniProt or adapt our system to the finer granularity of UniProt. In that case integration with a species name

identification system like LINNAEUS (Gerner et al. 2010) could facilitate mapping of identified enzymes to UniProt entries. We would have to determine at which textual level the species will be determined. The sentence level is in most cases too narrow and the document level is too broad if more than one species is mentioned. Complications arise if model species are used for the experiment yet conclusions are drawn for another species (e.g. mouse model for human study) or when studies are performed with genetically modified organisms, in which genes from one species have been transferred to another.

Weighting recovered reactions

Text mining results from a collection of literature provide a view of research and opinions spanning the fields and time span included in the corpus. Results may converge but inconsistencies are also possible. Is there a way to resolve inconsistencies? We distinguish between internal factors and external factors. Internal factors deal with the text in isolation, for example the wording of the text. Currently we only distinguish between positive interactions (e.g. 'convert') and outright negative interactions (e.g. 'does not convert'), but language affords many more shades of affirmation between these two extremes of assertions (Blake 2010). Making use of the possibilities of the complete syntactic analysis of the text it should be possible to distinguish between a number of gradations in the assertions made (for example, 'not', 'not likely', 'probably', 'very likely', 'we assume') and thereby increasing the resolution of the results. In combination with the external ranking this should provide a more complete picture of the reliability of each identified metabolic reaction. Yet given the current applications in the field of text mining such a refined system is far off in the future.

Turning now to external factors, one relatively simple method is to weight contributions stemming from different sources. In order to accurately weight the results a number of ranking or weighting features can be imagined to indicate our confidence in the veracity of a recovered reaction. If for example results regarding a particular enzyme are contradictory we solicit three approaches to break the tie. The simplest way is to count the number of occurrences of each assertion and follow the majority. Yet this tie-breaking approach in some ways disregards the dependence of publications on previous work through citations. When an investigated corpus spans a range of time it is only fair to put greater weight on newer insights even if they are (so far) less established as science advances. The same temporal order of results can give us an overview of the history of a certain research topic.

Likewise not all publications are considered equal. Peer-reviewed articles are generally valued higher than those that are not and the general perception is that research published in Science or Nature carries more weight than that published in journals with lower impact factor. Lin et al. studied the value of journal impact and citation measurements for ranking Medline search results and conclude that citation count per year is the best metric for ranking (Lin et al. 2007). So using this measurement

would provide a possibility to rank results.
Of course each of the three rankings (temporal, peer-reviewed versus not peer-reviewed, impact factor) only present a partial view of the truth. It would seem advantageous to create a combined measurement weighting the three rankings with respect to each other to obtain the best possible result after careful evaluation of the quality of such a combined ranking.

Access to literature

The usability and comprehensiveness of a text mining application depends on the amount of literature it is able to analyse. Currently we use MEDLINE abstracts as our main source of text, as do most other text mining applications. Use of only abstracts will limit our ability to find new and meaningful information. Studies have shown that the information content and density greatly varies between article abstract and body and between the different sections of the body (Cohen et al. 2010). Blake (Blake 2010) shows that on average only around 8% of the scientific claims of a paper are made in the abstract. This clearly illustrates the need to extend text mining analyses to full text.
The availability of full text is a problem, though. We also parsed a release of BioMed Central, an open access full-text database. This is the only current collection of full-text papers freely available for text mining. Other, larger datasets include the articles of the BioCreative and TREC genomics challenges, yet they only provide a snapshot and are not updated. Commercial publishers are largely hesitant to allow access to their literature stores for text miners, fearing undue strain on their network resources and theft of their intellectual property. Standard copyright licenses prohibit large scale download (Van Noorden 2012; Van Noorden 2013). Elsevier currently allows access to their data through web services and encourages researchers to contact them for access (Van Noorden 2012 comment), but recent talks on a more global scale for computational access have broken down (Van Noorden 2013). Hopefully, in the future a solution can be found that is satisfactory to both parties and that enables the research community to make full use of the riches of scientific literature.
Another aspect of the availability issue is the format in which literature is being provided. Nowadays papers exist in an online HTML version on the publisher's website and a PDF version for print. BioMed Central provides its paper collection in SGML, a language written in XML. Papers published before around 1995 are usually only available in hard-copy form or a scanned version, usually PDF, of the printed article. Each of these formats provides its own particular challenges before it can be used as input to text mining applications. The HTML versions are not standardised across publishers and journals and each version requires a different parser to separate mark-up from content. As added challenge, menus and other page content have to be separated from the actual article. PDF is a presentation format, meant for visual layout and not for extracting its content. The flow of the text in columns with interspersed headers and footers plus images and tables makes

it difficult to identify the correct textual content in its correct order. Even more complicated is the extraction of text from older, scanned copies of articles in PDF form. This was obvious in the corpora used in the benchmark of Chapter 4. In addition to the difficulty to extract the correct flow of text it is also challenging to recognize and convert special characters. In the conversion to flat text, as supplied by Czarnecki et al., superscript numbers indicating references had been converted to normal numbers making them indistinguishable from correctly occurring in-line numbers. The chemical reaction formulas breaking up the text flow similar to a table or figure had been integrated within the surrounding sentences during conversion, breaking both the sentence grammar and losing the information described in the actual reaction. We changed the documents by hand to assess the influence of the different confounding features on parser performance. These changes would have been impossible without being able to look at the original documents.

The SGML format (Goldfarb&Rubinsky 1990) provided in BioMed Central is standardised where it concerns tags indicating the flow of an article, but it is still both a content and layout format combined, and extensible. This means that different publishers add their own tags within sections to indicate or highlight specific portions of text, leaving the need to write a specific pre-processor for each journal. However using the BioMed Central corpus as input data, automated application of the changes required for the experiments of Chapter 4 was feasible. In the SGML-format references to cited papers and tables and figures are tagged with '<xref>' tags allowing simple removal. As mentioned, Greek characters can be easily replaced, in this case by replacing the XML-replacement character (e.g. *α* with *alpha*). The same approach can be taken for chemical reactions by identifying the pattern created by the XML characters and replacing the pattern with the corresponding sentence.

In general, text mining applications would be greatly aided with easy access to publications provided in a document format focussing only on content and not on layout, and by mark-up shared across the publishing world.

