Next generation text-mining applied to toxicogenomics data analysis = Next-generation text-mining toegepast op toxicogenomics data analyse

Chapter 7

Summary and general discussion

Summary of main findings

Concept profiling is a thesaurus-based text mining technique that has been developed for literature-based discovery and analysis of gene expression data. The technique had however not been applied to toxicogenomics before, and was not integrated into a framework for gene set analysis. Toxicogenomics is a promising *in vitro* method that might reduce the use of laboratory animals in research. The biomedical part of the thesaurus used for generating the concept profiles needed to be adapted for text mining purposes, and the coverage of chemical concepts was insufficient. As outlined in **chapter 1**, we hypothesized that concept profiling could be applied when interpreting toxicogenomics data. In the first part of this thesis we describe how we adapted the biomedical part of the thesaurus for text mining purposes, and how we created and evaluated a thesaurus of chemical concepts. In the second part, we describe how we incorporated concept profiling into the statistical framework of the weighted global test (a gene set analysis method), and further generalized the technology to be used together with other gene set analysis methods for interpreting toxicogenomics data.

The main findings of these investigations are described below.

The experiments described in **chapter 2** aimed at making the biomedical part of our thesaurus more useable for text-mining purposes. We hypothesized that this could be done by removing and adding synonyms to the thesaurus, and implemented a number of rewrite rules and suppress rules for this purpose. When we manually evaluated the impact of the rules on a MEDLINE corpus, we noted an increase of 2.8% in the number of terms and an increase of 3.4% in the number of concepts recognized when using the rewrite rules. When applying the suppression rules, thousands of undesired terms were suppressed in the corpus and the thesaurus was cut back with 25% in megabyte, positively influencing the performance of the concept identification software. We concluded that applying the five rewrite rules and seven suppression rules that passed our evaluation would positively influence the performance of biomedical term identification of MEDLINE abstracts when the UMLS is to be used as a source of concepts. A software tool to apply these rules to the UMLS is freely available at http://biosemantics.org/casper.

In **chapter 3** we merged multiple chemical databases, and evaluated their individual performance as well as the performance of the merged chemical thesaurus named Jochem (JOint CHEMical dictionary) in terms of recall and precision on a manually annotated corpus. We adapted the rules for rewrite and suppression described in **chapter 2** to fit chemical terms and tested how the application of these rules influenced the performance. In addition, we evaluated the impact of the use of disambiguation rules and limited manual curation (in terms of manual inspection of frequent terms). We concluded that all these automatic or semi-automatic curation actions increase precision with a minor loss of recall.

After creating and evaluating the combined chemical thesaurus described in **chapter 3**, it remained to be investigated which impact an extensive manual curation would have on chemical term identification in text. In **chapter 4**, we therefore applied the same automatic and semi-automatic curation actions (rule-based term filtering, semi-automatic manual curation, and disambiguation rules) that we had investigated in **chapter 3,** to ChemSpider, a manually curated multi-source chemical database. We noticed that also for ChemSpider, our curation actions were needed to achieve a high precision. After applying the curation actions, ChemSpider achieved the best precision but our chemical thesaurus Jochem had a higher recall.

In **chapter 5**, we set out to create concept profiles with our updated thesaurus with the aim to integrate these into the statistical framework of the weighted global test (a gene set analysis method). The weights are the concept profile matching scores. We showed that the concept profile matching scores reflect the importance of a gene for the concept of interest (for example a Gene Ontology category), and that literature-based associations provided a deeper insight into the gene expression experiment compared to

an analysis using classical Gene Ontology-based gene sets. We also demonstrated the possibilities of the literature-weighted global test for linking of gene expression data to patient survival in breast cancer and the action and metabolism of drugs.

In **chapter 6** we further explored the use of concept profile matching for gene set creation and its application in toxicogenomics. Using our thesaurus, we were able to create 25 times more chemical response-specific gene sets using text mining than what was possible using a method based on chemical-gene interaction information in the Comparative Toxicogenomics Database. By testing the differential expression of the text mining-derived gene sets in data sets of chemicals from experimental models we demonstrated that we could predict the chemical treatment, and by using three different gene set testing methods to evaluate the gene sets we demonstrated that our method is generalizable. We also demonstrated that gene sets created using concept profile matching could be used to identify embryotoxic effects of triazoles already at the gene expression stage, and discriminate triazoles from other chemicals in a principal component analysis.

Based on the findings reported in this thesis, we conclude that concept profiling can be integrated in the framework of gene set testing and as such be used to relate chemical information to gene expression data, identify toxic effects already at the gene expression stage, and discriminate between compound classes.

General discussion

In this section we highlight further points for discussion based on the contents of chapter 2-6. We then direct our focus to specific attention areas in biomedical text mining and its application to toxicogenomics, and end with conclusions based on the main outcomes of this thesis and implications for future work.

