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

NEXT-GENERATION TEXT-MINING APPLIED TO TOXICOGENOMICS DATA ANALYSIS

Kristina Maria Hettne

ISBN: 978-94-6182-203-1

Cover design: Sara Creson
Thesis layout: Jorg Brunner
Printed by: Off Page, Amsterdam, The Netherlands

NEXT-GENERATION TEXT-MINING APPLIED TO TOXICOGENOMICS DATA ANALYSIS

Next-generation text-mining toegepast op toxicogenomics data analyse

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Universiteit Maastricht,
op gezag van de Rector Magnificus, Prof dr. L. L.G. Soete
volgens het besluit van het College van Decanen,
in het openbaar te verdedigen
op donderdag 20 december 2012 om 12.00 uur

door

Kristina Maria Hettne

geboren te Råda, Sweden op 2 maart 1978

Promotores
Prof. dr. J.C.S. Kleinjans
Prof. dr. J. van der Lei (Erasmus MC)

Copromotor
Dr. J.A. Kors (Erasmus MC)

Beoordelingscommissie
Prof. dr. P. Lambin (voorzitter)
Dr. C. Evelo
Prof. dr. H. van Loveren
Prof. dr. B. Mons (Leids Universitair Medisch Centrum)
Dr. S. Muresan (AstraZeneca, Zweden)

This research was financially supported by the Dutch Technology Foundation STW (MFA6809).

Financial support of the printing costs of this thesis by the Erasmus MC, the foundation BAZIS, and the foundation Stimuleringsfond Alternatieven voor Proefdieren is gratefully acknowledged.

Acknowledgements

I almost can't believe that I finally completed the journey. It would never have been possible without the help of support of many people, and I can only hope that I do not forget to mention someone in the following text.

My copromotor and daily supervisor Jan, this section have to start with you. Scientifically, I admire your sense for detail and your ability to find the missing parts and cracks in any argument or piece of text. You have always given incredibly fast and constructive feedback on my work, and also supported me beyond your duty as supervisor and copromotor during my periods of bad health, all for which I thank you deeply.

Jos and Johan, my promotor and second promotor, I wish to thank you for taking the time to listen to me if I asked for your feedback on my research. My collaborators in the STW project: Rob, André, Bob, Peter H, Dorien, Sandra, Peter T, Petra, and Christina, I thank you for our fruitful discussions.

My Erasmus MC colleagues: I thank you for making working in the "bunker" of the old building a pleasure! Rob, Martijn and Antoine, my "roommates" during the first part of my PhD: you guys made me laugh every day and always took the time to help me when I was banging my head at the computer for the 10th time that day while trying to improve my programming skills (the answer "look in the codebase" was by the way the most helpful one, since it forced me to understand the underlying technology): thank you! Kang, Bharat, Zubair and Rogier, you helped me through the final stages of my PhD by always greeting me every morning with a smile. And I'm so impressed that you didn't manage to kill the plants when I left the room and department! Erik, you are always keen on the translational part of research, and an inspiration in that regard.

My closest colleagues and collaborators at the LUMC: Marco, Eleni, Herman, Erik, Mark, Reinout, Harish, Peter-Bram and Barend, without your support the past 12 months I seriously don't think I would have been able to finish this thesis, and I really mean that. Thank you for believing in me!

Scott, you got me on the text mining track and sparkled my interest in toxicology by hireing me what feels like ages ago for a bioinformatics position at AstraZeneca. I thank you for giving me the courage to jump on the research train and for your continuos interest in my work. Sorel and Antony, thank you for encouraging my work on chemical term identification.

Anna G and Femke, my dear paranymphs, who would imagine that an encounter in the locker room at a gym would lead to such solid friendships! Isabelle, two big bellies at "zwangerschapszwemmen" lead to beautiful kids and an equally beautiful friendship. Sofia, my dear Portugese cat lover, and Lukas, my Swiss friend who shares my passion for dancing, our friendships began in the Netherlands and continued while both of you left the country, for which I am grateful. Rob, Martijn, Geertrui, Marissa, Fedde, Marc, Marja, and Seppe: thank you for making me feel very welcome in your country and for your friendship.

Mamma och pappa, att ha en dotter och barnbarn så långt hemifrån är något ni inte hade önskat. Ni har alltid sagt att det allra viktigaste är att jag är lycklig, men jag vet och förstår vad ni får offra för min lycka. Tack för ert aldrig sviktande stöd! Magnus och Martin, mina kära bröder, jag saknar er varje dag. Jag vill tacka er för att ni alltid tar er tid att träffas när jag kommer till Sverige och för att ni fortsätter vara intresserade av min forskning även ifall jag är urdålig på att förklara vad jag precis gör!

Anna S, Jenny, Sara, Sheila, Maria, Emma, Karin, Mariette, Christoffer, Elie, Anders, Katja, och Anette, vi ses inte ofta nog men när vi ses är det bara att ta upp tråden som om ingen tid passerat. Det är något jag verkligen uppskattar!

Erna, Piet-Hein, Esger en Karg, jullie hebben mij met open armen opgenomen in de familie Brunner, dank ervoor!

Ronja, min lilla älskling. Jag kan fortfarande knappt förstå att du är min dotter. Med din komst blev jag inte bara moder, utan även en bättre människa. Du har lärt mig så mycket och det enda jag kan göra för dig är att vara där när du behöver mig.

Jorg, het is eindeljk zo ver. We hebben samen veel beleeft de afgelopen jaren en jij bent bij me blijven staan. Je bent mijn rustpunt en grote liefde!

Table of Contents

Chapter 1

Introduction

This thesis concerns the application of text mining to the analysis of toxicogenomics data. We begin this chapter by introducing toxicogenomics and text mining as disciplines, and then we continue with describing work already done in integrating text mining with toxicogenomics data analysis methods, and the problems encountered. We then end with an outline of the work done in this thesis.

Toxicogenomics

To predict the toxicity of a compound in humans is essential when assessing the safety of a medical drug, a food ingredient or pesticide. Before compound administration to humans, toxicity testing is traditionally performed in animal models, where acute toxic endpoints such as mortality and behavior, and chronic toxic endpoints such as reproduction, long-term survival and growth, can be studied. Such studies are time-consuming and costly, and not in line with growing emphasis on the 3Rs principle: the reduction, replacement and refinement of animal use [1]. As an alternative to animal testing, toxicogenomics approaches are being developed. In toxicogenomics, gene and protein activity within a particular cell or tissue of an organism in response to toxic substances is studied with the aim to predict in vivo effects from in vitro models [2]. One important technology in toxicogenomics is the DNA microarray: a collection of microscopic DNA spots attached to a solid surface. DNA microarrays are used to measure the activity (the expression) of thousands of genes at different time points at once to create a global picture of cellular function, for example in response to a treatment. Such a global picture of gene expression is called a gene expression profile. The assumption that similar gene expression profiles dictate similar physiological responses underlies the use of gene expression profiling in toxicogenomics to discern the toxicological properties of a chemical entity. It motivates the use of clustering approaches, where chemical entities with similar gene expression profiles are grouped together with the aim to define chemical classes, and connectivity mapping, which produces a ranked list of compounds with a similar gene expression profile to the query compound. Gene expression profiling is also used in mechanistic analysis, where the biology behind a toxicological endpoint at the genomics level is of interest, and in prediction analysis, where gene signatures that can act as early predictors of toxicity are sought. In the following section we will describe more closely the areas of class discovery and separation, connectivity mapping, mechanistic analysis, and identification of early predictors of toxicity.

Class discovery and separation

One of the first approaches used to analyze toxicogenomics data was the clustering of genes within or between samples based on their expression levels with the aim to group compounds with similar toxic mechanisms [3]. However, the study design of a classic gene expression experiment in for example cancer biology where tumor samples are compared with nontumor samples, is different from a typical toxicogenomics study. In a toxicogenomics study, time series and dose-response relationships play a much larger role, and the use of several compounds for comparison further complicates the analysis. These compounds may have unique or common expression signatures and low-dose exposures that may show very small and/or early gene expression changes difficult to distinguish from controls. Nevertheless, clustering approaches have been successfully used to separate two or more classes of samples according to a particular exposure condition or phenotype [4-6]. Other statistical methods that have been used in the area of toxicogenomics to perform class separation and/or prediction include analysis of variance (ANOVA) and linear discriminant analysis methods [7], t-test [8], linear and logistic regressions [9], principal component analysis [10], and support vector machines [11].

Connectivity mapping

Connectivity mapping is performed with the aim to infer toxicity about a new chemical entity via chemicals for which the toxicity is already known (Figure 1). A gene expression profile is compared to reference gene expression profiles collected in databases such as the commercial ToxExpress [12] from GeneLogic, the DrugMatrix [13] database from Entelos, and the publicly available Chemical Effects in Biological Systems (CEBS) [14] database from the National Institutes of Health in the U.S. Pattern matching techniques are used to produce a ranked list of chemical entities with gene expression changes similar to the one induced by the query compound (the suspected toxic agent) (pioneered in [15]).

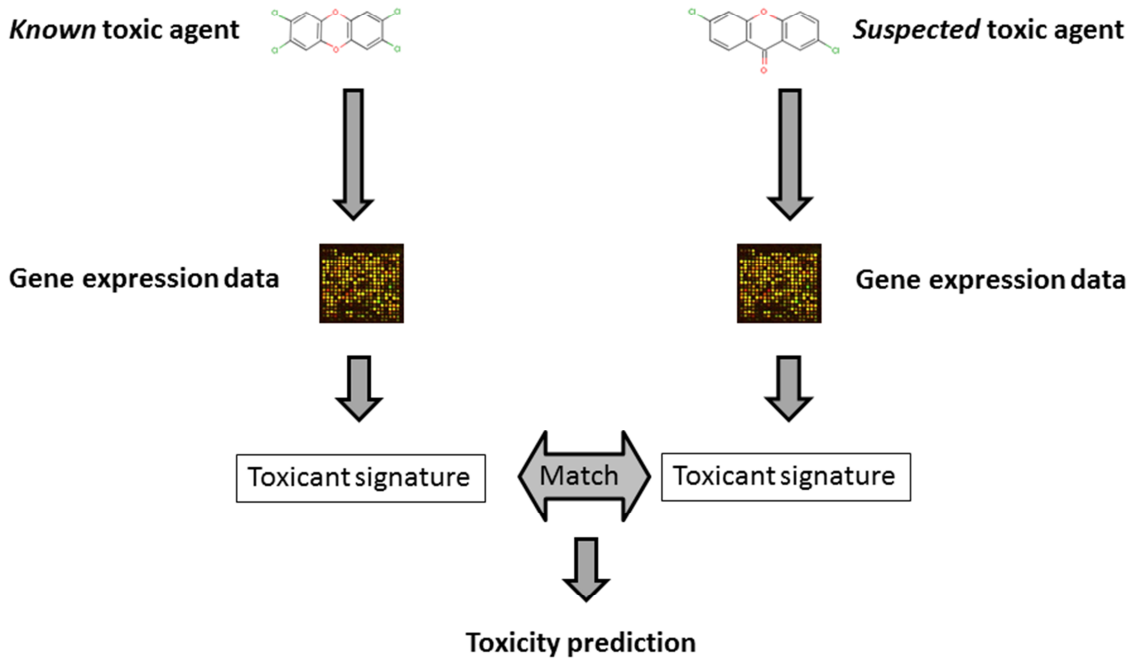

Figure 1: Connectivity mapping. A gene expression signature of a suspected toxic agent is compared to gene expression signatures of known toxic agents.