Accessibility of the results of the text mining system

The purpose of a text mining system like ours is of course to gain scientific insights from the results and to enable other scientists to access and use its findings. Currently access to our dataset is limited and only the two thesauri are available for download. The results of the extraction of metabolic reactions from the evaluation corpora and BioMed Central are stored in an Oracle 11.2g database. While this format makes the data accessible to most bioinformaticians it is hardly accessible for biologists. One can imagine different ways to enable access to the data held in our database. One of the more simple solutions would be a web service enabling users to integrate our data with other bioinformatics applications and pipelines. A potential use of such a web service would be the integration of extracted metabolic reactions in the putative pathway parts plug-in for PathVisio (van Iersel et al. 2008), suggesting potential

pathway extensions for a selected entity from different data sources.
Another way to facilitate access would be a more visual approach in a stand-alone application or web interface in which the user can navigate through or build up a metabolic network.
In all these considerations the interaction with intended users should serve to determine the requirements of such an application or web service. In *Text Mining – the way forward* (Altman et al. 2008) several of those interviewed noted that user access, meaning access by biologists, to the results of text mining applications has to be at the forefront of future development. Up to now most systems are only of use to information technology specialists and bioinformaticians at best. PathText (Kemper et al. 2010) is a system trying to break that mould and includes various user interfaces. It is actively used by biologists. The uptake of an application by end users is usually strongly correlated with the convenience of accessing and installing the application and the ease-of-use of the interface and features (Untergasser et al. 2007; Gehlenborg et al. 2010). To ensure this it is vitally important to include the end user in any interface development using techniques like user-centred design (Pavelin et al. 2012; de Matos et al. 2013) and future developments of interfaces for our application should follow these principles.

The role of text mining in bioinformatics

Despite the increase of research published on text mining the future role of text mining has still to be determined. The need to somehow access the knowledge held in scientific literature is obvious and alternatives to text mining for this purpose, like machine readable abstracts or nano-publications, have not yet arrived. This leaves text mining in all its facets as the scientific field to unlock the large volumes of unstructured text. Yet to what extent we can label and extract information held within the text remains to be seen. The figures for inter-annotator agreement show that 100% precision and 100% recall are impossible for humans, and machines will do worse for some time to come. All patterns describing entities and relations in text also follow Zipf's law which inherently limits the fraction of information that can realistically be extracted (Rebholz-Schuhmann et al. 2005). Natural language is a complex thing and scientists use and abuse every possible feature it presents. Scientists, although generally accurate in their research, are sometimes not so accurate in their writings and leave room for ambiguities and misunderstandings; some findings are over exaggerated and others left intentionally vague. Also scientists in this global community have different language backgrounds with English often not their first language. Those different backgrounds mean different interpretations in the nuances of words and syntax (Englander&López-Bonilla 2011; Alejandro 2012). Different research communities also use words in different meanings. Infamously the fruit fly community uses common English words as gene names (Proux et al. 1998).
These problems of natural language, even when interpreted by expert humans, indicate that there is a theoretical maximum for precision and recall of text mining

applications, although we currently do not know that maximum. It then remains up to the user to judge how far to use and trust the results of automated natural language analysis, just has he/she has to judge any other scientific method.
Hearst defines text mining in its strictest sense as extracting and linking of facts from text to gain new insights and knowledge (Hearst 2003) what is text mining]. Andronis et al. (Andronis et al. 2011) make a strong case for the ability to gain new insights from the disjointed scientific literature by linking evidence through joint connections based on Swanson's ABC model (Swanson 1990). This conversion of implicit links into explicit relations is exactly what we have been trying to attempt when we link enzymes and metabolites together into pathways. Others have shown that there are hidden gems to be found in text (Swanson 1987; Waagmeester et al. 2009). A specific use-case for the metabolic reactions extracted with our approach would be where a biologist studies a single enzyme, metabolite or pathway and wants to gain an overview of all its known reactions and associated publications. Van Landeghem et al. (Van Landeghem et al. 2013) conclude that text mining will become an indispensable addition to the analysis of results of large-scale biological studies, as *in-vivo* validation of individual results is infeasible and all available information, including that from homologous species, should be used.
I see the main role of a text mining system like ours mainly in gathering evidence for existing knowledge and giving insights into the nuances of the research landscape of a given topic. When using the results of our reaction extraction system for the identification of 'new' reactions it is important to go back to the actual evidence presented for extra validations and to cross-validate the predictions with other resources or experiments. When a ranking system has been implemented and the reliability of each reaction can be evaluated the results may be used with higher confidence. But ideally text mining will be used for generation of hypotheses, in which the researcher uses text mining findings to get ideas on, in our case, new connections between metabolites and enzymes; the researcher then goes back to the original texts for further study. In this role text mining should become an essential tool on the workbench of the molecular biologist.

References

Alejandro B. 2012. Running Like Alice and Losing Good Ideas: On the Quasi-Compulsive Use of English by Non-native English Speaking Scientists. *AMBIO: A Journal of the Human Environment*: 1-4.

Altman RB, Bergman CM, Blake J, Blaschke C, Cohen A et al. 2008. Text mining for biology-the way forward: opinions from leading scientists. *Genome Biol* **9**(Suppl 2): S7.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of molecular biology* **215**(3): 403-410.

Ananiadou S, McNaught J. 2006. *Text mining for biology and biomedicine*. Artech House Boston, London.

Andronis C, Sharma A, Virvilis V, Deftereos S, Persidis A. 2011. Literature mining, ontologies and information visualization for drug repurposing. *Briefings in Bioinformatics* **12**(4): 357-368.

Baeza-Yates R, Ribeiro-Neto B. 2011. Modern information retrieval. In *Modern information retrieval*, Vol 463. ACM press New York.

Barthelmes J, Ebeling C, Chang A, Schomburg I, Schomburg D. 2007. BRENDA, AMENDA and FRENDA: The enzyme information system in 2007. *Nucleic Acids Research* **35**(SUPPL. 1).

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ et al. 2013. GenBank. *Nucleic Acids Res* **41**(Database issue): D36-42.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN et al. 2000. The Protein Data Bank. *Nucleic Acids Research* **28**(1): 235-242.