Building and evaluating a thesaurus of domain-relevant concepts

In this section we focus on further discussion points related to the research conducted in the first part of the thesis, namely chapter 2, 3 and 4. The identification of domain-relevant terms in natural language is essential for biomedical text mining. Naturally, the success of the thesaurus-based approach depends on the coverage of terms in the thesaurus for the particular domain and how well the terms are suited for natural language processing. Genes, chemicals, pathways and toxicological endpoints are obviously important concept categories when applying text mining in toxicogenomics. In this thesis, the thesaurus used to find concepts in text is composed of four parts: biomedical concepts, genes, chemicals, and concepts related to toxicology. Previously, a lot of effort had gone into the creation and evaluation of the gene part of the thesaurus [51, 139, 141, 236], while the biomedical part, made up of the UMLS, had not been adapted for text mining purposes. Also, the UMLS contained some chemical and toxicological concepts, but nothing was known about the coverage or performance of these types of concepts.

Biomedical concepts: preparing the UMLS Metathesaurus for text mining

In **chapter 2**, we describe how we set out to adapt the UMLS version available at that time (2007AA) for concept identification in text. Earlier experience with creating concept profiles for genes combined with the results from our investigations described in **chapter 2** resulted in a protocol for building the UMLS part of the thesaurus, which we will here outline and discuss.

First, we gather the synonyms and the definition for each concept, and record the place of the concept in the semantic hierarchy of the UMLS. Then we execute a number of rewriting and suppression rules based on term structure, and perform a manual analysis step of the top 250 terms from a MEDLINE-indexation using Peregrine. Next, terms in the thesaurus are checked for in-thesaurus homonyms, and the 250 terms with the most homonyms are inspected manually. As a final step, terms that are not found

when performing a whole MEDLINE-indexation using Peregrine are removed from the thesaurus. This is done for efficiency purposes.

The time-consuming and biased manual curation of most frequent terms and homonyms is a drawback of the protocol described above. A list of terms and concepts to remove is kept and updated every time a new version of the thesaurus is created. There are however no guidelines as to which terms within the 250 most frequent terms should be removed. Sometimes the choice may seem obvious, such as the removal of common English words, but one should keep in mind that for some biomedical concepts these types of words are actually a correct synonym and even though the overall precision will increase, the recall for that specific concept will actually go down. Thus, the balance is delicate and there is a need for guidelines regarding how to judge the "correctness" of a synonym already at the thesaurus creation process. Increased transparency would also help. One way to provide transparency is to provide access to not only the concept name but also all synonyms attached to that concept in applications where the thesaurus is used. For example in Anni it is possible to go directly to PubMed via a hyperlink and read the abstracts annotated to a specific concept. The actual concept occurrence is however not marked in the abstract since PubMed does not allow that. A possibility would be to show the abstract with the annotated concepts and its synonyms in a viewer different from PubMed, similar to the iHOP interface [237]. If a scientist using Anni had access to all synonyms for a concept, or even better would know which particular synonym of a concept that was actually found in the text, he or she would be much better equipped to judge if he or she were looking at a true or false positive. If the scientist would then be able to edit the synonym list, for example in a similar way as has been implemented for chemical names in ChemSpider [238], the feedback loop would be complete.

Although the term rewrite and suppress rules together with the manual curation steps have been shown to increase the number of times a term is found in text and to suppress erroneous terms, we have only evaluated these steps for source vocabularies having the lowest restriction level (the UMLS has five grades of restrictions, because some sources are subject to costs and other restrictions such as use only in the USA). If higher-level source vocabularies are added, one should keep in mind that the impact of the term rewrite and suppress rules and the manual curation steps might be different. When the work in **chapter 2** was published we suggested that the impact of the rules should be tested on another corpus than MEDLINE, for example electronic patient records. This has since then been done by Roque and coworkers [239]. They used a thesaurus-based text mining approach to extract information from the free text part of Danish electronic patient records. The thesaurus used by the authors was based on the Danish translation of the WHO International Classification of Diseases (ICD10) and they augmented existing terms with variants as described in **chapter 2**, and in work by Hersh et al. [240]. They noticed that generated term variants were responsible for 24% of the total number of hits in the records. It is clear that term variant generation greatly increase the number of hits in electronic patient records, much more so than what we found for MEDLINE (2.8% more terms and 3.4% more concepts). Since the authors do not report the performance per rule, this difference in performance is difficult to explain, but might be due to the different structure of MEDLINE abstracts compared to the free text part of electronic patient records. It might also be the case the terms in the ICD10 dictionary are specifically good examples of terms that need to be rewritten in order to be found in free text.