Mechanistic analysis

Mechanistic analysis is performed to unravel the biology behind a toxicological endpoint at the genomics level, and includes investigating how the differential expression of genes affects biological pathways and processes. By describing a pathway or biological process as a group of genes and/or gene products, one can test the involvement of the particular pathway or biological process in the toxic response to a chemical. If for example the toxicological endpoint is hepatotoxicity (liver toxicity) in response to the drug acetaminophen, one would expect the biological process named "oxidative stress" to appear in the bioinformatic analysis results since that is one of the mechanisms by which acetaminophen causes hepatotoxicity [16]. Gene annotation databases such as the Gene Ontology (GO) database [17] and pathway databases such as Kyoto Encyclopedia of Genes and Genomes (KEGG) [18] are commonly used in mechanistic analyses since they contain such groupings of genes into biological processes and pathways, respectively.

Such a group of genes can also be referred to as a gene set, and analysis methods that utilize gene sets are called gene set analysis (GSA) methods. An alternative to the GO and KEGG databases, more directed at toxicogenomics, is the Comparative Toxicogenomics Database (CTD) [19], which includes curated data from the published literature describing cross-species chemical-gene/protein interactions and chemical- and gene-disease associations. GSA has become standard when analysing gene expression data during the last years and a wide variety of methods have been proposed that use different statistical tests to associate gene sets with a gene expression data set. In brief, three different types of statistical tests are available: (i) tests for the overrepresentation of a gene set in a list of differentially expressed genes using a hypergeometric or equivalent test; (ii) methods that use the P-values of all the genes. Well-known is the gene set enrichment analysis (GSEA) [20], which uses ranked P-values and tests whether the ranks of genes in a gene set differ from a uniform distribution; (iii) regression analyses that use the actual expression levels of the genes in the gene set and test whether these are associated with the studied phenotype, for example the global test [9, 21]. Software programs for performing GSA or other bioinformatic analyses specifically aimed at toxicogenomics data include the commercial IPA-Tox from Ingenuity Pathway Analysis [22] and MetaDrug from GeneGo [23]. A public alternative is ToxProfiler from the Netherlands Organization for Applied Scientific Research [24].

Identifying early predictors of toxicity

The identification of predictive biomarkers is another important application area for toxicogenomics. The path to finding such biomarkers by expression profiling is treacherous, but the search can be aided by bioinformatics approaches (for a recent review, see [25]). Normally a battery of bioinformatics techniques is used, as illustrated by the two following examples. In an early study aimed at predicting nephrotoxicity, Fielden and coworkers [26] produced a gene signature that could predict the future development of renal tubular degeneration weeks before it appeared histologically. To derive a signature, a three-step process of data reduction, signature generation and cross-validation was used. The signature used to predict the presence or absence of future renal tubular degeneration was derived using a sparse linear programming algorithm, which is a classification algorithm based on support vector machines. The authors noticed that many of the genes in the signature were of unknown function, and as a result the mechanism of action was difficult to infer. In another example, this time aimed at hepatotoxicity, Huang and coworkers [27] found that apoptosis-related genes could predict liver necrosis observed in rats exposed to a compendium of hepatotoxicants. They first grouped the liver samples by the level of necrosis exhibited in the tissue. Next, the level of necrosis was derived according to three different methods: 1) the severity scores of the injury; 2) the differentially expressed genes from an ANOVA model; and 3) the GO biological processes enrichment shared by adjacent necrosis levels. They then used a Random Forest classifier with feature selection to identify informative genes (a so called "gene signature"). The gene signature was analyzed for gene function using Ingenuity Pathway Analysis for pathway enrichment and network building.

Text mining and its application in toxicogenomics

The bioinformatics approaches to toxicogenomics data analysis outlined in the previous sections make use of manually curated knowledge bases such as GO, KEGG and CTD, either as a step in the analysis process or at the end when it is time to interpret the results. For example, gene signatures from an experiment where compounds belonging to different toxicological classes are separated based on gene expression changes might be investigated for significantly different GO processes. However, information in manually curated knowledge bases is not sufficient in coverage and this situation is unlikely to change in the near future [28]. Text mining is therefore seen by many as a valuable tool

to provide meaning to data [29]. Text mining can be defined as the use of automated methods for exploiting the enormous amount of information available in the biomedical literature [30]. In the following sections we will first introduce the basics of text mining, then we will continue to describe how it has been used to find explicit and implicit relationships between biomedical concepts, how it has been applied in the analysis of gene expression data, and end by suggesting how it can be applied in toxicogenomics. Box 1 explains common text-mining jargon.

Box 1. Text-mining jargon

Concept

A concept is commonly defined a cognitive unit of meaning - an abstract idea or a mental symbol sometimes defined as a "unit of knowledge". In text-mining a concept is uniquely identifiable.

Thesaurus

A thesaurus is a list containing the preferred term referring to a concept and all its synonyms. One commonly used thesaurus in biomedical text-mining is the UMLS metathesaurus. It is organized by concept, and each concept has specific attributes (i.e. terms and definition) defining its meaning and is linked to the corresponding concept names in the various source vocabularies.

Indexing/tagging

Indexing is the process of scanning all documents for relevant terms referring to a concept. It is also referred to as tagging. The process also usually includes storing the term statistics per document in a database.

Text corpus

A text corpus is a set of documents containing unstructured text (usually electronically stored and processed). Text corpora are used to do information retrieval and extraction.

Precision and recall

Precision is defined as the number of relevant concepts retrieved by a search divided by the total number of concepts retrieved by that search (true positives /(true positives + false positives)), while recall is defined as the number of relevant concepts retrieved by a search divided by the total number of existing relevant concepts (true positives/(true positives + false negatives)). In broader terms: precision is the fraction of retrieved instances that are relevant, while recall is the fraction of relevant instances that are retrieved.

Concept identification in free text

A typical concept identification pipeline is shown in Figure 2. The first step in the concept identification pipeline concerns the retrieval of documents with relevant information from databases that predominantly contain textual information, a task commonly referred to as information retrieval [31]. A common resource used by the biomedical text mining community is the MEDLINE database. A scientist seeking information usually queries MEDLINE through the web-based interface and search engine named PubMed, which is provided by the National Library of Medicine (NLM). For larger scale searches, programming utilities are offered by the NLM. However, due to risks of server overload, the NLM places different limits on these services, and bioinformaticians are often faced with the need to obtain a local version of MEDLINE. A local copy also gives software developers greater control over how they use the data, and facilitates the development of customizable interfaces. A common way to store a local copy of MEDLINE is in a MySQL database, which then can be queried from the bioinformaticians programming environment, for example in Eclipse using the Java programming language. Once the local copy is in place, an update procedure needs to be configured so that the database remains up-to-date. The local copy of the MEDLINE database can then be used for text

mining purposes. Depending on the question that is asked, different subsets of the MEDLINE document collection can be created. For example, if we are interested in all documents describing gene annotation in mice, we can formulate a query that will retrieve documents that only concern genes and mice. This will form our text corpus.

The next step in the concept identification pipeline concerns the identification of relevant terms in the document collection. To continue our example about mouse genes, we now would like to know which genes are mentioned in the text.

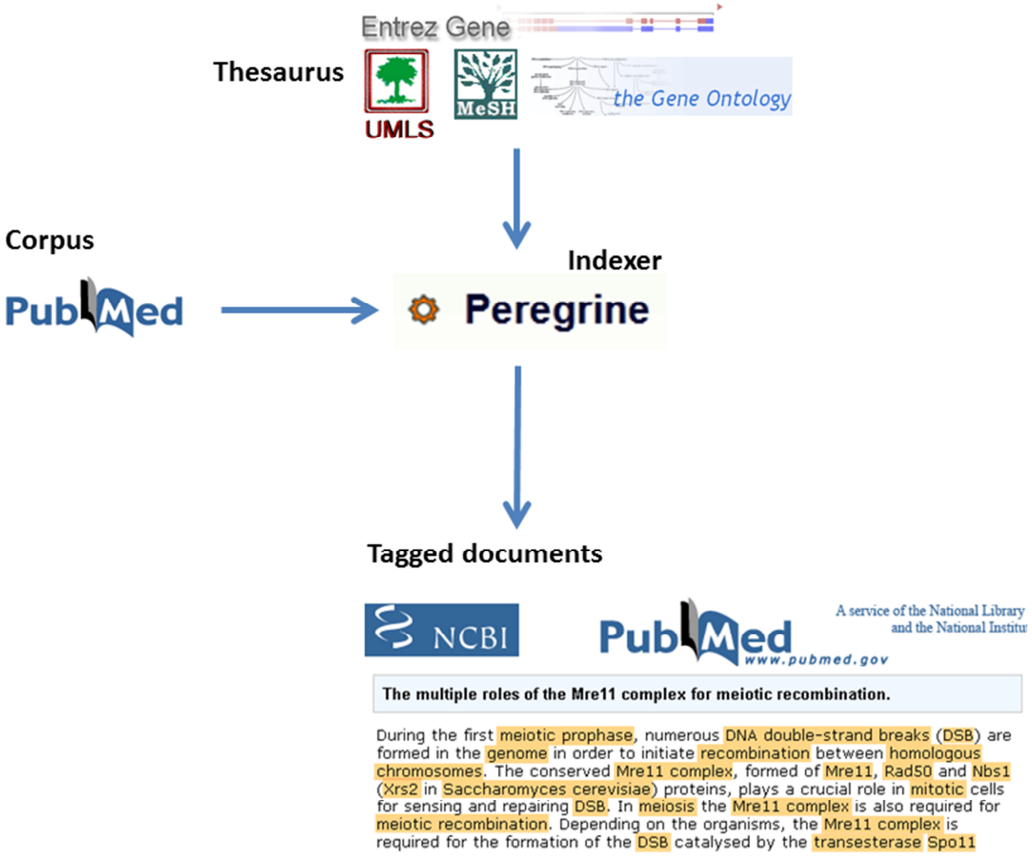

Figure 2: Thesaurus-based concept identification pipeline. An indexing engine/tagger identifies terms referring to concepts in free text using a thesaurus of biomedical concepts.

Approaches to term identification generally fall into one of three categories: thesaurus-based systems, rule-based systems, and statistics-based systems. All approaches have their disadvantages: thesaurus-based systems are dependent on fast updates and large coverage of the underlying thesauri or dictionaries; to craft the rules for a rule-based system is time consuming and requires a high level of domain knowledge, and statistics-based systems, which use machine learning techniques, need annotated corpora to train the classifiers. One important advantage of thesaurus-based systems is the possibility to perform term mapping, where an identified term is linked to a main concept and to reference data sources. A typical record in a thesaurus contains the concept listed together with its synonyms and referent data sources. A thesaurus-based system has two parts: a thesaurus and an indexing engine or tagger: a piece of software that recognizes the concepts from the thesaurus in free text. The tagger that is used in this thesis is called *Peregrine* [32], and has recently been released under an open source license as detailed in [33]. In addition to term identification, usually some form of disambiguation procedure is implemented in the tagger to map the term in the text to the correct concept. This is especially important for gene symbols and chemical name

abbreviations, which are notoriously ambiguous. Disambiguation of terms is important since terms can have different meanings ("word senses") (e.g. "BAP" is a shared synonym between the two chemicals "Benzo(a)pyrene" and "Benzyladenine", and has an additional 58 meanings according to Acronym Finder [34], including "Blood Agar Plate", "BiP-Associated Protein", and "British Association of Psychotherapists"). Word-sense disambiguation algorithms can be distinguished as supervised, unsupervised, or using established knowledge [35]. Peregrine uses established knowledge to disambiguate terms on the fly during the indexation process.