Blake C. 2010. Beyond genes, proteins, and abstracts: Identifying scientific claims from full-text biomedical articles. *Journal of Biomedical Informatics* **43**(2): 173-189.

Brants T. 2000. Inter-annotator Agreement for a German Newspaper Corpus. In *LREC*.

Castano J, Zhang J, Pustejovsky J. 2002. Anaphora resolution in biomedical literature. In *Proceedings of International Symposium on Reference Resolution for NLP*, Alicante, Spain.

Ceol A, Chatr-Aryamontri A, Licata L, Cesareni G. 2008. Linking entries in protein interaction database to structured text: The FEBS Letters experiment. *FEBS letters* **582**(8): 1171-1177.

Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV et al. 2007. MINT: the Molecular INTeraction database. *Nucleic Acids Research* **35**(suppl 1): D572-D574.

Cohen KB, Johnson H, Verspoor K, Roeder C, Hunter L. 2010. The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC Bioinformatics* **11**(1): 492.

de Matos P, Cham J, Cao H, Alcantara R, Rowland F et al. 2013. The Enzyme Portal: a case study in applying user-centred design methods in bioinformatics.

BMC Bioinformatics **14**(1): 103.

Engelken H, Golebiewski M, Bittkowski M, Hamm F, Saric J et al. 2009. Flache und semantische Verarbeitung von Namen biochemischer Verbindungen. *INFORMATIK - Im Focus das Leben*: 687-692.

Englander K, López-Bonilla G. 2011. Acknowledging or denying membership: Reviewers' responses to non-anglophone scientists' manuscripts. *Discourse Studies* **13**(4): 395-416.

Enzyme Nomenclature [http://www.chem.qmul.ac.uk/iubmb/enzyme/]

Fiehn O. 2002. Metabolomics - The link between genotypes and phenotypes. *Plant Molecular Biology* **48**(1-2): 155-171.

Gehlenborg N, O'Donoghue SI, Baliga NS, Goesmann A, Hibbs MA et al. 2010. Visualization of omics data for systems biology. *Nature methods* **7**: S56-S68.

Gerner M, Nenadic G, Bergman CM. 2010. LINNAEUS: a species name identification system for biomedical literature. *BMC Bioinformatics* **11**(1): 85.

Gerstein M, Seringhaus M, Fields S. 2007. Structured digital abstract makes text mining easy. *Nature* **447**(7141): 142.

Goldfarb CF, Rubinsky Y. 1990. *The SGML handbook*. Clarendon Press Oxford.

Hahn U, Wermter J, Blasczyk R, Horn PA. 2007. Text mining: powering the database revolution. *Nature* **448**(7150): 130-130.

Hastings J, de Matos P, Dekker A, Ennis M, Harsha B et al. 2013. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Research* **41**(D1): D456-D463.

What is Text Mining? [http://people.ischool.berkeley.edu/~hearst/text-mining.html]

Hettne KM, Stierum RH, Schuemie MJ, Hendriksen PJM, Schijvenaars BJA et al. 2009. A dictionary to identify small molecules and drugs in free text. *Bioinformatics* **25**(22): 2983-2991.

Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M et al. 2006. From genomics to chemical genomics: new developments in KEGG. *Nucl Acids Res* **34**(suppl_1): D354-357.

Kemper B, Matsuzaki T, Matsuoka Y, Tsuruoka Y, Kitano H et al. 2010. PathText: a text mining integrator for biological pathway visualizations. *Bioinformatics* **26**(12): i374-i381.

Keseler IM, Collado-Vides J, Santos-Zavaleta A, Peralta-Gil M, Gama-Castro S et al. 2011. EcoCyc: a comprehensive database of Escherichia coli biology. *Nucleic Acids Research* **39**(suppl 1): D583-D590.

Kopp J, Schwede T. 2006. The SWISS-MODEL Repository: new features and functionalities. *Nucleic Acids Res* **34**(Database issue): D315-318.

Koster CHA, Verbruggen E. 2002. The AGFL Grammar Work Lab. In *Proceedings of the FREENIX Track: 2002 USENIX Annual Technical Conference*. USENIX Association.

Krauthammer M, Rzhetsky A, Morozov P, Friedman C. 2000. Using BLAST for identifying gene and protein names in journal articles. *Gene* **259**(1-2): 245-

252.

Leaman R, Gonzalez G. 2008. BANNER: an executable survey of advances in biomedical named entity recognition. In *Pacific Symposium on Biocomputing*, Vol 13, pp. 652-663.

Lee J, Kim S, Lee S, Lee K, Kang J. 2012. High precision rule based PPI extraction and per-pair basis performance evaluation. In *Proceedings of the ACM sixth international workshop on Data and text mining in biomedical informatics*, pp. 69-76. ACM, Maui, Hawaii, USA.

Levenshtein V. 1965. Binary codes capable of correcting spurious insertions and deletions of ones. *Problems of Information Transmission* **1**(1): 8-17.

Li W. 1992. Random texts exhibit Zipf's-law-like word frequency distribution. *Information Theory, IEEE Transactions on* **38**(6): 1842-1845.

Lin Y, Li W, Chen K, Liu Y. 2007. A document clustering and ranking system for exploring MEDLINE citations. *Journal of the American Medical Informatics Association* **14**(5): 651-661.

McNaught AD, Wilkinson A. 1997. *Compendium of chemical terminology*. Blackwell Science Oxford.

Mons B, Velterop J. 2009. Nano-Publication in the e-science era. In *Workshop on Semantic Web Applications in Scientific Discourse (SWASD 2009)*.

Mueller LA, Zhang P, Rhee SY. 2003. AraCyc: a biochemical pathway database for Arabidopsis. *Plant Physiol* **132**(2): 453-460.

Nobata C, Dobson P, Iqbal S, Mendes P, Tsujii Ji et al. 2011. Mining metabolites: extracting the yeast metabolome from the literature. *Metabolomics* **7**(1): 94-101-101.

Ono T, Hishigaki H, Tanigami A, Takagi T. 2001. Automated extraction of information on protein-protein interactions from the biological literature. *Bioinformatics* **17**(2): 155-161.