Chemical concepts: combining public online databases

Before the work done in this thesis, studies describing dictionary-based chemical text mining based on public resources had reported disappointing performance figures (see **chapter 1**), and the chemical part of the UMLS had not been tested on an annotated corpus. To investigate if the UMLS alone would be sufficient as a resource for chemical and toxicological concepts, we performed a small study in which we indexed two full-length toxicogenomics-focused articles using our current version of the thesaurus that included the UMLS and a gene thesaurus, and let a toxicologist (Rob Stierum, PhD) check the results. The toxicologist indicated many missing concepts, and was concerned with

the fact that the chemical names in the thesaurus were not linked to any identifier such as a CAS number or InChI string. We noticed that a few of these "missing" concepts were actually in the thesaurus, but were not recognized due to the current implementation of the indexing engine Peregrine. Peregrine had only been designed to recognize gene and protein names, and the complicated nature of many chemical names caused them to be either incorrectly recognized or not recognized at all. For example, the exact placement of tokens such as commas, spaces, hyphens, and parentheses plays a much larger role for chemical names than for gene names. This initial study led us to believe that a larger chemical thesaurus was needed, and that Peregrine needed to be tuned to work better with chemical names. We therefore set out to create and evaluate a chemical thesaurus and to adapt Peregrine, as described in **chapter 3** and **chapter 4**.

The work in **chapter 3** and **chapter 4** resulted in a protocol for creating and evaluating a thesaurus of chemicals. We will briefly describe and discuss this protocol below. In short, the different chemical vocabularies are downloaded locally and concepts are extracted together with their synonyms, definition and links to online databases if available. Concepts from the different vocabularies are merged if they have the same CAS number, InChI string or online database link. Before and after the merging of concepts, the chemical thesauri are processed by applying slightly modified versions of the rewrite and suppress rules described in **chapter 2**. The resulting chemical thesaurus Jochem is manually curated by removing frequent terms and homonyms in a similar way as for the UMLS that we described in the previous section about biomedical concepts. In contrast to the research around the UMLS in **chapter 2** where we could not evaluate the rewrite and suppress rules on an annotated corpus and thus not provide a recall and precision value for the biomedical part of our thesaurus, the access to an annotated corpus of chemical entities [111] made it possible to do exactly this for the chemical thesaurus. Even though we were able to achieve a reasonable performance in terms of recall and precision on the chemical entity corpus using the thesaurus together with the Peregrine tagger, the process of downloading, cleaning and merging the different dictionaries is time consuming and error-prone. In addition, when putting the thesaurus into practical use, we have noted that the merging of concepts based on CAS numbers, InChI string or online database links can result in chemical concepts with tens of different CAS numbers and InChI strings due to the different levels of granulation used by the CAS and InChI systems and by the individual chemical vocabularies. Also, errors originating from the source vocabularies will propagate to the merged thesaurus. Our chemical thesaurus Jochem does not focus on the chemical structure but on chemical names and database identifiers (the InChI strings are not used when mining the text, only for merging purposes). Obviously, the chemical names are presumed to link to the structure.

The need for better quality chemical databases was addressed in two recent publications by Williams et al. [241, 242] where they also provided suggestions on how to improve the quality. Williams and coworkers suggest a combination of manual curation, possibly with the help of crowdsourcing (a strategy that combines the effort of the public to solve one problem or produce one particular thing), and automated mechanisms to ensure structures and data are correct. Williams and coworkers have implemented these steps for the ChemSpider database and claim to have managed to address thousands of inherited errors. We tested the curated part of the dictionary of chemical names behind ChemSpider in **chapter 4**. When correcting for errors in the corpus, ChemSpider had a precision of 91% compared to 82% for Jochem (chapter 4), showing the benefit of the curation steps in ChemSpider when it comes to quality. The recall for ChemSpider was however much lower (19%, compared to 40% for Jochem), confirming the trade-off between quality and quantity. Another recent database that claims to deal with inherited errors is the database behind the 'NPC browser' from the NIH Chemical Genomics Center [243], but when tested for quality by Williams and coworkers [241, 242] multiple errors were found. Williams and coworkers analyzed the structures for a random selection of 50 of the top selling US drugs as represented in the NCGC database and found that 40% were incorrect. Clearly more effort is needed to ensure the quality in public chemical databases. A related study comparing human metabolic pathway databases [244] also emphasize the need for standardizing

metabolite names and identifiers, and stresses that the conceptual differences between the databases should be resolved. The recently published database MetRxn [245] claims to have dealt with these problems using for example chemical structure analysis procedures during the matching process, but their system is not aimed at text mining and has not been tested on an annotated corpus. An initiative aiming at integrating available data resources is the OpenPHACTS (Open Pharmacological Concepts Triple Store) [246]. The ChemSpider database is serving the chemical services to the project and OpenPHACTS has agreed on the need for a set of structure standardization rules that will be used to process all incoming chemical compounds. Future will tell if this approach proved successful.