Information extraction

Once the document collection is in place and has been indexed by the tagger, meaningful associations can be extracted from the text, such as relationships between genes and toxicological endpoints, or chemicals and toxicological endpoints. There are currently two different approaches to this problem, one based on co-occurrence of terms and the other on natural language processing (NLP) techniques. The idea behind co-occurrence is that if two concepts are mentioned together, in for example the abstract of a scientific article or a sentence or phrase within that abstract, there might be a relationship between these two concepts. Since two concepts, for example a gene and a chemical, can co-occur without there being a meaningful relationship between them, most co-occurrence based approaches make use of algorithms that take the occurrence frequency of the concepts into account in some way [36-39]. NLP techniques focus on the extraction of precise relationships between genes and other biomedical concepts, using techniques varying from the detection of simple patterns such as "chemical A - action X - gene B" that is used by for example Pharmspresso [40], to the complete parsing of whole sentences (e.g. [41]). Hybrid approaches exist as well [42-44]. As a general trend, a system based on co-occurrence will have a higher recall and lower precision compared to a system based on NLP, which in turn will have a higher precision but lower recall.

Literature-based discovery

Literature-based discovery builds on the techniques for concept identification and information extraction, but adds the extra dimension of trying to discover or predict previously unknown relationships between biomedical entities. Swanson was the pioneer in this field already in the 1980's, publishing studies that were able to predict connections years before they were established in clinical trials [45, 46]. He proposed a model commonly referred to as Swanson's ABC model. The model states that if 'A influences B' and 'B influences C', then 'A may influence C'. Others have built upon this model and managed to reproduce the studies by Swanson, and/or identify novel hypothetical relationships [47-57]. A few studies have been reported where new relationships have actually later been confirmed in animal models or in vitro assays: Wren et al. suggested that chlorpromazine may reduce cardiac hypertrophy, and validated this hypothesis in a rodent model [58]; Hettne et al. related the transcription factor nuclear factor kappa B to Complex Regional Pain Syndrome type I (CRPS-I) [59], a relation that was later confirmed in an animal model of CRPS-I [60]; Van Haagen et al. confirmed a predicted protein-protein interaction between Calpain 3 and Parvalbumin B experimentally [61]; and Frijters et al. validated newly predicted relationships between compounds and cell proliferation in an in vitro cell proliferation assay [62].

To relate two concepts to each other several authors have used the vector space model, as vectors can be compared efficiently and transparently, and the model yields a measure of the strength of the relationship [63]. An example of a tool that uses the vector space model is Anni 2.0 [64]. In Anni 2.0, a concept is represented by a concept profile. A concept profile is a list of concepts with for every concept a weight to indicate the importance of its association to the main concept, based on co-occurrence. To construct a concept profile, first PubMed records are associated to a concept using the indexing engine Peregrine equipped with a thesaurus. For all concepts except genes the PubMed records are comprised of the texts in which the concept is mentioned. For genes

only a subset of PubMed records are used in order to limit the impact of ambiguous terms and distant homologs. GO terms are sometimes given as words or phrases that are infrequently found in the normal texts. To still provide broad coverage of GO terms, the PubMed records that were used as evidence for annotating genes with this GO term are added. For every concept in the thesaurus that is associated to at least five PubMed records, a concept profile is created. This concept profile is in reality a vector containing all concepts related to the main concept (direct co-occurrence), weighted by their importance using the symmetric uncertainly coefficient [64]. Figure 3 presents an overview of the concept profile generation process. Concept profiles can be matched to identify similarities between concept profiles via their shared concepts (indirect relations), for instance to identify genes associated with similar biological processes. Any distance measure can be used for this matching such as the inner product, cosine, angle, Euclidean distance or Pearson's correlation. By next-generation text mining we mean knowledge discovery based on indirect relations, in comparison to first-generation text mining where only known relations are extracted.

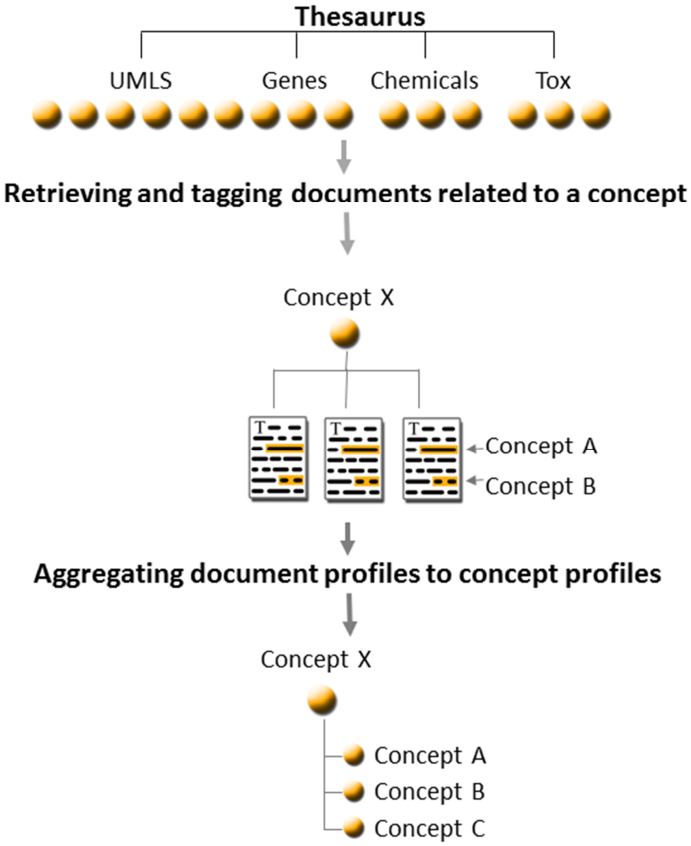

Figure 3: Overview of the concept profile generation process.

Assigning gene function: text mining applied to gene expression data

Techniques for information extraction and literature-based discovery have been applied previously to support the analysis of gene expression data. The majority of these techniques work with gene lists. They can retrieve concepts or terms strongly associated to the selected genes and/or cluster the genes to retrieve functionally coherent subclusters [65]. The main differences between these methods concern the following two aspects: i) usage of text words (all words in the texts or only words that map to concepts in a thesaurus); ii) relationships considered (only direct co-occurrence or also indirect

relationships). Apart from the tools that focus on retrieving gene relations from a list of genes, some methods retrieve functional associations shared between gene lists of different experiments [49, 66].

The analyses performed by the mentioned text mining-based approaches are of an exploratory nature, and do not provide a statistical evaluation for the identified associations in the context of the performed experiment. However, the algorithms can readily be combined with the three previously mentioned classes of statistical approaches for GSA. An approach for the first class of tests, for the overrepresentation of a gene set in a list of differentially expressed genes, could simply entail the creation of gene sets, for instance by applying a threshold on the literature-derived association scores between genes and biomedical concepts. Sartor et al. provide literature-based gene sets in their tool ConceptGen [67], which uses Gene2MeSH [68] to identify gene and MeSH term pairs with a significantly higher number of co-occurrences than expected by chance. Frijters and coworkers [69] and Leong and Kipling [70] calculate biomedical term overrepresentation for a set of regulated genes in a similar fashion to standard class one over-representation tools.

Several text-mining approaches have been published that resemble the earlier mentioned class two approaches that use the P-values of all the genes. Kueffner and coworkers [71] integrate the rank of the genes after sorting on P-value with an analysis of the literature. However, their approach is based on factorization, which complicates the interpretation of their results, and does not include formally testing retrieved associations. Minguez and coworkers [72] test if a ranked list of genes shows a significant correlation with the genes' associations to a biomedical term. These associations are based on the literature and reflect the extent to which a gene and a biomedical term that occur together in documents exceed the co-occurrence rate expected by chance.

One could also imagine a class three, regression-based test, which uses the actual expression levels of the genes in the gene set, in combination with text-mining derived information. The association scores between two concepts based on concept profile matching could provide such a source of literature information.

Application in toxicogenomics

Toxicogenomics could benefit from the use of text mining through the whole process from experimental setup to the final analysis of the data. In the setup of a toxicogenomics experiment, text mining can help answering questions regarding the selected compounds. Examples of such questions are: "which genes are known to be affected?", "which pathologies are known to be induced?", and "what other chemicals are known to have a similar effect?". Traditionally, these type of questions would be answered by searching databases such as the Toxicology Data Network "TOXNET" [73], available from the NLM, PubMed, or the CTD. This usually works fine when information is sought for well-investigated and "fashionable" chemicals such as Bisphenol A, for which a search in the CTD results in 1,245 manually curated gene interactions (14 March 2012), but for chemicals that are unfashionable and/or not very well investigated, the results are often disappointing. For example, a search on the fungicide Flusilazole in the CTD resulted in only two manually curated genes associated with the compound. One cannot exclude the possibility that more information is locked away in the literature, confirming the non-completeness of manually curated knowledge bases. Here, thesaurus-based text mining can prove valuable since it can find information about chemicals in the literature, and importantly also map the chemical to its identifier (like a Chemical Abstract Service (CAS) number or InChI string) via information found in the thesaurus. Naturally, the success of a thesaurus-based text mining approach depends on the coverage of terms in the thesaurus for toxicogenomics and how well the terms can be resolved by NLP. Genes, chemicals, pathways, and toxicological endpoints are obviously important when applying text mining in toxicogenomics. The bioinformatics community has spent a lot of effort in developing text-mining tools for genes and pathways (mostly represented by GO concepts from the ontologies on biological processes and molecular functions). The

identification of these entities in text remains a challenge, but dictionary and rule-based methods as well as machine learning techniques are now well established (the proceedings of the BioCreative challenges [74] give a good overview about the state-of-the-art methods and their performance). The Peregrine tagger equipped with a dictionary of gene names has been reported to have a precision of 0.75 and a recall of 0.76, placing the system in the above average performance range according to the Proceedings of the BioCreative II workshop. However, due to a lack of annotated chemical compound test corpora before year 2008, the performance for dictionary-based chemical name recognition was measured in only one study before that time: Zimmermann and coworkers [75] reported a 80% precision and 99% recall on a modified version of the GENIA corpus [76]. They, however, made it very clear that the high recall was due to the artificial nature of the test corpus (since chemical entities in the GENIA corpus mostly consist of ion names (e.g. Ca+), these entities were removed and replaced by compounds randomly picked from their own small-molecule dictionary). In 2008, Kolarik and coworkers [77] created a test corpus consisting of 100 manually annotated PubMed abstracts. Using a simple case insensitive string search, ignoring hyphens, they tested the recall and precision for a number of public resources of chemical names and a combined version of the dictionaries. The best recall was achieved using a combination of all resources (precision 0.13, recall 0.49) and the best precision was achieved using KEGG drug (precision 0.59, recall 0.12). These modest performance numbers made clear that dictionary-based chemical text mining did not measure up to the numbers achieved for gene name identification, and important links in the chain of events from a chemical exposure to toxicological endpoint would therefore for a large part be missing when applying text mining to toxicogenomics.