Pavelin K, Cham JA, de Matos P, Brooksbank C, Cameron G et al. 2012. Bioinformatics Meets User-Centred Design: A Perspective. *PLoS Comput Biol* **8**(7): e1002554.

Pico AR, Kelder T, van Iersel MP, Hanspers K, Conklin BR et al. 2008. WikiPathways: pathway editing for the people. *PLoS biology* **6**(7): e184.

Proux D, Rechenmann F, Julliard L, Pillet V, Jacq B. 1998. Detecting gene symbols and names in biological texts: a first step toward pertinent information extraction. *GENOME INFORMATICS SERIES*: 72-80.

Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J et al. 2012. The Pfam protein families database. *Nucleic Acids Research* **40**(D1): D290-D301.

Pyysalo S, Ginter F, Pahikkala T, Boberg J, J\ J et al. 2004. Analysis of link grammar on biomedical dependency corpus targeted at protein-protein interactions. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pp. 15-21. Association for Computational Linguistics, Geneva, Switzerland.

Rebholz-Schuhmann D, Kirsch H, Couto F. 2005. Facts from Text—Is Text Mining

Ready to Deliver? *PLoS Biol* **3**(2): e65.
Rzhetsky A, Iossifov I, Koike T, Krauthammer M, Kra P et al. 2004. GeneWays: A system for extracting, analyzing, visualizing, and integrating molecular pathway data. *Journal of Biomedical Informatics* **37**(1): 43-53.
Sayle R, Xie PH, Muresan S. 2011. Improved chemical text mining of patents with infinite dictionaries and automatic spelling correction. *Journal of chemical information and modeling* **52**(1): 51-62.
Scherf M, Epple A, Werner T. 2005. The next generation of literature analysis: Integration of genomic analysis into text mining. *Briefings in Bioinformatics* **6**(3): 287-297.
Schuemie MJ, Mons B, Weeber M, Kors JA. 2007. Evaluation of techniques for increasing recall in a dictionary approach to gene and protein name identification. *Journal of Biomedical Informatics* **40**(3): 316-324.
Schwartz AS, Hearst MA. 2003. A simple algorithm for identifying abbreviation definitions in biomedical text. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing*: 451-462.
Seringhaus M, Gerstein M. 2008. Manually structured digital abstracts: A scaffold for automatic text mining. *FEBS letters* **582**(8): 1170.
Settles B. 2005. ABNER: An open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics* **21**(14): 3191-3192.
Sigrist CJ, de Castro E, Cerutti L, Cuche BA, Hulo N et al. 2013. New and continuing developments at PROSITE. *Nucleic Acids Res* **41**(Database issue): D344-347.
Swanson DR. 1987. Migraine and magnesium: eleven neglected connections. *Perspectives in biology and medicine* **31**(4): 526-557.
Swanson DR. 1990. Medical literature as a potential source of new knowledge. *Bulletin of the Medical Library Association* **78**(1): 29.
Tanabe L, Wilbur WJ. 2002. Tagging gene and protein names in biomedical text. *Bioinformatics* **18**(8): 1124-1132.
Thiele I, Swainston N, Fleming RMT, Hoppe A, Sahoo S et al. 2013. A community-driven global reconstruction of human metabolism. *Nat Biotech* **31**(5): 419-425.
Torii M, Vijay-Shanker K. 2007. Sortal Anaphora Resolution in Medline Abstracts. *Computational Intelligence* **23**(1): 15-27.
Tsuruoka Y, Tsujii Ji. 2004. Improving the performance of dictionary-based approaches in protein name recognition. *Journal of Biomedical Informatics* **37**(6): 461-470.
UniProt. 2013. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Research* **41**(D1): D43-D47.
Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R et al. 2007. Primer3Plus, an enhanced web interface to Primer3. *Nucleic Acids Research* **35**(suppl 2): W71-W74.
van Iersel MP, Kelder T, Pico AR, Hanspers K, Coort S et al. 2008. Presenting and

exploring biological pathways with PathVisio. *BMC Bioinformatics* **9**: 399.
Van Landeghem S, De Bodt S, Drebert ZJ, Inzé D, Van De Peer Y. 2013. The potential of text mining in data integration and network biology for plant research: A case study on Arabidopsis. *Plant Cell* **25**(3): 794-807.
Van Noorden R. 2012. Trouble at the text mine. *Nature* **483**(7388): 134-135.
Van Noorden R. 2013. Tensions grow as data-mining discussions fall apart. *Nature* **498**(7452): 14-15.
Véronis J. 1998. A study of polysemy judgements and inter-annotator agreement. In *Programme and advanced papers of the Senseval workshop*, pp. 2-4.
XML Path Language (XPath) [http://www.w3.org/TR/1999/REC-xpath-19991116]
Resource Description Framework [http://www.w3.org/RDF/]
Extensible Markup Language (XML) 1.1 (Second Edition) [http://www.w3.org/TR/2006/REC-xml11-20060816]
XQuery 1.0: An XML Query Language [http://www.w3.org/TR/2007/REC-xquery-20070123/]
XQuery and XPath Full Text 1.0 [http://www.w3.org/TR/xpath-full-text-10/]
Waagmeester A, Pezik P, Coort S, Tourniaire F, Evelo C et al. 2009. Pathway Enrichment Based on Text Mining and Its Validation on Carotenoid and Vitamin A Metabolism. *OMICS: A Journal of Integrative Biology* **13**(5): 367-379.
Zhu F, Patumcharoenpol P, Zhang C, Yang Y, Chan J et al. 2012. Biomedical text mining and its applications in cancer research. *Journal of Biomedical Informatics*.

Summary

Science relies on data in all its different forms. In particular in molecular biology and bioinformatics, large scale data generation has taken centre stage in the form of high-throughput experiments. In parallel with this exponential increase of experimental data has been the near exponential growth of scientific publications. Access of information found in scientific literature is still limited to search engines allowing searches on the level keywords, titles and abstracts. However, large amounts of knowledge about biological entities and their relations are held within the body of articles. When extracted, this data can be used as evidence for existing knowledge or hypothesis generation making scientific literature a valuable scientific resource. In order to unlock the information inside the articles a dedicated set of techniques and approaches tailored to the unstructured nature of free text is required. Analogous to data mining for the analysis of structured data, text mining has emerged for unstructured text and a number of applications have been developed in that field.