There have been a few other studies regarding thesaurus-based identification of chemical names published after the works in **chapter 3** and **chapter 4** were published, and these focus on using one source for the chemical names instead of combining many. Zhang and coworkers presented a system based on the chemical part of MeSH, which performed comparable to the results we presented for MeSH in **chapter 3**. The recently published Compounds In Literature (CIL) system [247] uses the pre-processing steps described in **chapter 3** together with a self-generated stop word list for compound synonyms when adapting the PubChem dictionary to screen for compounds and relatives in PubMed. A notable difference is that the authors of CIL only used the first five synonyms of each compound. CIL achieved a lower precision (52%) but higher recall (72%) compared to the results based on PubChem described in **chapter 3** (precision: 73%, recall: 35%). Clearly, the precision reported for the CIL system needs improvement, underlining the need for better disambiguation of chemical names. The authors did not report using the disambiguation rules described in **chapter 3**, and they used a different indexer than Peregrine. It would be interesting to see how their dictionary performs when using Peregrine, which implements the disambiguation rules from **chapter 3**. Their dictionary is publicly available at ftp://ftp.ncbi.nlm.nih.gov/pubchem/Compound/Extras/.

In **chapter 3**, we suggested that thesaurus-based and machine learning methods should be combined for a better performance and underlined the need for name-to-structure mapping. This approach was later taken by Nobata and coworkers [248] for metabolite names, and they showed that combining a thesaurus-based named entity recognition system with a system that learns from linguistic cues using an annotated corpus results in increased performance of metabolite name recognition. Interestingly, they mapped the names recognized in text against ChemSpider to identify appropriate structures. They identified structures for 55% of their unique names, and also found many real yeast metabolites among unmatched names, which made good candidates to extend metabolite databases. It would be interesting to see how such an approach would work for other types of chemicals than metabolites. Such a system could for example make use of the recently published Open-Source Chemistry Analysis Routines (OSCAR) software version 4, a toolkit for the recognition of named entities and data in chemistry publications [249].

The master thesaurus

To supply our thesaurus with toxicology-related concepts and terms, we converted the IUPAC glossary of terms used in toxicology to our thesaurus format. When forming the master thesaurus, the UMLS, gene, chemical and toxicity thesauri are merged based on term overlap and a number of patterns for recognizing gene and protein names. The different steps are performed by a series of coupled java scripts. For the latest release of Anni (2.1) [250], we used the new improved master thesaurus to make the concept profiles. Continuing our Flusilazole example from the Introduction where a search on the term in the CTD only retrieved two associated genes, matching the concept profile of Flusilazole with all human genes in Anni 2.1 not only adds 13 genes that co-occur with Flusilazole in the literature to the ones also found in the CTD, but also gives suggestions of other, probable gene interactions.

Even though the master thesaurus seems to work satisfactory and there seems to be no pressing need to change the content of it, the merging process is time-consuming and not very flexible. It would be interesting to contrast the process of creating, and the final performance of, the master thesaurus against a thesaurus based on ontologies from the Open Biological and Biomedical Ontologies (OBO) Foundry [251]. The OBO Foundry is a collaborative experiment involving developers of science-based ontologies who are establishing a set of principles for ontology development with the goal of creating a suite of orthogonal, interoperable reference ontologies in the biomedical domain. The OBO Foundry provides discussion fora, technical infrastructure, and other services to facilitate ontology development such as mappings between, logical definitions for, bridging, and relations for combining, ontologies. One would presume that the guidelines presented by the OBO Foundry would indeed make the thesaurus-creation process faster and less error-prone. Actually, 32 of the 161 UMLS vocabulary sources (statistics are from the 2011AB release) are already available in the library of ontologies called BioPortal [252], which is supplied by the OBO Foundry. The BioPortal currently (6 April, 2012) contains 297 ontologies. The UMLS vocabularies are assigned to the Ontology Group "UMLS" in the BioPortal, and include vocabularies commonly used in text mining such as the Gene Ontology, MeSH, and the Online Mendelian Inheritance in Man (OMIM). Unfortunately, for chemicals and genes/proteins, only a small number of dedicated ontologies are available on BioPortal. The largest chemical ontology available in the category "Chemical" is the Chemical entities of biological interest (CHEBI) ontology, but with only 31,470 terms it hardly covers the chemical domain. The largest included ontology for gene names seems to be the Human Genome Organisation (HUGO) ontology with 32,917 terms. It therefore seems likely that, at this point in time, the information about genes and chemicals contained in BioPortal does not match up to the information about these entity classes currently contained in the master thesaurus. However, for other biomedical concepts it can provide additional sources of concepts next to the UMLS, and may even be able to replace the UMLS part of the master thesaurus since many of the commonly used vocabularies are already included in the BioPortal. Again, research contrasting the creation and performance of the master thesaurus against a thesaurus based on the resources available in the BioPortal would be valuable.