Continuing with the step after performing the toxicogenomics experiment, i.e., the data analysis, one can imagine text mining playing a role in all the bioinformatic approaches that we described earlier (i.e. class discovery and separation, connectivity mapping, mechanistic analysis, and identifying early predictors of toxicology). Frijters and coworkers showed that text mining applied to expression data from toxicogenomics experiments can separate compounds that have distinct biological activities and yield detailed insight into the mode of toxicity [78]. They created keyword profiles for compounds based on co-occurrence in MEDLINE between the mentions of the differentially expressed genes in the experiments and keywords from an in-house thesaurus. The pathology thesaurus used in their study was however limited to liver pathologies and co-occurrences between genes and chemicals were not included. Van Dartel and coworkers used gene sets created with Anni 2.0 to show that dedicated gene sets allow for detection of cardiomyocyte differentiation-related effects in the in embryonic stem cells in the early phase of differentiation [79], but only direct co-occurring concepts were used. Some attempts have been done to relate chemical structures to gene expression patterns in microarray experiments using literature-derived chemical response-specific gene sets. Minguez and coworkers [72] used their tool MarmiteScan to associate chemicals with the characteristics of acute myeloid leukemia cell differentiation. However, there is no information about the size and scope of the chemical dictionary they used to mine the literature and their gene sets are not separately available, thus forcing the user to use their GSA method. There is also no possibility to test a subset of the gene sets, for example only those relevant for evaluation of developmental toxicity. Patel and Butte [80] associated chemical response-specific gene sets derived from the CTD with six gene expression data sets selected based on their diversity with respect to species, chemical exposure and cell type. Manual curation of chemical-gene interactions from publications is however a very time-consuming process producing high-quality information but with limited coverage, reflected by the number of chemicals for which Patel and Butte could create gene sets (1,338 chemicals). We hypothesize that many more gene sets can be created using text mining.

Aim and outline of this thesis

Toxicogenomics is a new discipline with its own challenges, and few attempts have been done to apply text mining in toxicogenomics [78]. The bioinformatics community has reported low performance numbers for chemical name identification in text, resulting in the omission of important links in the chain of events from chemical exposure to toxicological endpoint. There is need for research detailing how useful the information in current public chemical databases is for text mining aimed at interpreting toxicogenomics data. We aim to improve the concept profiling technique to work with chemical information, and to apply the technique to interpret data from toxicogenomics experiments. Concept profiling is an established technique, but its use in toxicogenomics has not previously been explored. We hypothesize that concept profiling can be used together with GSA methods for compound ranking à la connectivity mapping, for mechanistic analysis, for toxic class discrimination, and for prediction analysis.

Since a thesaurus forms the base for the technique, the first part of the research described in this thesis was aimed at constructing and evaluating a thesaurus fitted for mining textual data relevant to toxicogenomics. In **Chapter 2** we describe a number of rules that can be used to make a dictionary of biomedical terms ready for text-mining purposes, and we present a tool that implements these rules. In **Chapter 3** and **Chapter 4**, we describe the creation of a chemical dictionary with minimal use of human intervention, and show that this chemical dictionary performs well when compared to manual efforts. We also highlight areas in the field of chemical name identification in free-text needing improvement.

The updated thesaurus was used to create concept profiles, and the second part of the research concerns the application and evaluation of the technique. In **Chapter 5** the developed methods and dictionaries are applied for the first time in a GSA framework with the aim to associate biological processes and drugs with a gene expression data set, and to predict survival. In **Chapter 6** we generalize the method by producing gene sets that can be used by any GSA tool, and test these gene sets on a wide variety of gene expression data sets from toxicogenomics experiments.

Chapter 2

Improving biomedical term identification

Kristina M Hettne[1,2], Erik M van Mulligen[1], Martijn J Schuemie[1], Bob JA Schijvenaars[3] and Jan A Kors[1]

[1] Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, the Netherlands
[2] Department of Health Risk Analysis and Toxicology, Maastricht University, Maastricht, The Netherlands
[3] Collexis Holdings Inc, Columbia SC, USA

Published as: **Rewriting and suppressing UMLS terms for improved biomedical term identification** in *Journal of Biomedical Semantics* 2010, **1**:5

Abstract

Background

Identification of terms is essential for biomedical text mining. We concentrate here on the use of vocabularies for term identification, specifically the Unified Medical Language System (UMLS). To make the UMLS more suitable for biomedical text mining we implemented and evaluated nine term rewrite and eight term suppression rules. The rules rely on UMLS properties that have been identified in previous work by others, together with an additional set of new properties discovered by our group during our work with the UMLS. Our work complements the earlier work in that we measure the impact on the number of terms identified by the different rules on a MEDLINE corpus. The number of uniquely identified terms and their frequency in MEDLINE were computed before and after applying the rules. The 50 most frequently found terms together with a sample of 100 randomly selected terms were evaluated for every rule.

Results

Five of the nine rewrite rules were found to generate additional synonyms and spelling variants that correctly corresponded to the meaning of the original terms and seven out of the eight suppression rules were found to suppress only undesired terms. Using the five rewrite rules that passed our evaluation, we were able to identify 1,117,772 new occurrences of 14,784 rewritten terms in MEDLINE. Without the rewriting, we recognized 651,268 terms belonging to 397,414 concepts; with rewriting, we recognized 666,053 terms belonging to 410,823 concepts, which is an increase of 2.8% in the number of terms and an increase of 3.4% in the number of concepts recognized. Using the seven suppression rules, a total of 257,118 undesired terms were suppressed in the UMLS, notably decreasing its size. 7,397 terms were suppressed in the corpus.

Conclusions

We recommend applying the five rewrite rules and seven suppression rules that passed our evaluation when the UMLS is to be used for biomedical term identification in MEDLINE. A software tool to apply these rules to the UMLS is freely available at http://biosemantics.org/casper.

Background

Biomedical text mining has been shown to be valuable for diverse applications in the domains of molecular biology, toxicogenomics, and medicine. For example, it has been used to functionally annotate gene lists from microarray experiments [71, 81-83], create literature-based compound profiles [78], generate medical hypotheses [50, 84], find new uses for old drugs [45, 46, 85], and measure protein similarity [86, 87]. The identification of biomedical terms in natural language is essential for biomedical text mining. The process of term identification consists of three tasks: term recognition, term classification and term mapping [88, 89]. Approaches to term identification generally fall into three categories: lexicon-based systems, rule-based systems, and statistics-based systems making use of different machine learning techniques [90]. All approaches have their disadvantages: lexicon-based systems are dependent on fast updates and large coverage of the underlying lexicons; to craft the rules for a rule-based system is time consuming and requires a high level of domain knowledge, and statistics-based systems need annotated corpora to train the classifiers. Term mapping, in which terms are linked to reference data sources, is the last step in the term identification process. Term mapping is only possible using lexicon-based term identification and is the focus of this paper (for comprehensive reviews on term identification see for example [88-92]). In addition, the lexicon-based approach deals with general medical terms for which it is difficult to design general matching patterns that are used by rule-based systems. It provides information concerning the semantic relations between terms and supports synonym and referent data source mapping, which is not possible using rule-based or statistically-based term identification. Specifically, we use the Unified Medical Language System (UMLS) meta-thesaurus provided by the U.S National Library of Medicine (NLM) [93]. The 2007AA edition of the UMLS contains more than 1.3 million concepts and 6.4 million terms referring to these concepts from more than 100 different vocabularies. These vocabularies cover different aspects of the biomedical field and have been developed for such different purposes as disease and procedure coding, adverse event reporting, literature indexing, billing, and gene function identification. NLM checks terms from different vocabularies for synonymy, assigns a unique concept identifier (CUI) and assigns concepts to one or more semantic types from the UMLS Semantic Network.

Naturally, the usefulness of the lexicon-based approach depends on the coverage of terms in the vocabulary for the particular domain and how well the terms are suited for natural language processing. The UMLS is not primarily intended as a resource for text mining, so not all of its terms are suitable for this purpose. For example, terms for coding of concepts can include specialized syntax (e.g., brackets) that is not suitable for text mining solutions ("undesired terms"). The following definition of a term in the UMLS can be found in the UMLS Glossary [94]: "A word or collection of words comprising an expression. In the Metathesaurus, a term is the class of all strings that are lexical variants (made singular and normalized to case) of each other". his definition allows for expressions that are not terms according to certain theories of terminology, in which terms are expressions that are actually used in domain-specific communication [95-97]. In fact, the UMLS abounds of expressions that are not expected to occur in any written or oral communication but are intended to precisely paraphrase the exact meaning of a concept. This has been illustrated by, for example, Srinivasan et al. [98, 99], who found that by using normalized matching (i.e. ignoring case variation, punctuation, possessive markers, inflectional variation and word order) only a total of 34.3% of the 1,451,824 terms in the January 2002 version of the UMLS (non-English terms and terms with a suppressible term type excluded) could be found in a 11.5 million MEDLINE abstract corpus. McCray et al [99] could only find 10% of the UMLS terms when using a smaller corpus of 439,741 MEDLINE abstracts (UMLS 2001 version, 1,397,429 English terms, string features retained except for case variation). The lower match result in comparison with Srinivasan et al. might be explained by the difference in the matching methods, the UMLS version, and by the smaller corpus used. McCray et al. [99, 100] also investigated the nature of the strings in the UMLS and evaluated them for their use in natural language processing. The investigation resulted in a number of properties that could be

used to filter unwanted strings from the UMLS. Rogers and Aronson [101] identified a number of filtering rules and term types which help in filtering the UMLS for the update of the MetaMap program [102].

This paper is inspired by McCray et al. [99, 100] and Rogers and Aronson [101] in that we aim to make the UMLS more useable for text-mining purposes. We do this by removing and adding synonyms to the UMLS, which are supposed to increase the accuracy and efficiency of biomedical term identification using the UMLS. We manually evaluate the impact of the rules on a MEDLINE corpus.

Methods

The rewrite rules were implemented to increase the recall of UMLS concepts in text. The suppression rules on the other hand were implemented to rid the UMLS of terms that are undesired when it comes to term identification either because they affect the precision of the term identification, e.g. the synonym "2" for the term "clinical class", the synonym "EC 2.7.1.-" for the concept "human CDC7 protein", or because they affect the efficiency of the term identification, i.e. long and vague terms that are unlikely to be found in text such as the term "poisoning by other and unspecified drugs and medicinal substances" or terms that are useless for concept identification such as the concept with the single term "WHILE". We applied the rules to the 2007AA version of the UMLS in UTF8 coding and then indexed citations from the MEDLINE database (1965-2007) (we refer to this as "the corpus" in the rest of the paper). Finally, the identified rewritten terms were manually assessed for their correspondence to the original UMLS terms and the identified suppressed terms were manually assessed for their usefulness for automatic text mining purposes. A detailed description of the procedure follows.