This thesis is about text mining in the field of metabolomics. Metabolic reactions are important for our understanding of metabolic processes within cells and that information provides an important link between genotype phenotype. Furthermore information about metabolic reactions stored in databases is far from complete making it an excellent target for text mining applications. Our work focusses on strategies for accessing large collections of scientific text and on the text mining steps required to extract metabolic reactions and their constituents, enzymes and metabolites, from scientific text.

In order to access the body of scientific publications for further analysis they can be used in a flat text form or loaded into database systems. In **Chapter 2** we assessed and discussed the capabilities and performance of XML-type database systems to store and access very large collections of XML-type documents in the form of the Medline corpus, a collection of more than 20 million of scientific abstracts. XML data formats are common in the field of bioinformatics and are also at the core of most web services. With the increasing amount of data available in XML format comes the need for storing and accessing the data. The database systems were evaluated on a number of aspects broadly ranging from technical requirements to ease-of-use and performance. The performance of the different XML-type database systems was measured using Medline abstract collections of increasing size and a number of different queries. One of the queries assessed the capabilities of each database system to search the full-text of each abstract, which would allow access to the information within the text without further text analysis. The results show that all database systems cope well with the small and medium datasets, but that the full dataset remains a challenge. Also the query possibilities varied greatly across all studied databases. This led us to conclude that the performances and possibilities of the different database types vary greatly, also depending on the type of research question. There is not a single system that outperforms the others; instead different circumstances can lead to a different optimal solution. Some of these scenarios are

presented in the chapter.
Among the conclusions of **Chapter 2** is that conventional data mining techniques do not work for the natural language part of a publication beyond simple retrieval queries based on pattern matching. The natural language used in written text is too unstructured for that purpose and requires dedicated text mining approaches, the main research topic of this thesis. There are two major tasks of text mining: named entity recognition, the identification of relevant entities in the text, and relation extraction, the identification of relations between those named entities. For both text mining tasks many different techniques and approaches have been developed. For the named entity recognition of enzymes and metabolites we used a dictionary-based approach (**Chapter 3**) and for metabolic reaction extraction a full grammar approach (**Chapter 4**).
In **Chapter 3** we describe the creation of two thesauri, one for enzymes and one for metabolites with the specific goal of allowing named entity identification for metabolic reaction extraction. Here synonymous named entities are mapped to a common identifier, in the case of the enzyme thesaurus these identifiers are EC numbers, in the case of the metabolite thesaurus KEGG metabolite identifiers. Both thesauri were created from existing data sources by a series of automated steps followed by manual curation. Compared to a previously published chemical thesaurus, created entirely with automated steps, our much smaller metabolite thesaurus performed on the same level for F-measure with a slightly higher precision. The enzyme thesaurus produced results equal to our metabolite thesaurus. The compactness of our thesauri permits the manual curation step important in guaranteeing accuracy of the thesaurus contents, whereas creation from existing resources by automated means reduces the effort required for creation. We concluded that our thesauri are compact and of high quality, and that this compactness has only a minor effect on recall.
In **Chapter 4** we studied the applicability and performance of a full parsing approach using the two thesauri described in **Chapter 3** for the extraction of metabolic reactions from scientific full-text articles. For this we developed a text mining pipeline built around a modified dependency parser from the AGFL grammar lab using a pattern-based approach to extract metabolic reactions from the parsing output. Results of a comparison to a previously published modified rule-based approach using three metabolic pathways from the EcoCyc database show a slightly lower recall compared to the rule-based approach, but higher precision. We concluded that despite its current recall our full parsing approach to metabolic reaction extraction has high precision and potential to be used to (re-)construct metabolic pathways in an automated setting. Future improvements to the grammar and relation extraction rules should allow reactions to be extracted with even higher specificity.
To identify potential improvements to the recall, the effect of text pre-processing steps on the performance was tested in a number of experiments. The one experiment that had the most effect on performance was the conversion of schematic chemical formulas to syntactic complete sentences allowing them to be analysed by the parser. In addition to the improvements to the text mining approach described in **Chapter**

4 I make suggestions in **Chapter 5** for potential improvements and extensions to our full parsing approach for metabolic reaction extraction. Main focus here is the increase of recall by optimising each of the steps required for the final goal of extracting metabolic reactions from the text. One of the discussed improvements is to increase the coverage of the used thesauri, possibly with specialist thesauri depending on the topic of the analysed literature. Another potential target is the grammar, where there is still room to increase parsing success by taking into account the characteristics of biomedical language. On a different level are suggestions to include some form of anaphora resolution and across sentence boundary search to increase the amount of information extracted from literature.

In the second part of **Chapter 5** I make suggestions as to how to maximise the knowledge gained from the text mining results. One of the first steps should be integration with other biomedical databases to allow integration with existing knowledge about metabolic reactions and other biological entities. Another aspect is some form of ranking or weighting of the results to be able to distinguish between high quality results useful for automated analyses and lower quality results still useful for manual approaches. Furthermore I provide a perspective on the necessity of computational literature analysis in the form of text mining. The main reasoning here is that human annotators cannot keep up with the amount of publications so that some form of automated analysis is necessary. Lastly I discuss the role of text mining in bioinformatics and with that also the accessibility of both text mining results and the literature resources needed to create them. An important requirement for the future of text mining is that the barriers around high-throughput access to literature for text mining applications should be removed. With regards to accessing text mining results, there is a long way to go for many applications, including ours, before they can be used directly by biologists. A major factor is that these applications rarely feature a suitable user interface and easy to use setup.

To conclude, I see the main role of a text mining system like ours mainly in gathering evidence for existing knowledge and giving insights into the nuances of the research landscape of a given topic. When using the results of our reaction extraction system for the identification of 'new' reactions it is important to go back to the actual evidence presented for extra validations and to cross-validate the predictions with other resources or experiments. Ideally text mining will be used for generation of hypotheses, in which the researcher uses text mining findings to get ideas on, as in our case, new connections between metabolites and enzymes; subsequently the researcher needs to go back to the original texts for further study. In this role text mining should become an essential tool on the workbench of the molecular biologist.