Using concept profile technology to create gene sets relevant for toxicogenomics

In this section we focus on further discussion points related to the research conducted in the second part of the thesis, namely chapter 5 and 6. In **chapter 5** we introduced the literature-weighted global test that uses concept profile matching scores to weigh the contribution of the genes in a gene set. Even though the literature-weighted global test works well, it is not particularly user friendly for biologists. It has a command line interface and knowledge of the bioconductor framework and the statistical language R is desired. It requires the loading of large amounts of data at the same time, which makes it slow in comparison to the "normal" global test (without the weights). Also, the process of updating the concept association scores is non-trivial. A conversion of the software package to a web service environment, preferably with a web interface, would greatly enhance its usability. In addition, if the loading of large data sets and the computations could be placed in a distributed environment, the speed of use would increase. We expect that a more flexible environment for updating the concept profile association scores, maybe based on linked data, and a transfer of the technology to a web service environment, would make it easier to keep the concept profile association scores up-to-date with the current literature.

In **chapter 6** we further generalized the concept profile-based gene set creation method to be used with other gene set analysis tools. We described how to create such gene sets by matching concept profiles, and provided chemical response-specific gene sets for download in a generic format. The gene sets however capture the information available in the literature at that specific point in time, and the same arguments regarding the updating of the concept profile association scores in **chapter 5** apply to

these gene sets. When investigating the differences between the gene sets created using text mining and the CTD in **chapter 6**, we noticed that some of the genes that were missing in the text-mining based gene sets but present in the CTD-based gene sets came from tables or supplemental material listing differentially expressed genes from a gene expression experiment. This indicates that the text-mining based gene sets would benefit from using information from the full text and supplements instead of only abstracts. One could also consider mining information from other resources than the scientific literature, including wiki's such as the Wikipedia [253], the ConceptWiki [254], and the WikiPathways [255], drug labels [256], and databases such as the pharmacogenomics database PharmGKB [257] and GeneCards [258]. In contrast, some gene-chemical interaction information was missing from the CTD, causing the text-mining based gene sets to perform better in some cases and demonstrating the limitations of manual curation. A combination of both approaches might prove beneficial and is something that could be investigated further.

The text-mining based gene sets are created using a concept profile length cutoff of 200 concepts, and a cutoff of 1000 genes from the concept profile matching procedure. These cutoffs were empirically determined, and might be suboptimal. Further investigation of the best cutoffs should include detailed analysis of the relation between the concept profile matching scores between genes and chemicals and gene expression levels induced by a chemical treatment.

Another issue is the specificity of the text-mining generated gene sets. The fact that they are based on the whole of MEDLINE can be seen as both an advantage and a disadvantage. It can be an advantage because chemical-gene interaction information can be present in other types of journals than chemical-focused ones, but on the other hand more noise might be expected. It would be interesting to try to build the concept profiles of the chemicals on a limited corpus of chemical-related documents, and to include patents. This could be accomplished for example by mining patent collections [259], or the citations and patents linked to compounds in the ChemSpider database. Adding patent information would also make the concept profiles even more up-to-date since patent information usually becomes available in the scientific literature only at a later stage. Recent developments in chemical text mining aimed at patents include the use of finite state machines to encode the rules used for systematic naming, effectively creating an infinite dictionary [260]. Such a finite state machine covers the vast majority of systematic chemical names likely to be found in medicinal chemistry papers or pharmaceutical patents. The drawback is that such grammars encode only the syntax of the chemical structure naming rules but not the semantics, allowing the finite state machine to also accept chemically nonsense strings. This requires the use of name-to-structure conversion tools to filter out such false positives. It would be interesting to see how such a method would perform if combined with a thesaurus of non-systematic names.

The gene sets for embryotoxicity described in **chapter 6** proved useful for detecting the embryotoxic properties of triazoles at the gene expression stage, and for discriminating triazoles from other compounds based on gene expression changes. Further study of the performance of these embryotoxicity gene sets for other compound classes forms a topic for future research, as does creating and testing text-mining based gene sets for other types of toxicities such as carcinogenicity and neurotoxicity.

Special attention areas

This section highlights areas needing special attention. These areas were identified during the whole research track.