UMLS extraction

The UMLS 2007AA version was downloaded from the UMLS knowledge source server [103] and installed locally using the MetamorphoSys tool provided by NLM for customizing the UMLS. The default settings in MetamorphoSys were used to create the UMLS subset, using the option to include all vocabularies in the English language. Strings marked as suppressible by the NLM as well as strings longer than 255 characters were not included in the analysis. This approach resulted in 2,844,004 strings, based on the String Unique Identifier (SUI) field in the UMLS. These strings belonged to 1,294,936 concepts, based on the CUI field in the UMLS. Duplicate strings within a concept were removed by comparing strings after conversion to lower case and removal of punctuation; 2,696,820 strings remained and these are henceforth referred to as "terms".

Corpus creation

All MEDLINE citations (title and abstract) available at the time of this study, with publication dates ranging from January 1965 to December 2007 (17,674,805 citations, of which 9,446,335 have an abstract) were used as a test corpus.

Creation of rules

A set of nine rewrite rules and eight suppression rules were given. A description of the rules together with motivation and differences in comparison to original source (when applicable) is provided below. In order to avoid introducing duplicates and homonyms when applying the rewrite rules, a new term was not added to the concept if it could already be found among the synonyms for that concept or any other concept (case insensitive matching after removal of punctuation).

1) Rewrite rules

Syntactic inversion [99, 101]: add syntactic inversion of term if a term contains a comma followed by a space and does not contain a preposition or conjunction (e.g. "Failure, Renal"). We added the condition that only one such pattern of a comma followed by a space is to be found in a term for the rule to be executed.

Possessives [101]: remove the possessive "'s" at the end of a word (e.g. "Alzheimer's disease") and add the rewritten term.

Short form/long form [104]: add short form and long form of term (e.g. "Selective Serotonin Reuptake Inhibitors (SSRIs)"). Schwartz and Hearst's algorithm [104] achieved 96% precision and 82% recall on a standard test collection, which was as good as existing approaches at the time [104] and still competitive according to recent comparison studies [105, 106]. An advantage of the algorithm is that, unlike other approaches, it does not require any training data. Two extra conditions were added to the original rule by Schwartz and Hearst: 1) the short form must be found at the end of the term, and 2) the first letter of the short form should be the same as the first letter of the long form. These conditions were added in order to adjust the rule to extract abbreviations from a dictionary instead of from biomedical text.

Angular brackets [101]: remove expressions within angular brackets anywhere in a term. This pattern was previously used in the UMLS to denote polysemy or homonymy of a term, i.e. a term having different meanings. Terms having this property still exist in the UMLS, even though the property is not assigned to new terms. We have adjusted the rule to remove expressions within angular brackets anywhere in a term since these expressions usually contain meta-information about a term, which is unlikely to be found in text (e.g. "Chondria <beetle>").

Semantic type : remove expressions within parentheses that match the list of semantic types in the UMLS (e.g. "Surgical intervention (finding)"). This rule was developed by our group based on the observation that the semantic type to which the term belongs to is often added as meta-information about the term.

Non-essential parentheticals [99, 101] has been split into four rules in order to make the error analysis more transparent:

1. ***Begin parentheses*** removes expressions within parenthesis at the beginning of a term (e.g. (protein) methionine-R-sulfoxide reductase)
2. ***Begin brackets*** removes expressions within brackets at the beginning of a term (e.g. [V] Alcohol use)
3. ***End parentheses*** removes expressions within parenthesis at the end of a term (e.g. flagellar filament (sensu Bacteria))
4. ***End brackets*** removes expressions within brackets at the end of a term (e.g. Gluten-free foods [generic 1])

In addition, we have added the condition that the rule does not apply to terms belonging to the semantic group Chemicals & Drugs. The reason for this condition is that chemical expressions by nature often contain both brackets and parentheses at the beginning or end of a term.

2) Suppression rules

Short token [99, 101]: remove term if the whole term after tokenization and removal of stop words is a single character, or is an arabic or roman number. For this rule, the stop word list from PubMed [107] was used. This rule differs from the one in [99, 101] in that it takes each token into account separately (e.g. the term "10*9/L" would be tokenised to "10 9 L" and removed by this rule since every token either is a number or a single character).

Dosages [99]: the original rule addressed terms belonging to certain term types defined by the NLM in the UMLS, namely BD (Fully-specified drug brand name that can be prescribed), CD (Clinical Drug) or MS (Multiple names of branded and generic supplies or supplements). This rule was further refined by us to remove all terms that contain a dosage in percent, gram, microgram or milliliter (e.g. Oxygen 2%).

At-sign : this rule was implemented by us to remove terms that contain the @-character (e.g. ADHESIVE @@ BANDAGE).

EC numbers [101]: Remove terms that contain enzyme classification numbers as defined by IUPAC (e.g. EC 2.7.1.112). The justification for this rule is that an EC number in the UMLS usually is mapped to a specific enzyme while it actually refers to a class of enzymes.

Any classification [99]: remove terms containing the following properties: "NEC" at the end of a term and preceded by a comma, "NEC" within parentheses or brackets at the end of a term and preceded by a space, "not elsewhere classified", "unclassified", "without mention" (e.g. "Unclassified sequences").

Any underspecification [99, 101]: remove terms containing the following properties: "not otherwise specified", "not specified", or "unspecified"; "NOS" at the end of a term and preceded by a comma, or "NOS" within parentheses or brackets at the end of a term and preceded by a space (e.g. "Other and unspecified leukaemia").
Miscellaneous [99, 101]: remove terms containing the following properties: "other" at the beginning of a term and followed by a space character or at the end of a term and preceded by a space character; "deprecated", "unknown", "obsolete", "miscellaneous", or "no" at the beginning of a term and followed by a space character (e.g."Other").
Words > 5 [100]: remove terms that contain more that five words (e.g. "Head and Neck Squamous Cell Carcinoma"). This rule is not applied to terms belonging to the semantic group Chemicals & Drugs.

Term and concept recognition

For the term and concept recognition we used our concept recognition software Peregrine [108]. For this study, Peregrine was set up to mimic a minimal, general-purpose concept recognizer performing case-insensitive string lookup (ignoring punctuation), similar to, for instance, TextPresso [109]. Largest match was turned off, meaning that nested terms were counted both as a match for a longer and for a short term. Our choice of set-up was based on the fact that we clearly wanted to see the effect of the rewrite and suppression rules.

Evaluation

Each rule was evaluated separately. To assess the effect of a rule, the difference in the set of terms identified in the corpus before and after applying the rule was determined. For rewrite rules, the number of different additional terms found was determined. In addition, for each term its frequency of occurrence in the corpus was computed. For the suppression rules, the number of different suppressed terms was determined and for each term the number of times it was suppressed in the corpus. A manual analysis of the top 50 most frequent terms and 100 randomly selected terms was performed for each rule. This analysis was used to determine the size of the effect and to judge its quality.

Results

Generation of new synonyms and suppression of undesired ones

The number of new terms generated by the rewrite rules and number of terms suppressed by the suppression rules are shown in Table 1. The *syntactic inversion* rule generated the highest number of new terms (231,976 terms). The number of homonyms generated for every rule is shown in Table 2. The homonyms were not used in the MEDLINE indexation. The *words > 5* rule suppressed the highest number of terms in the thesaurus (653,128 terms). When excluding the *words > 5* rule, a total of 257,118 undesired terms was suppressed in the UMLS, thereby decreasing its size in megabyte by 25%.

Table 1. New terms generated by the rewrite rules and terms suppressed by the suppression rules.

Rule	Terms in thesaurus
Original	2,696,820
Rewrite rules	
Syntactic inversion	231,976
Possessives	10,388
Short/long form	288
Angular brackets	2,824
Semantic type	7,231
Begin parentheses	376
End parentheses	45,265
Begin brackets	11,402
End brackets	17,620

Suppression rules	
Dosages	171,369
Short token	2,044
At-sign	123
EC numbers	161
Any classification	5,299
Any underspecification	40,237
Miscellaneous	37,885
Words > 5	653,128

"Terms in thesaurus" indicates the number of new terms generated by the rewrite rules and the number of terms suppressed by the suppression rules, for every rule. The row "Original" indicates the total number of terms in the thesaurus when no rewrite or suppression rule was applied.

Impact on number of identified terms in the MEDLINE corpus

Of the 2,696,820 original UMLS terms, 651,268 (24.2%) were uniquely identified in the corpus, with an occurrence count of roughly 4 billion; 397,414 of the 1,294,936 distinct concepts (30.6%) were identified. The different rewrite and suppression rules had a different impact on the number of identified terms (Table 3). *Syntactic inversion* (12,433 distinct terms) had the highest impact on number of distinct terms found in the MEDLINE corpus.

Words > 5 (5,734 distinct terms) had the highest impact on the number of distinct terms suppressed in the MEDLINE corpus.

Table 2. Number of homonyms (%) generated for every rewrite rule.

Rewrite rule	**No of homonyms (%)**
Syntactic inversion	303 (0.1)
Possessives	40 (0.4)
Short/long form	321 (52.7)
Angular brackets	218 (7.2)
Semantic type	130 (1.8)
Begin parentheses	28 (6.9)
End parentheses	5,505 (10.8)
Begin brackets	249 (2.1)
End brackets	37,083 (67.8)

The percentage is relative to the total number of rewritten terms for every rule.

In addition, terms suppressed by the *short token* rule and the *miscellaneous* rule are found with an extremely high frequency (*short token*: roughly 2 billion times, *miscellaneous*: 91,576,083 times).The rewrite rules also had different impact on the coverage regarding unique concepts. By rule *syntactic inversion* we have improved coverage by 2.8%, by rule *possessives* the improvement was 0.2%, by rule *short/long form* the improvement was 0.05%, by rule *angular brackets* the improvement was 0.2%, by rule *semantic type* the improvement was 0.07%, by rule *begin parentheses* the improvement was 0.006%, by rule *end parentheses* the improvement was 1.1%, by rule *begin brackets* the improvement was 0.06%, by rule *end brackets* the improvement was 0.06%, overall 5.0%

Manual error analysis of identified terms

A sample of the 50 most frequent terms in the corpus and 100 random terms were analyzed for every rule (see additional file 1: The 50 most frequent and 100 random terms). Based on a manual analysis of the sample terms, we found that six of the nine rewrite rules resulted in incorrectly rewritten terms: *angular brackets*, *short/long form*, *begin parentheses*, *end parentheses*, *begin brackets*, and *end brackets* (Table 4).

The three incorrect terms generated by the *angular brackets* rule were the terms: "<timing>C (_cum_)<meal>" rewritten as "C (_cum_)", "every <integer> weeks" rewritten as "every weeks", "every <integer> minutes" rewritten as "every minutes". Projecting the results from the random sample, the three incorrect terms would correspond to 22 terms (3% of 743 terms) found in the corpus by this rule.