Samenvatting

Wetenschap heeft data in al zijn vormen als basis. Grootschalige data productie in de vorm van high-throughput experimenten vormen steeds meer de kern van de moleculaire biologie en de bioinformatica in het bijzonder. Naast de exponentiële groei van de experimentele data is ook de hoeveelheid wetenschappelijke publicaties nagenoeg exponentieel gestegen. Toegang tot de informatie in wetenschappelijke publicaties is nog steeds beperkt tot zoekmachines op het niveau van steekwoorden, titels en samenvattingen. Maar een groot deel van de kennis over met name biologische entiteiten en hun onderlinge relaties staat beschreven in de body van de publicaties, niet alleen in de samenvattingen. Deze data kunnen naar extractie worden gebruikt als bevestiging van bestaande inzichten of als basis voor nieuwe wetenschappelijke hypotheses. Dit maakt literatuur een waardevolle informatiebron voor wetenschappelijk onderzoek. Om deze kennis beter toegankelijk te maken is een eigen set aan technieken en benaderingen die rekening houd met de ongestructureerde vorm van vrije tekst vereist. Analoog aan het gebied van data mining voor gestructureerde data heeft zich voor ongestructureerde tekst het veld van tekst mining ontwikkelt en er zijn verschillende software tools in dit gebied gepubliceerd.

Dit proefschrift gaat over tekst mining specifiek in het vakgebied van de metabolomics. Metabolereacties zijn erg belangrijk in ons begrip van de metabole processen die plaats vinden in cellen en deze kennis vormt een belangrijke link tussen genotype en phenotype van organismen. De informatie over dit soort metabole reacties in bestaande databases is niet altijd compleet en dat maakt dit soort reacties een excellente toepassing voor tekst mining. Ons werk focust op de strategieën noodzakelijk voor de toegang tot grote collecties wetenschappelijke tekst en de tekst mining stappen benodigd om metabole reacties en hun bestanddelen, enzymen en metabolieten, uit de tekst te extraheren.

De inhoud van wetenschappelijke publicaties kan, voor verdere analyse, als platte tekst benaderd of in database systemen geladen worden. In het tweede hoofdstuk onderzoeken en bespreken wij de mogelijkheden en performance van XML database systemen voor het opslaan en ondervragen van heel grote collecties XML geformatteerde documenten in de vorm van het Medline corpus, een verzameling van meer dan 20 miljoen samenvattingen van wetenschappelijke artikelen. Het XML data format wordt veel gebruikt in de bioinformatica en vormt ook het hart van de meeste web services. Met de toegenomen hoeveelheid data in het XML formaat is ook de behoefte aan opslag en toegang tot dit soort data gegroeid. De XML database systemen in deze studie zijn beoordeeld op een aantal aspecten variërend van de technische vereisten tot gebruiksgemak en performance. De performance is gemeten met behulp van Medline documenten verzamelingen van toenemende grootte in combinatie met verschillende database queries. Een van de queries bepaald de mogelijkheden van elk database system in het doorzoeken van de tekst van elk document. Dit is belangrijk voor de toegang tot de inhoud van de teksten zonder

verdere tekst analyse. De resultaten laten zien dat alle database systemen goed kunnen omgaan met de kleine en gemiddeld grote datasets, maar dat de volledige verzameling Medline documenten een uitdaging blijft. De query mogelijkheden verschillen ook in grote mate tussen de verschillende systemen. Dit bracht ons tot de conclusie dat de performance en capaciteiten van de onderzochte database systemen verschillen, ook afhankelijk van de onderzoeksvraag. Er is geen systeem dat op alle vlakken beter is dan de andere. In plaats daarvan kunnen verschillende omstandigheden zorgen voor een andere optimale keuze. Sommige van deze scenario's worden in dit hoofdstuk verder uitgewerkt.

Een van de conclusies van hoofdstuk 2 is dat conventionele data mining technieken voor het natuurlijke taal deel van een publicatie uitsluitend werken voor simpele retrieval queries op basis van patronen. De natuurlijke taal in geschreven tekst is simpelweg te ongestructureerd en vereist specifieke tekst mining benaderingen. Deze benaderingen vormen het hoofdonderwerp van dit proefschrift. Er zijn twee hoofdtaken in tekst mining: named entity recognition, het identificeren van relevante entiteiten in tekst, en relation extraction, de identificatie van relaties tussen deze entiteiten. Voor beide taken zijn al vele verschillende technieken en benaderingen ontwikkeld. Wij gebruiken voor de identificatie van enzymen en metabolieten een woordenboek benadering (hoofdstuk 3) en voor de extractie van de metabole reacties een grammaticale analyse benadering (hoofdstuk 4).

In hoofdstuk 3 beschrijven wij de creatie van twee thesauri, een voor enzymen en een voor metabolieten, met het doel om named entity recognition toe te passen voor de extractie van metabole reacties. In deze thesauri zijn synonieme entiteiten gegroepeerd onder een gezamenlijke identifier. In het geval van de enzym thesaurus zijn dat EC nummers, in het geval van de metaboliet thesaurus KEGG metaboliet nummers. Beide thesauri zijn van bestaande bronnen gemaakt met een combinatie van automatische stappen gevolgd door manuele curatie. In vergelijking met een reeds gepubliceerde, geheel met automatische stappen gecreëerde, chemische thesaurus functioneert onze aanzienlijk kleinere metaboliet thesaurus op het zelfde niveau wat betreft F-measure met daarbinnen een iets hogere precisie. De resultaten voor de enzym thesaurus zijn van de zelfde orde van grootte als die van de metaboliet thesaurus. De compactheid van onze thesauri staat handmatige curatie stappen toe en deze zijn belangrijk voor de precisie en kwaliteit van de thesaurus inhoud. Aan de andere kant beperken de automatische stappen de tijd die benodigd is voor de creatie. Wij concludeerden dat onze thesauri compact en hoog kwalitatief zijn, en dat de compactheid maar een gering effect heeft op de recall.