The need for annotated corpora

The rewrite and suppression rules described in **chapter 2** were not tested on an annotated corpus, simply because at the time there was no large corpus annotated with a variety of biomedical entities available. Such a corpus would be of great value,

because the impact on recall and precision could then be tested before and after execution of the rules. Many manually annotated corpora exist (see for example [261]) but they are either limited to one specific domain (e.g. cancer), or the entities that are annotated are limited to only one type of concept (e.g. genes or diseases), or they are small (hundreds of abstracts). Due to the high costs involved, there are few initiatives aiming at producing a large and broad manually annotated corpus of biomedical and chemical entities. In **chapter 3**, we pointed out many problems with the corpus of annotated chemical entities that was used to test the chemical thesaurus and stressed the need for other efforts. Crowdsourcing might be of use here [262], and one could imagine a call for scientists to annotate a large number of documents for biomedical entities in a similar way as the call that came out in 2008 for scientists to participate in the community annotation of Wikiproteins [263]. An example of a tool directed at supporting such annotation efforts is BioNotate-2.0 [264]. BioNotate-2.0 also builds upon the Semantic Web, facilitating the dissemination of annotated facts into other resources and pipelines.

The Collaborative Annotation of a Large Biomedical Corpus (CALBC) project [265] partners approach the problem from a different angle. The project organized two public challenges, with the aim to produce a large-scale annotated biomedical corpus with four different semantic groups through the harmonization of annotations from automatic text mining solutions [266], thus sidestepping manual curation. The four semantic groups are chemical entities and drugs, genes and proteins, diseases and disorders, and species. The final annotated corpus contained about 1,000,000 MEDLINE abstracts. Kang and coworkers showed that this corpus can be a viable alternative for, or a supplement to, a manually annotated corpus when training NLP software in a biomedical domain [267].

Resolving ambiguities

Word Sense Disambiguation (WSD) is the task of automatically identifying the appropriate sense (or concept) of an ambiguous word based on the context in which the word is used. The disambiguation procedure that we used in this thesis (described in **chapter 3**) is knowledge-based and thus based on the knowledge source (in our case the thesaurus) and the textual context in which the word is found (in our case the MEDLINE abstract). Many other WSD methods exist, and all have their advantages and disadvantages (for a recent comparison of methods, see [268]). WSD using statistical learning approaches actually achieve better performance than knowledge-based methods [35, 268]. On the other hand, statistical learning approaches require manually annotated training data for each ambiguous word, which is an infeasible task for a large resource such as the thesaurus used in this thesis. WSD is an active research field and recent advances include incorporating different types of collocations (a sequence of words or terms that co-occur more often than would be expected by chance) in the disambiguation algorithms [269], and crowdsourcing [263, 270]. The impact of these techniques on the quality of concept profiles is unknown, and a topic for future research.

Concept profiles and linked data

The current process for creating concept profiles (as described in **chapter 1** and **chapter 5**) has proved to be functional, but has its limits. Concept profiles are based on information from the scientific literature, but there are also other information sources such as patents and web resources that could provide useful information. Also, the current implementation for generating and using concept profiles is not very flexible with regards to adding and removing knowledge sources, making it time-consuming to adapt the technology to a new domain. In contrast, a technology that is known for its flexibility is the Semantic Web [271]. The Semantic Web is a collaborative movement led by the World Wide Web Consortium (W3C) that promotes common formats for data on the World Wide Web. By encouraging the inclusion of semantic content in web pages, the Semantic Web aims at converting the current web of unstructured documents into a "web of data". It builds on the W3C's Resource Description Framework (RDF). One could imagine querying relevant resources on the Semantic Web and storing the information

from these queries in a graph database, which could then be used to infer relationships between concepts. This would also enable relations further away than the current concept profile implementation based on Swanson's ABC model described in **chapter 1** (briefly, the model states that if 'A influences B' and 'B influences C', then 'A may influence C'). If the concept profile generation and analysis process could be translated into a workflow with nested, but independent, components, and implemented in a workflow management system such as for example Taverna [272], this would allow for even more flexibility. Components of the concept profile generation and analysis process such as the text resource or specific indexing engine, and parameters such as the specific statistic used for concept profile matching, would be more transparent and easier to manipulate. To support in-silico experimentation, Taverna contains a suite of tools used to design and execute scientific workflows, that together with BioCatalogue [273] (a curated catalogue of life science Web Services) and myExperiment [274] (a virtual research environment for sharing workflows) forms a foundation to store, interpret, analyse and network data to other work groups. Integrating the concept profile generation and analysis process into such a framework would probably lead to increased speed, flexibility, collaboration and visibility.

Examples of initiatives that have tackled the problem of linking biological data and drug data respectively using RDF are Bio2RDF [275] and LODD [276]. Combining these resources, Chen and coworkers created a new semantic systems chemical biology resource (Chem2Bio2RDF), and demonstrated its potential usefulness in specific examples of polypharmacology, multiple pathway inhibition and adverse drug reaction to pathway mapping [277]. A comparison of the relations found using concept profile technology against relations found using Chem2Bio2RDF would provide more insight into the benefits and pitfalls of the two technologies. One could for example expect that the quality of the data in peer-reviewed scientific publications is higher than non peer-reviewed Semantic Web content, but that remains to be investigated.