The two incorrect terms generated by the *short/long form* rule in the sample were the terms "Control of skeletal myogenesis by HDAC & calcium/calmodulin-dependent kinase (CaMK)" which gave the long form "calmodulin-dependent kinase" and "Polibar Rapid (P/P)" which gave the short form "P/P". These terms do not correspond to their original UMLS terms, since the first UMLS term describes a process which is incorrectly rewritten as an enzyme, and "P/P" is not a short form of Polibar Rapid. Projecting the results from the random sample, the two incorrect terms would correspond to four terms (2% of 216 terms) found in the corpus by this rule.

Only 26 terms generated by the *begin parentheses* rule were found in the MEDLINE corpus (Table 3) and only one was correct: "(protein) methionine-R-sulfoxide reductase" rewritten as "methionine-R-sulfoxide reductase". Almost all other terms corresponded to the activity of enzymes where the application of the rule resulted in a less specific term, e.g. "(2-5')oligo(A) synthetase activity" rewritten as "oligo(A) synthetase activity". The number of terms in the random sample was equal to the total number of terms found in the corpus by this rule. There is therefore no need to project the results from the random sample.

Table 3. Rewritten or suppressed terms and concepts found in the corpus.

Rule	Terms in corpus (all)	Terms in corpus (distinct)	Concepts in corpus (distinct)
Original	3,992,662,340	651,268	397,414
Rewrite rules			
Syntactic inversion	529,058	12,433	11,291
Possessives	34,211	1,134	946
Short/long form	305,541	216	182
Angular brackets	30,124	743	731
Semantic type	218,838	259	259
Begin parentheses	523	26	25
End parentheses	8,916,764	4,776	4,494
Begin brackets	176,791	274	251
End brackets	65,873	241	236
Suppression rules			
Dosages	109,246	5,014	4,885
Short token	1,906,901,846	1009	945
At-sign	0	0	0
EC numbers	45,138	149	146
Any classification	6,972	42	36
Any underspecification	9,470	322	290
Miscellaneous	91,576,083	1,257	1,095
Words > 5	179,051	5,734	4,665

"Terms in corpus (all)" indicates the number of occurrences of the new terms generated by the rewrite rules and the terms suppressed by the suppression rules in the corpus. "Terms in corpus (distinct)" and "Concepts in corpus (distinct)" indicate the number of unique terms and concepts produced or suppressed by the rules that were found in the corpus. The row "Original" indicates the total number of terms found in corpus when no rewrite or suppression rule was applied.

The *end parentheses* rule had four incorrect terms in the random sample. The incorrect terms in the random sample all corresponded to loci on a chromosome, e.g. "t(3;6)(p13;q25) " rewritten as "t(3;6)". Terms generated by this rewrite rule are found with a high frequency in the corpus (Table 3), which can be explained by the fact that the removal of end parentheses can result in very general terms. For example, rewriting the term "Controls (Instrument)" results in the general term "Controls" that is found 609,492 times in the MEDLINE corpus. Projecting the results from the random sample, the four incorrect terms would correspond to 191 terms (4% of 4,776 terms) found in the corpus by this rule.

The high error rate for the rule *begin brackets* is due to the fact that many of the incorrectly rewritten terms correspond to biological activities of proteins such as the term " [pyruvate dehydrogenase (lipoamide)] phosphatase activity", which is incorrectly rewritten as "phosphatase activity". On the other hand, many of the correctly rewritten terms corresponded to terms that start with a code, e.g. " [D]Respiratory abnormalities", which is correctly rewritten as "Respiratory abnormalities". Projecting the results from the random sample, the nine incorrect terms would correspond to 25 terms (9% of 274 terms) found in the corpus by this rule.

Almost all incorrect terms produced by the *end brackets* rule corresponded to antigens of a specific bacterial strain, e.g. "Shigella flexneri 2a [II:3,4]" incorrectly rewritten as "Shigella flexneri 2a". Terms generated by this rewrite rule are found with a high frequency in the corpus (Table 3), which can be explained by the fact that the removal of end brackets from terms such as "Abstracts [Publication Type]" results in the very general term "Abstracts", which is found 25,082 times in the MEDLINE corpus. Projecting the results from the random sample, the five incorrect terms would correspond to 12 terms (5% of 241 terms) found in the corpus by this rule.

Table 4. Number of correct and incorrect terms for each of the rewrite and suppression rules.

Rule	**Most frequent**		**Random**	
Rewrite rules	***Correct***	***Incorrect***	***Correct***	***Incorrect***
Syntactic inversion	50	0	100	0
Possessives	50	0	100	0
Short/long form	49	1	98	2
Angular brackets	50	0	97	3
Semantic type	50	0	100	0
Begin parentheses	1	25	-	-
End parentheses	49	1	96	4
Begin brackets	38	12	91	9
End brackets	46	4	95	5
Suppression rules				
Dosages	50	0	100	0
Short token	50	0	100	0
At-sign	-	-	-	-
EC numbers	50	0	99	0
Any classification	50	0	100	0
Any underspecification	50	0	100	0
Miscellaneous	50	0	100	0
Words > 5	0	50	5	95

The calculations are based on the, for every rule, 50 most frequently found terms in the corpus and 100 randomly selected terms in the corpus (if available). The At-sign rule has no values because terms suppressed by this rule were not found in the corpus.

Most terms suppressed by the *words > 5* rule were found to be valuable terms that did not need to be suppressed, e.g. "Carcinoma of the Head and Neck", "insulin-like growth

factor binding protein 1". Projecting the results from the random sample, the 95 incorrect terms would correspond to 4,432 terms (95% of 4,665 terms) found in the corpus by this rule.

None of the suppression rules except the *words > 5* rule caused any correct term to be suppressed in the sample (Table 4). All suppressed terms were either too generic (e.g. "Unspecified conditions", "Of"), highly unlikely to be found in the literature (e.g. "Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified"), or suppressed for specific needs (terms from the *dosages* and *EC numbers* rules).

Discussion

To make the UMLS more suitable for biomedical text mining we implemented and evaluated nine term rewrite and eight term suppression rules. In the creation of these rules, we used and refined many of the UMLS string properties identified by McCray et al. [99, 100] and Rogers and Aronson [101], together with an additional set of new properties discovered by our group during our work with the UMLS. Our work complements the work by McCray et al. and Rogers and Aronson in that we measured the impact on the number of terms identified by the different rules on all of MEDLINE (1965-2007), whereas the others only reported the number of strings in the UMLS that were affected by the specific string properties, and in that we also performed a manual analysis of the rewritten terms retrieved from the corpus and of the terms that were suppressed in the corpus. This was done in order to establish that the rewritten terms indeed correspond to the original terms and that only undesired terms were suppressed by the suppression rules.

The goal set for the rewrite rules was to increase the number of synonyms for a UMLS concept and thereby also increase the number of times a concept will be correctly identified in text, thus increasing the recall of biomedical term identification. Good rewrite rules should not generate terms that do not correspond to the original term. This holds true for the rules *syntactic inversion*, *possessives* and *semantic type*. The *angular brackets* rule and the *short/long form* rule generated a few incorrect cases. The incorrect cases from the *angular brackets* rule represent repeat patterns used in coding meal-related timings in patient records. These terms are in fact compositional grammar representing a class of "terms" in which various parts of a complex term are separated to their primitive codes and then put together through, for example, qualifiers. Hypothetically, such template terms could yield instance terms that have matches in the text. Rewriting these terms alters the template pattern and therefore the meaning of the term. Despite the incorrect cases generated by the *angular brackets* rule we recommend it to be used, but with a manual check of the results. We argue that this is feasible considering the small number of incorrect cases. We also recommend the use of the *short/long form* rule together with a manual check of the results. We find this advisable since the number of terms generated by the rule is relatively small (288 terms) but significant: it for example adds the commonly used abbreviation "SSRIs" to the term "Selective Serotonin Reuptake Inhibitors". It can also be noted that about half of the terms generated by this rule were homonyms. This indicates that the rule gives rise to quite ambiguous terms, which is another reason why we recommend a manual check of the results of this rule. Our analysis revealed that even though the different rules for rewriting terms with parentheses or brackets had an impact on the number of rewritten terms found, the quality of the rewritten terms was not perfect. The terms giving most problems were the names of biological entities, such as a genetic locus or the activity of an enzyme. These problems might be solved by introducing the criteria that parentheticals at the beginning (end) of a term should only be removed if they are followed (preceded) by a white space. This however would cause the rules to miss obvious cases without a white space where rewriting is necessary, such as " [M]Lymphoid leukaemias" (where the [M] is specific for the Read Codes vocabulary). A more promising way to tackle this problem is to analyze what kind of strings are found between parentheticals in the UMLS and based on these findings try to rewrite the terms. Our

group has in this manner found that in 7,231 cases in the UMLS, the string between parentheses at the end of a term actually corresponds to the semantic type to which the term belongs. The implementation of this *semantic type* rule caused 259 new distinct terms to be found 218,838 times in our corpus and thus improves the recall of the UMLS-based term matching. Using the rewrite rules that passed our evaluation (rules 1-5) we were able to identify 1,117,772 new occurrences of 14,784 rewritten terms in the corpus. Projecting the results from the manual evaluation of 100 random rewritten terms per rule, 26 of these 14,784 terms would be incorrect.

Removal of erroneous synonyms improves the precision of the UMLS terms and removal of unnecessary terms and synonyms reduces the size of the UMLS, thus improving its efficiency. The evaluation of the suppression rules showed that all except one, namely the *words > 5* rule, are safe to apply when the UMLS is to be used for concept identification in text.

Use scenarios

The following use scenarios illustrate the usefulness of the rules:

1) Concept-based biomedical information extraction

The rules that passed the evaluation in this work can be used to prepare the UMLS for use in a lexicon-based information extraction pipeline with the goal of identifying biomedical concepts in text. For illustration, we used the rules to prepare the UMLS for the indexing of MEDLINE abstracts with MetaMap. It is worth mentioning that most of the rules are already partly or fully implemented in MetaMap and that it is not possible to measure the exact effect of the different rules on MetaMap since MetaMap also performs a number of steps (part-of-speech tagging, shallow parsing and normalization) that all affect its performance. MetaMap also has an internal rule engine that cannot be switched on or off for specific rules. Indeed, by applying the rules to the UMLS and subsequently indexing a random set of 10,000 MEDLINE abstracts using MetaMap, only a minor increase in recall and precision was gained (31 additional concepts were recognized and 95 concepts were suppressed, all manually checked and found to be correctly recognized or suppressed). It is worth noting that MetaMap is not designed to work on large corpora: indexing the 10,000 abstracts took 33 hours on a medium performance computer. Using Peregrine with the settings described in this paper, 17,674,805 citations (9,446,335 of these have an abstract) were indexed within the same amount of time.

2) Chemical name identification

In a separate study, we used the rules suitable for chemical terms as a pre-processing step in the creation of a multi-source chemical dictionary [110]. We used the suppression rules *short token*, *dosages*, *at-sign*, *any underspecification*, and *miscellaneous*, and the rewrite rules *syntactic inversion*, *possessives*, and *short form/long form* for this purpose. The dictionary was tested on a corpus annotated with chemical entities [111] and recall and precision was calculated. The rules doubled the precision, leaving the recall practically unchanged. From this use case it is obvious that the suppression rules played a large role in increasing the precision of a chemical dictionary by removing highly ambiguous terms that are rarely used as synonyms for chemicals in text. Examples of such synonyms are single letter acronyms and general English words. The rewrite rules played a less important role, only generating a few extra hits that did not influence the recall much.