In hoofdstuk 4 bestudeerden wij de toepasbaarheid en performance van een full parsing benadering, gebruikmakend van de twee thesauri uit hoofdstuk 3, voor de extractie van metabole reacties uit de volledige tekst van wetenschappelijke publicaties. Hiervoor hebben wij een tekst mining pipeline gebouwd, gebaseerd op een gemodificeerde dependency parser van het AGFL grammar lab en een patroon herkenning benadering om vervolgens de metabole reacties in de parsing uitvoer te herkennen. De resultaten van een vergelijking met een reeds gepubliceerde regel-

gebaseerde methode toegepast op drie metabole routes uit de EcoCyc database laten een licht lagere recall en licht hogere precisie voor onze methode zien. Wij concludeerden dat ondanks deze lagere recall onze full-parsing benadering voor de extractie van metabole reacties een hoge precisie heeft en de potentie heeft om in een automatische setting metabole netwerken te (re-)construeren. Toekomstige verbeteringen van de grammatica en de regels voor reactie extractie zouden zelfs nog hogere precisie mogelijk moeten maken.

Om mogelijke aanknopingspunten voor de verbetering van de recall te identificeren hebben wij het effect van een aantal tekstvoorbereidingsstappen bestudeerd met behulp van experimenten. Het experiment met de meest veelbelovende resultaten betrof de conversie van chemische formules naar syntactische constructies die door de parser konden worden geanalyseerd. Aanvullend aan de verbeteringen voor de tekst mining benadering zoals beschreven in hoofdstuk 4, doe ik in hoofdstuk 5 suggesties voor mogelijke verbeteringen en uitbreidingen van het full parsing systeem voor de extractie van metabole reacties. Het belangrijkste aandachtspunt is de verbetering van recall door elke stap op weg naar het einddoel te optimaliseren. Een van de besproken verbeteringen is het vergroten van het bereik van de thesauri, mogelijk ook in combinatie met het gebruik van specialistische thesauri afhankelijk van het onderwerp van de geanalyseerde literatuur. Een ander aanknopingspunt voor verbetering is de grammatica. Hier zijn nog verbeteringsmogelijkheden met het oog op de specifieke eigenschappen van biomedische taal. Op een ander vlak suggereer ik om een mogelijkheid voor het oplossen van anaforen toe te voegen. Dit heeft, in combinatie met zoekmogelijkheden over de grenzen van zinnen heen, de potentie de hoeveelheid uit de literatuur geëxtraheerde informatie te vergroten.

In het tweede deel van hoofdstuk 5 doe ik suggesties hoe de uit tekst mining resultaten verkregen kennis gemaximaliseerd zou kunnen worden. Een van de eerste stappen zou integratie van de gegevens met andere biomedische databases zijn om integratie van bestaande kennis over metabole reacties en biologische entiteiten mogelijk te maken. Een ander suggestie is een vorm van rangschikking van de resultaten om zo onderscheid te kunnen maken tussen hoge kwaliteit resultaten geschikt voor automatische analyses en lagere kwaliteit resultaten die nog steeds bruikbaar kunnen zijn voor handmatige gecontroleerde benaderingen. Verder geef ik mijn visie op de noodzaak van geautomatiseerde literatuuranalyse in de vorm van tekst mining. Belangrijkste aanleiding hiervoor is dat menselijke curatoren de vloed aan nieuwe wetenschappelijke publicaties niet kunnen bijhouden en dat daardoor een vorm van automatische analyse noodzakelijk is. Tot slot bespreek ik de rol van tekst mining in de bioinformatica en daarmee ook de toegankelijkheid van zowel tekst mining resultaten alsook de literatuur nodig voor hun creatie. Een belangrijke vereiste voor de toekomst van tekst mining is dat de barrières rond de hoge capaciteit toegang tot wetenschappelijke literatuur voor tekst mining doeleinden worden geslecht. Met betrekking tot de bruikbaarheid van tekst mining resultaten voor biologen is nog een lange weg te gaan voor de meeste toepassingen, inclusief de onze.

Een grote rol speelt hierbij het gebrek aan gebruikersvriendelijke interfaces en makkelijke installatie mogelijkheden.
Tot slot zie ik de hoofdrol voor een tekst mining systeem als het onze in het verzamelen van aanvullende informatie bij bestaande kennis en het vergaren van inzichten in de nuances in het onderzoekslandschap van een gegeven onderwerp. Wanneer de resultaten van ons reactie extractie systeem worden gebruikt voor de identificatie van 'nieuwe' reacties is het belangrijk om terug te gaan naar het daadwerkelijke bewijs voor validatie en om aanvullende bronnen te gebruiken ter bevestiging. Idealiter zal tekst mining worden gebruikt voor het ontwikkelen van nieuwe onderzoekshypotheses waarbij een onderzoeker tekst mining resultaten gebruikt voor het verkrijgen van ideeën over, zoals in ons geval, nieuwe relaties tussen metabolieten en enzymen. In vervolgstappen moet de onderzoeker dan terug naar de originele teksten voor nader onderzoek. Op deze wijze kan tekst mining een belangrijk stuk gereedschap voor een moleculair bioloog worden.

Acknowledgements

This is the place for me to say thank you to all the people who helped me to get here. I would like to apologise in advance to any one I might have inadvertently forgotten, I would like to thank you just as much.

This thesis, although ultimately written and defended by me, has come into being only with the help of others.

First and foremost, Jack. This work started with you taking me on as a PhD student. I would have liked to thank you for that. Working with you was always a pleasure, as well as all the diversions into life, the universe and everything. This project wasn't easy from the start, but I managed to get there in the end. Sadly we couldn't finish it together.

Paul, you came in as a knight in shining armour. Sometimes I wish you had come along sooner. With you for the first time since Jacks illness, I had someone who really understood the content of my work, and this was very much appreciated. Without that and you're kicks up my backside this book would never have gotten printed and I'd probably still be slogging away.

Ton, thank you for stepping in as promotor and trusting that this endeavour would work out. I would also like to thank you and Ernst for enabling Paul to spend time on my supervision.

Harm and Sandra, thank you as well, for the company, the discussions and the support. It was a pleasure sharing an office with you. Harm, also thanks for the IT support and software installations used in much of this work. Although the move broke up the gang of three, may be in the future I can share an office again with at least one of you.

I'm also indebted to all the other members, past and present, of both the chair group of Bioinformatics and the Applied Bioinformatics group. You know who you are and you made my stay a very enjoyable experience. Special thanks goes to Jan, for being the good spirit of both groups and my swimming buddy.