Application of concept profiling to toxicogenomics

In **chapter 1** we introduced four different areas in toxicogenomics where conventional bioinformatics solutions might be assisted by concept profiling: class discovery and separation, connectivity mapping, mechanistic analysis, and identification of early predictors of toxicity. The work described in **chapter 5** and **chapter 6** and further discussed previously in this chapter (section "Using concept profile technology to create gene sets relevant for toxicogenomics") shows that concept profiling can aid bioinformatic approaches in all these areas. The characteristics of the data from a toxicogenomics study, such as time series analysis, dose-response relationships, and the use of multiple compounds for comparison showing very small and/or early gene expression changes, make the analysis complicated but this is not something that in our opinion influences the concept profiling technique itself. In our experience, concept profiling can help toxicogenomics data interpretation, just like concept profiling can help interpret data from other omics areas. All omics areas have their specific features, and domain adjustments will always be needed with regards to thesaurus content and corpus selection, but the concept profile technology does not seem to be domain-dependent.

Concluding remarks

We have enhanced the concept profile creation pipeline by greatly improving the performance of biomedical and chemical concept identification in text, and made these results available not only via scientific publications but also by implementing these improvements in a new release of the Anni tool (2.1) for text-mining based knowledge discovery. We have shown that concept profiling can be integrated in the framework of gene set testing and as such be used to relate chemical information to gene expression data, identify toxic effects already at the gene expression stage, and discriminate between compound classes.

Future works surrounding the master thesaurus used in this thesis include the development of guidelines for manual curation, and a comparison of the creation process and the performance of the master thesaurus with a thesaurus based on the ontologies in the BioPortal. With regards to the identification of biomedical and chemical terms in text in general, the creation of large-scale corpora for benchmarking, and the disambiguation of entities remains a challenge.

A comparison between the current implementation of the concept profile creation pipeline and a Semantic Web approach is a topic for future research, as well as a benchmark procedure for measuring general concept profile quality and the impact of concept profile length on the concept profile matching score. More research is needed to ensure and measure the quality of the information in the chemical concept profiles. Ways to improve the information quality could include limiting the corpus used for chemical concept profile generation to chemical-specific information only, and to include patents in the corpus. The concept profile is only as good as the underlying data sources, and continuing efforts to ensure the quality in public chemical databases is essential.

Regarding the use of concept profiles for the generation of gene sets, it would be interesting to investigate whether a combination of manually curated information as provided by for example the CTD and the concept profile generated associations lead to improved performance. Also, the potential of the embryotoxicity-specific gene sets to detect embryotoxic signals already at the gene expression stage could be explored further by testing these gene sets on more compounds and compound classes. It would also be interesting to see if text-mining based gene sets for other types of toxicities than embryotoxicity, such as carcinogenicity or neurotoxicity, can be used to detect early signals of these types of toxicities as well.

Samenvatting (Summary in Dutch)

Concept profiling is een op thesauri-gebaseerde text-mining techniek die is ontwikkeld voor het ontdekken en analyseren van gen-expressie data op basis van reeds bekende literatuur. Deze techniek is niet eerder gebruikt in toxicogenomics en was ook niet geïntegreerd in een framework voor het doen van gen-set analyse. Toxicogenomics is een *in vitro* techniek die tot vermindering en vervanging van dierproeven kan leiden. Het biomedische gedeelte van de thesaurus die wordt gebruikt om de concept profielen te maken moest worden aangepast om te worden gebruikt voor text-mining. Verder was ook de dekking van chemische concepten onvoldoende. Zoals beschreven in **hoofdstuk 1** gingen we er vanuit dat concept profiling bruikbaar zou zijn voor het interpreteren van toxicogenomics data. In het eerste gedeelte van dit proefschrift laten we zien hoe we het biomedische gedeelte van de thesaurus hebben aangepast voor gebruik in text-mining, maar ook hoe we de thesaurus voor chemische concepten hebben gemaakt en geëvalueerd. In het tweede gedeelte beschrijven we hoe we concept profiling hebben ingepast in het statistische framework van de weighted global test, die wordt gebruikt als gen-set analyse methode. Daarnaast beschrijven we hoe we de technologie hebben gegeneraliseerd zodat het samen met andere gen-set analyse methodes gebruikt kan worden om toxicogenomics data te interpreteren.

Hieronder staan de belangrijkste bevindingen van het onderzoek.