Limitations

A limitation for the generalizability of our study is that we restricted ourselves to MEDLINE and did not include other types of text such as electronic patient records, which have a different structure that might influence the performance of the rules.

Future work

A restriction on size or on the type of content within parentheticals might lead to additional useful rewrite or suppression rules. Furthermore, a vocabulary-based suppression of terms in the UMLS might also be applicable since each vocabulary has been independently developed and adheres to its own rules. One could for example question the use of the vocabulary NCI modified Common Terminology Criteria for

Adverse Events v3.0, 2003 (NCI-CTCAE), for which only two out of the 4504 terms in the vocabulary were found in the corpus. A quick analysis of the terms in NCI-CTCAE showed that many of them may be useful for clinical applications but not for knowledge discovery aiming at for example finding links between chemicals and adverse events in free text. An example is the term "CTCAE Grade 1 Supraventricular extrasystoles (Premature Atrial Contractions; Premature Nodal/Junctional Contractions)", which is very specific but will not be found in free text. Another example comes from Read codes where an axis indicator as [M] is often used before a term.

To further investigate the generalizability of the rules they should be tested on another type of text than MEDLINE, for example electronic patient records.

Conclusions

We recommend the usage of the five rewrite rules and seven suppression rules that passed our evaluation when the UMLS is to be used for term identification in free text. Using these five rewrite rules we were able to identify 1,117,772 new occurrences of 14,784 rewritten terms in MEDLINE. Without the rewriting, we recognized 651,268 terms belonging to 397,414 concepts; with rewriting, we recognized 666,053 terms belonging to 410,823 concepts, which is an increase of 2.8% in the number of terms and an increase of 3.4% in the number of concepts recognized. Using the seven suppression rules, a total of 257,118 undesired terms were suppressed in the UMLS, thereby decreasing its size in megabyte by 25%, and 7,397 terms were suppressed in the corpus. By rewriting and suppressing the UMLS (and thereby increasing its recall and precision) it becomes more suitable for biomedical text mining purposes, such as information retrieval and knowledge discovery.

All the rules evaluated in this paper can be applied to UMLS data by using the software program Casper, which is available online at http://www.biosemantics.org. Casper takes a UMLS data file as input and gives a rewritten and suppressed UMLS data file as output. This UMLS data file can then be used together with any concept recognition software of choice. Please note that Casper operates on UMLS data, for which a license is needed.

Additional files

The additional files of this chapter are freely available at the Journal of Biomedical Semantics website: http://www.jbiomedsem.com/content/1/1/5/additional

Chapter 3

Improving chemical term identification

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

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

Published as: **A dictionary to identify small molecules and drugs in free text** in *Bioinformatics* 2009, **25** (22): 2983-2991

Abstract

Motivation
From the scientific community, a lot of effort has been spent on the correct identification of gene and protein names in text, while less effort has been spent on the correct identification of chemical names. Dictionary-based term identification has the power to recognize the diverse representation of chemical information in the literature and map the chemicals to their database identifiers.
Results
We developed a dictionary for the identification of small molecules and drugs in text, combining information from UMLS, MeSH, ChEBI, DrugBank, KEGG, HMDB and ChemIDplus. Rule-based term filtering, manual check of highly frequent terms and disambiguation rules were applied. We tested the combined dictionary and the dictionaries derived from the individual resources on an annotated corpus, and conclude the following: (i) each of the different processing steps increase precision with a minor loss of recall; (ii) the overall performance of the combined dictionary is acceptable (precision 0.67, recall 0.40 (0.80 for trivial names); (iii) the combined dictionary performed better than the dictionary in the chemical recognizer OSCAR3; (iv) the performance of a dictionary based on ChemIDplus alone is comparable to the performance of the combined dictionary.
Availability
The combined dictionary is freely available as an XML file in Simple Knowledge Organization System format on the web site http://www.biosemantics.org/jochem.

Introduction

Biomedical text mining has been shown to be valuable for diverse applications in the domains of molecular biology, toxicogenomics, and medicine. The techniques behind current text mining applications focus, however, for a great part on the ability of the system to correctly identify gene and protein names in text, while less effort has been spent on the correct identification of chemical names [91, 112]. Indeed, the domains of genomics and chemistry have developed quite separate from each other, until now, with the important difference that genomic databases and the bioinformatics tools used to mine them arise from an open-source and open-access friendly community while chemistry has a long tradition of closedness and restricted access to data [75, 113]. This is, however, about to change as more and more chemical resources are becoming freely available [e.g. the chemistry search engine ChemSpider (http://www.chemspider.com)], giving rise to the new research field of chemical genomics [114-116]. ChemSpider has an internal dictionary containing links to many public chemical databases and provides web services to access the data. The dictionary is, however, not downloadable and there is no information published on how the dictionary was created and evaluated, which makes it difficult to include it in text mining applications.

Finding biomedical terms in natural language is essential for biomedical text mining. Biomedical named entity recognition (NER) is the task of identifying the boundary of a substring and then map the substring to a predefined category [92]. Approaches to NER generally fall into three categories: dictionary-based systems, rule-based systems and statistically based systems making use of different machine learning techniques [90]. The challenges of chemical name identification differ from the ones in the genomics field in the sense that the exact placement of tokens such as commas, spaces, hyphens and parentheses plays a much larger role. Chemical NER in general has been reviewed by Banville [117] and methods for confidence-based chemical NER have been evaluated by Corbett and Copestake [118]. According to Klinger *et al*. [77], the only chemical NER software freely available to the academic community is OSCAR3 (http://sourceforge.net/projects/oscar3-chem) [119]. OSCAR3 uses a combined NER approach of overlapping 4 g together with a dictionary based on the Chemical Entities (CM) of Biological Interest (ChEBI) ontology [120].

In this article, we focus on the task of term *identification*, which goes beyond NER to also include term *mapping*, i.e. the linking of terms to referent data sources. To achieve this, a dictionary with database links is essential. For instance, the Whatizit system is able to directly link protein names to their respective UniProt-ID using a dictionary generated from the UniProt database [121]. Naturally, the usefulness of the dictionary approach depends on the coverage of terms in the dictionary for the particular domain and how well the terms are suited for natural language processing. Recently, resources such as DrugBank [122] and the Unified Medical Language System (UMLS) metathesaurus [93] have been applied for the identification of drug names in text [123, 124] [for a recent review of literature mining in support of drug discovery, see Agarwal and Searls [125]]. In this work, we aim at a broader level of chemical identification where also the organism's own biomolecules such as metabolites and signaling molecules are included, referred to as small molecules in the rest of this article. Dictionary-based approaches aiming at identifying small molecules in text have used different proprietary resources to create their dictionary: Singh *et al*. [126] used the proprietary Compound Knowledge Base system [127], and Zhu *et al*. [128] used the proprietary Chemical Abstracts Services (CAS) Registry numbers [129]. Due to a lack of annotated chemical compound test corpora, before the year 2008 only one study reported the recall and precision of a small-molecule dictionary: Zimmermann *et al*. [75] evaluated a dictionary consisting of the chemical part of the Medical Subject Headings (MeSH) [130] together with ChEBI using the ProMiner system [131] on a modified version of the GENIA corpus [76], and reported 80% precision and 99% recall. They, however, made it very clear that the high recall was due to the artificial nature of the test corpus (since CM in the GENIA corpus mostly consist of ion names (e.g. Ca^{+}), these entities were removed and replaced

by compounds randomly picked from their small-molecule dictionary). Recently, Kolarik *et al*. [111] created a test corpus consisting of 100 manually annotated PubMed abstracts. Using a simple case insensitive string search, ignoring hyphens, they tested the recall and precision for MeSH headings, MeSH supplementary concept records, ChEBI, PubChem [132], DrugBank, KEGG drug [133], KEGG compound [134], Human Metabolome database (HMDB) [135], and for a combined version of the dictionaries, with the goal of gaining knowledge about the suitability for a dictionary with curation effort. The best recall was achieved using the a combination of all resources (precision 13%, recall 49%) and the best precision was achieved using KEGG drug (precision 59%, recall 12%).

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

Methods

Choice of chemical resources

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

Chemicals from a broad chemical space

The UMLS (http://www.nlm.nih.gov/research/umls/) contains information about biomedical and health-related concepts, their various names and the relationships among them. It is provided by the US National Library of Medicine (NLM). All entities, henceforth referred to as concepts, in the UMLS are assigned a unique concept identifier (CUI); the terms belonging to the concept are, in turn, assigned a unique term identifier (LUI), a unique string identifier (SUI) and a unique atom identifier (AUI). We extracted concepts based on the CUI and terms based on the SUI. In addition, every concept has at least one semantic type from the Semantic Network (http://www.nlm.nih.gov/pubs/factsheets/umlssemn.html) assigned to it. These semantic types have been aggregated into semantic groups [136]. Similar to Wilbur *et al*. [137], we used the semantic types belonging to the semantic group 'Chemicals & Drugs' and removed the types T120 'Chemical Viewed Functionally', T122 'Biomedical and Dental Material' and T192 'Receptor'. In contrast, we excluded the semantic types T200 'Clinical Drug' (e.g. 'fluorescein 250 mg/ml injectable solution [fluorescein lite]'), T126 'Enzyme' (e.g. 'Kininase III'), T116 'Amino Acid, Peptide or Protein' (e.g. 'alpha 1-antitrypsin-leukocyte elastase complex') and T103 'Chemical' (e.g. 'Chemicals'), and in addition, we added the semantic type T129 'Immunologic Factor' (e.g. 'Efalizumab'). The different choice of removal or inclusion of semantic types compared with Wilbur *et al*. [137] was determined based on a manual analysis of a random set of 100 terms from each semantic type, with the criteria that the terms should mainly represent small molecules or drugs and be likely to be found in text. Since the UMLS does not contain CAS numbers or InChI strings, the concepts were mapped to CAS numbers via the MeSH identifier in the UMLS. The resulting dictionary will be referred to as UMLSchem.

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

MeSH supplemental concept records (http://www.nlm.nih.gov/mesh/) are used to index chemicals, drugs and other concepts for MEDLINE and are searchable by Substance Name [NM] in PubMed. We extracted records concerning small molecules from MeSH by

filtering for the same semantic types as we used for the UMLS and MeSH. We will refer to this dictionary as MeSHsupp.

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

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

Drug terminology

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

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

Metabolic substances

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

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

Toxic substances

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

Data extraction

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

Dictionary pre-processing

We have previously investigated the effect of a number of rewrite and suppress rules, collectively called *filtering rules*, on the terms in the UMLS [138]. The number of uniquely identified terms and their frequency in MEDLINE were computed before and after applying the rules. The 50 most frequently found terms together with a sample of 100 randomly selected terms were evaluated per rule. Using the rewrite rules that passed our evaluation, we were able to identify 1 117 772 new occurrences of 14 784 rewritten terms, and using the suppress rules that passed our evaluation, a total of 257 118 were suppressed in the UMLS. We also implemented a software tool to apply these rules to the UMLS (http://biosemantics.org/casper). We decided to use the rules suitable for chemical terms to rewrite and suppress terms in the chemical dictionaries. The rules are listed and explained below together with references to the original sources.