To my two paranymphs, Sandra and Rutger, thanks for being brave enough to sit on the stage with me.

To Rutger, Nynke, Yves, Iris, Martin, Petra, Saskia and Cathelijne, if you read this I'm really really finished!

Rolf, without your support this whole venture would have been a lot harder. Thank you very much for being there.

Curriculum Vitae

Judith Elisabeth Risse was born on February 2, 1979 in Dortmund, Germany.

In 1998 she passed the Abitur in Math, Physics, Dutch and Geography at the Freiherr vom Stein Gymnasium in Kleve, Germany.

Later that year she went to study Biology at Wageningen University and from 2002 Bioinformatics. In 2004 she interrupted her studies to organise the Annual Introduction Days of Wageningen University. As part of her MSc in Bioinformatics she worked on two Master theses. The first one she performed in a combined wet-lab/bioinformatics project with the topic of characterising the oxalic acid production of Aspergillus niger at the Fungal Genomics group of the laboratory of Microbiology at Wageningen University. The second Master thesis was done at the Applied Bioinformatics group of Plant Research International and involved the computational prediction of viral miRNAs targeting host genes. For her internship Judith worked in the IntAct group at the European Bioinformatics Institute in Hinxton, United Kingdom. Here she developed a web-application assisting scientific curators in the validation of text mining results for the BioCreative II challenge.

Judith graduated with an MSc in Bioinformatics from Wageningen University in August 2006.

In November of that year Judith started with her PhD project of 'Text Mining for Metabolic Pathways' under the auspices of Prof. Jack Leunissen in the Laboratory of Bioinformatics of Wageningen University.

From March 2014 Judith is working as Bioinformatics Data Analyst at Edinburgh Genomics of Edinburgh University.

List of Publications

Risse, Judith E., and Jack AM Leunissen. "A comparison of database systems for XML-type data." *In silico biology* 10, no. 3 (2010): 193-205.

Education Statement of the Graduate School
Experimental Plant Sciences

The Graduate School EXPERIMENTAL PLANT SCIENCES

Issued to: Judith E. Risse
Date: 7 April 2014
Group: Bioinformatics, Wageningen University & Research Centre

1) Start-up phase	date
► **First presentation of your project**	
Pathway reconstruction with the aid of textmining	Feb 15, 2007
► **Writing or rewriting a project proposal**	
Fatty acid pathway reconstruction with the aid of text mining	Feb 21, 2007
► **Writing a review or book chapter**	
► **MSc courses**	
► **Laboratory use of isotopes**	
Subtotal Start-up Phase	*3.0 credits**

2) Scientific Exposure	date
► **EPS PhD student days**	
EPS PhD student day, Leiden University	Feb 26, 2009
EPS PhD student day, Utrecht University	Jun 01, 2010
► **EPS theme symposia**	
EPS Theme 4 symposium 'Genome Biology', Leiden University	Dec 07, 2007
EPS Theme 4 symposium 'Genome Biology', Wageningen University	Dec 12, 2008
► **NWO Lunteren days and other National Platforms**	
NBIC/ISNB 2007	Apr 16-19, 2007
NBIC/ISNB 2008	Mar 05-06, 2008
BioRange Meeting, Arnhem	Oct 08, 2009
NBIC/ISNB 2009	Mar 17-18,2009
NBIC/BioRange 2010	Mar 30, 2010
► **Seminars (series), workshops and symposia**	
WUR Bioinformatics Day	Nov 13, 2006
WUR wide bioinformatics invited seminars	2006-2009
Machine Learning for NLP, Amsterdam	May 16, 2007
Bridging Ontologies and Textmining, EBI	Sep 12-14, 2007
WEES Seminars	2012-2013
► **Seminar plus**	
► **International symposia and congresses**	
BBC 2007, Leuven	Nov 12-13, 2007
BBC 2008, Maastricht	Dec 15-16, 2008
ISMB 2007 Vienna	Jul 21-25, 2007
ISMB 2007 Special Interest Group Meetings	Jul 19-20, 2007
ISMB 2009 Stockholm	Jun 27-Jul 02, 2009
► **Presentations**	
Presentation "Pathway Reconstruction and Textmining", UU	Sep 02, 2009
Presentation "Bioinformatics in Wageningen", HAN	May 28, 2008
Presentation "Pathway Reconstruction Progress" BioRange Meeting	Oct 08, 2009
Poster "XML database evaluation", BBC, NBIC	Nov 12-13, 2007
Poster "Pathway Reconstruction with Text Mining", NBIC	Mar 17-18,2009
► **IAB interview**	Dec 05, 2008
► **Excursions**	
Subtotal Scientific Exposure	*17.6 credits**

3) In-Depth Studies	date
► **EPS courses or other PhD courses**	
IP Math Biology, Paris	Jul 02-16, 2007
The Power of RNA-seq	Dec 16-18, 2013
► **Journal club**	
bioinformatics literature discussion	2006-2010
► **Individual research training**	
Subtotal In-Depth Studies	*7.9 credits**

4) Personal development	date
► **Skill training courses**	
Afstudeervak organiseren en begeleiden	Nov 08-09, 2007
Techniques for Writing and Presenting a Scientific Paper	Jun 29-Jul 02, 2010
► **Organisation of PhD students day, course or conference**	
Symposium Textmining for Dutch Genomics, Wageningen	Nov 23, 2007
► **Membership of Board, Committee or PhD council**	
Subtotal Personal Development	*2.8 credits**

TOTAL NUMBER OF CREDIT POINTS*	**31.3**

Herewith the Graduate School declares that the PhD candidate has complied with the educational requirements set by the Educational Committee of EPS which comprises of a minimum total of 30 ECTS credits

** A credit represents a normative study load of 28 hours of study.*

The research described in this thesis was financially supported by the Netherlands Bioinformatics Consortium (NBIC) as part of the BioRange SP 4.1.1 project.

This thesis was printed by Ipskamp Drukkers.

The cover depicts a word cloud of approximately the top 200 metabolites identified by us in the BioMed central corpus. The word cloud was generated with tagxedo (ww.tagxedo.com). The underlying pathway is a cutout from the human 'Metabolism of Carbohydrates Pathway' from wikipathways.org.