De experimenten in **hoofdstuk 2** waren er op gericht om het biomedische gedeelte van onze thesaurus aan te passen voor het gebruik in text-mining. We namen aan dat dit mogelijk zou moeten zijn door synoniemen toe te voegen en te verwijderen. Daarvoor implementeerden we een aantal herschrijf- en suppressieregels. Een manuele evaluatie van de impact van de regels op een MEDLINE corpus liet een 2.8% toename van het aantal herkende termen zien en een 3.4% toename van het aantal herkende concepten door het gebruik van de herschrijfregels. De supressieregels onderdrukten duizenden ongewenste termen in het corpus die daardoor 25% in megabytes kleiner werd, wat een positieve invloed had op de prestaties van de concept identificatie software. Onze conclusie was dat het gebruik van de vijf herschrijf- en zeven suppressieregels, die onze evaluatie doorstonden, een positieve invloed heeft op de prestatie van de biomedische term identificatie van MEDLINE abstracts met UMLS als de bron van de concepten. De software om deze regels toe te passen op de UMLS is vrij beschikbaar via http://biosemantics.org/casper.

In **hoofdstuk 3** hebben we meerdere chemische databases samengevoegd. We hebben geëvalueerd of deze samengevoegde chemische thesaurus Jochem (Joint CHEMical dictionary) beter presteerde op een handmatig geannoteerd corpus, gelet op recall en precisie, dan de individuele databases. We hebben de herschrijf- en suppressie regels uit **hoofdstuk 2** aangepast voor chemische termen en geëvalueerd wat de toepassing van deze regels voor effect had op de prestaties. Verder hebben we gekeken wat de invloed is van het gebruik van disambiguatie regels en beperkte manuele curatie (handmatige inspectie van veelvoorkomende termen). Onze conclusie was dat iedere handmatige en semi-handmatige curatie de precisie verhoogde met een minimal verlies van recall.

Na het maken en evalueren van de gecombineerde chemische thesaurus in **hoofdstuk 3** moest er nog gekeken worden naar de impact van een uitgebreide handmatige curatie op de identificatie van chemische termen in tekst. Daarom hebben we in **hoofdstuk 4** op ChemSpider, een handmatig gecureerde samengestelde chemische database, dezelfde handmatige en semi-handmatige curaties (regel gebaseerde termen filtering, semi-automatische handmatige curatie en disambiguatie regels) uitgevoerd als in **hoofdstuk 3**. We zagen hier dat ook voor ChemSpider onze curaties nodig waren om een hoge precisie te krijgen. Na het toepassen van de curaties, haalde ChemSpider de beste precisie, maar onze chemische thesaurus Jochem haalde een hogere recall.

In **hoofdstuk 5** hebben we concept profielen gemaakt met onze geupdate thesaurus om deze te integreren in het statistische framework van de weighted global test. De gewichten die hiervoor zijn gebruikt zijn de concept profiel overeenkomst scores (matching scores). We laten zien dat concept profiel overeenkomst scores het belang aangeven van een gen voor het doelconcept (bijv. een Gene Ontology categorie). Daarnaast laten we zien dat associaties op basis van literatuur een meer betekenis geven aan een gen-expressie experiment dan een analyse die gebruikt maakt van gen-sets gebaseerd op de klassieke Gene Ontology. Ook laten we de mogelijkheden zien van de op literatuur gebaseerde weighted global test om gen-expressie data te linken aan overleving van borstkanker patiënten en de werking en het metabolisme van medicijnen.

In **hoofdstuk 6** gaan we verder in op het gebruik van concept profile matching voor het maken van gen-sets en de toepassing op het gebied van toxicogenomics. Met behulp van onze thesaurus konden we een veel groter aantal chemische respons-specifieke gen-sets maken met behulp van text-mining dan wanneer we gebruik maakten van methoden gebaseerd op chemische stof-gen interactie informatie uit de Comparative Toxicogenomics Database. We laten zien dat we aan de hand van differentiële expressie van met text-mining verkregen gen-sets kunnen achterhalen wat voor chemische behandeling er in het experiment toegepast is. Daarnaast laten we door drie gen-set testmethoden zien dat onze aanpak ook generiek kan worden ingezet. We laten verder zien dat gen-sets verkregen met concept profile matching gebruikt kunnen worden om embryo-toxische effecten van triazolen al in het gen-expressie stadium te herkennen. Daarbij kunnen triazolen ook onderscheiden worden van andere chemicaliën met behulp van principal component analyse.

Aan de hand van de gegevens beschreven in dit proefschrift concluderen wij dat concept profiling inderdaad geïntegreerd kan worden in het framework van gen-set testen. Daardoor kan het ook gebruikt worden om chemische informatie te koppelen aan gen-expressie data en om toxische effecten al in het gen-expressie stadium te identificeren inclusief het onderscheiden van verschillende chemische klassen.