Short token filter rule [99, 101]: remove term if the whole term after tokenization and removal of stop words is a single character, or is an Arabic or Roman number (e.g. 'T' as an abbreviation for 'Tritium'). For this rule, the stop word list from PubMed (http://www.ncbi.nlm.nih.gov/books/bv.fcgi?highlight=stopwords&rid=helppubmed.table.pubmedhelp.T43) was used. This rule resembles the one mentioned in McCray *et al*. [136] and Rogers and Aronson [101] with the difference that it takes each token into account separately.

Dosages rule [136]: the original rule addressed terms belonging to certain term types in the UMLS, namely BD (fully specified drug brand name that can be prescribed), CD (Clinical Drug) or MS (Multiple names of branded and generic supplies or supplements). This rule was further refined by us to remove all terms that contain a dosage in percent, gram, microgram or milliliter (e.g. 'Theophylline 0.4% and dextrose 5% in plastic container' as a synonym for 'Theophylline').

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

Any underspecification rule [101, 136]: remove terms that contain any of the following features: 'not otherwise specified', 'not specified' or 'unspecified'; 'NOS' at the end of a term and preceded by a comma, or 'NOS' within parentheses or brackets at the end of a term and preceded by a space (e.g. 'unspecified phosphate of chloroquine diphosphate' as synonym for 'chloroquine diphosphate').

Miscellaneous rule [101, 136]: remove terms that contain the following features: 'other' at the beginning of a term and followed by a space character or at the end of a term and preceded by a space character, 'deprecated', 'unknown', 'obsolete', 'miscellaneous' or 'no' at the beginning of a term and followed by a space character (e.g. 'no stereochem' as synonym for 'Encainide').

Syntactic inversion rule [101, 136]: add syntactic inversion of term if a term contains a comma followed by a space and does not contain a preposition or conjunction (e.g. 'acid, gamma-vinyl-gamma-aminobutyric' is rewritten to 'gamma-vinyl-gamma-aminobutyric acid'). We added the condition that only one such pattern of a comma followed by a space is to be found in a term for the rule to be executed.

Possessives rule [101, 136]: remove the possessive 's' at the end of a term (e.g. 'Ringer's lactate' rewritten as 'Ringer lactate') and add the rewritten term.

Short form/long form rule [104]: add short form and long form of term [e.g. 'Hydrogen chloride (HCL)' is split into 'Hydrogen chloride' and 'HCL']. The rule is based on the abbreviation finding algorithm described by Schwartz and Hearst [104]. The algorithm achieved 96% precision and 82% recall on a standard test collection, which was as good as existing approaches at the time [104] and still competitive according to recent comparison studies [105, 106]. An advantage of the algorithm is that, unlike other approaches, it does not require any training data. Two extra conditions were added to the original rule by Schwartz and Hearst: (i) the short form must be found at the end of the term, and (ii) the first letter of the short form should be the same as the first letter of the long form. These conditions were added in order to adjust the rule to extract abbreviations from a dictionary instead of from biomedical text.

Manual check of highly frequent terms

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

Data resource combination

We merged entries if they had the same CAS numbers [similar to Zimmerman *et al*. [75]], database identifier, or InChI string.

Identification of chemical names

For the term and concept identification, we used our concept recognition software Peregrine [139]. The Peregrine system was designed with two goals in mind. First of all, it should be easy to maintain. There is only a single step (manual check of highly frequent terms) that requires human involvement when implementing a new lexicon. The second goal was speed. Because Peregrine does not rely on part-of-speech tagging or natural language parsing, it is very fast: 100 000 MEDLINE records can be processed in 213 s on a standard PC. The whole of MEDLINE can be processed within a single day [140]. The Peregrine system translates the terms in the dictionary into sequences of tokens or words. When such a sequence of tokens is found in a document, the term, and thus the chemical associated with that term, is recognized in the text. Some tokens are ignored, since these are considered to be non-informative ('of', 'the', 'and' and 'in'). We used Peregrine with the following settings: case-insensitive, word-order sensitive and largest match. In its default setting, the tokenizer in Peregrine considers everything that is not a letter or a digit to be a word delimiter. To fine-tune the tokenizer for chemical concept recognition we made the following adjustments: full stops, commas, plus signs, hyphens, single quotation marks and all types of parentheses ((, {, [) were excluded from the word delimiter list. After tokenization, the tokens were stripped of trailing full stops, commas and non-matching parentheses. Parentheses were also removed if they surrounded the whole token. In addition, a list of common suffixes was used to remove these suffixes at the end of tokens (Supplementary Material 2). The suffix list was obtained by scanning the whole UMLS (i.e. not just the chemical part) for suffixes that were English verbs or adjectives.

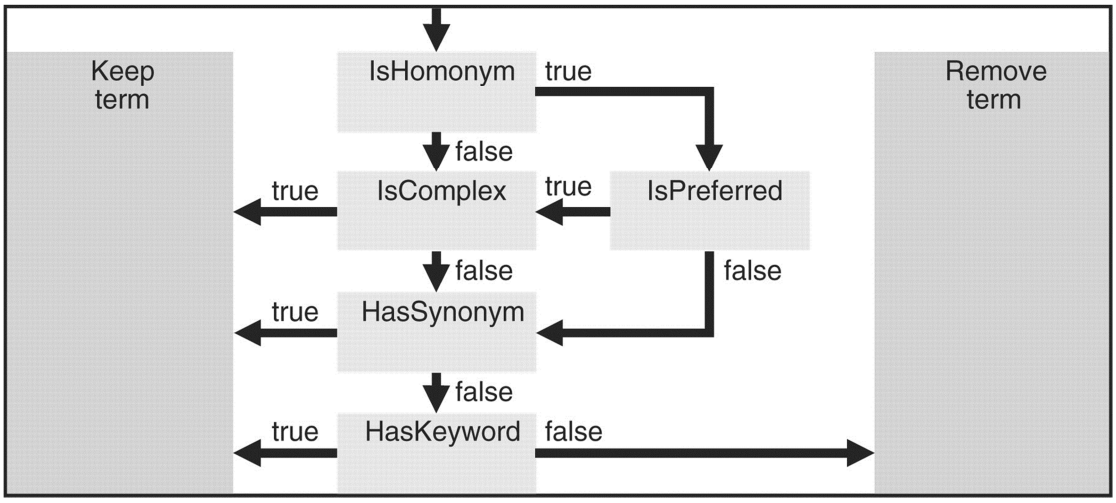

Figure 1. Term disambiguation scheme.

Disambiguation rules

Disambiguation of terms is important since terms not only can have different meanings ('word senses') in a dictionary but also in text (e.g. 'BAP' is a shared synonym between the two chemicals 'Benzo(a)pyrene' and 'Benzyladenine' and has an additional 44 meanings according to Acronym Finder (http://www.acronymfinder.com), including

'Blood Agar Plate', 'BiP-Associated Protein' and 'British Association of Psychotherapists'). Word-sense disambiguation algorithms can be distinguished as supervised, unsupervised or using established knowledge {Alexopoulou, 2009 #96;Edmonds, 2006 #97}. Peregrine uses established knowledge to disambiguate terms on the fly during the indexation process. Specifically, Schuemie *et al*. [108, 141] evaluated a number of rules to disambiguate gene names found in text. These disambiguation rules are potentially also applicable to chemical names. Disambiguation of terms found in text was carried out as follows (Fig. 1). We first determine whether a term is a dictionary homonym, i.e. if it refers to more than one entity in the dictionary. If the term is a dictionary homonym, but it is the preferred term of that entity, it is further handled as if it is not a dictionary homonym. If the term is not a dictionary homonym it still needs further processing since it can have many meanings in text. Therefore, terms that are not complex (i.e. longer than five characters or containing a number) are also considered potential homonyms, and require extra information to be assigned. A (potential) homonym is only kept if (i) another synonym of the entity is found in the same piece of text; (ii) a keyword (i.e. a word or 'token' that occurs in any of the long-form names of the small molecule, and appears less than 1000 times in the dictionary as a whole) is found in the same piece of text.

Annotated test corpus

The annotated corpus (http://www.scai.fraunhofer.de/chem-corpora.html) from Kolarik *et al*. (2008)[111] was used to test the chemical dictionaries. The corpus consists of 100 MEDLINE abstracts with 1206 annotated chemical occurrences divided into the following groups: multi-word systematic names (IUPAC, 391 occurrences), partial chemical names (PART, 92 occurrences), sum formulas (SUM, 49 occurrences), trivial names (including single word IUPAC names) (TRIV, 414 occurrences), abbreviations (ABB, 161 occurrences) and chemical family names (FAM, 99 occurrences). Larger drug molecules such as protein drugs were not annotated. See Kolarik *et al*. [77] for details on the creation of the corpus.

Results

Dictionary characteristics

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

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

Dictionary	Concepts	Terms	CAS numbers	InChI strings
ChEBI	11 428	65 409	6436 (6295)	11 212 (11 152)
ChemIDplus	260 393	1 378 808	260 393 (260 393)	–
DrugBank	4540	37 508	2240 (2218)	4381 (4208)
HMDB	6859	75 957	2683 (2537)	6857 (6734)
KEGGc	11 976	31 143	7695 (7661)	11 875 (11 738)

KEGGd	6927	18 697	6769 (6670)	6140 (6083)
MeSHchem	2897	29 023	2897 (2897)	–
MeSHsupp	19 137	92 918	19 137 (19 137)	–
PubChem	383 043	2 121 960	420 737 (395 108)	16 222 (16 108)
UMLSchem	47 508	126 470	47 509 (18 703)	–
Combined	377 849	2 600 445	400 899 (400 899)	50 254 (50 25

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

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

Dictionary performance

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

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

Table 2 shows the effect of preprocessing and disambiguation on precision and recall for each of the dictionaries. The values reported by Kolarik *et al*. [77] are also shown, if available. It is clear that the preprocessing steps and the disambiguation rules have a strong positive influence on the precision of all dictionaries. We also achieve higher recall and precision than Kolarik *et al*. [77] for most dictionaries even in the unprocessed stage, which may be explained by updates of the dictionaries since the study by Kolarik *et al*. [111], by our additional criteria to only include entities with a CAS number or InChI string, and by our refined search strategy. The combined version of all dictionaries after executing all the preprocessing steps and disambiguation rules had the highest recall (0.39) but the lowest precision (0.62) compared with all the separate dictionaries using disambiguation (Table 2), which led us to investigate the possibility to exclude resources with low precision to further improve the precision of the combined dictionary without loss of recall. PubChem had the lowest precision (0.58) of all dictionaries before application of the disambiguation rules, which raises questions about the quality of the data. Indeed, in a recent publication by Richard *et al*. [142] concerning chemical information available in databases and through search engines, the quality of chemical information in PubChem was described as 'user beware'. Also Williams [116] expressed concerns about the accuracy of some of the identifiers associated with PubChem compounds. In addition, all resources that we used claim to perform manual curation of the data except for PubChem. When PubChem was left out of the combined dictionary it achieved a precision of 0.67 and a recall of 0.40, both higher than for the combined dictionary without PubChem. When removing the dictionary with the second lowest precision before disambiguation (ChemIDplus: 0.60), the precision of the combined dictionary rose to 0.69 but at the cost of lower recall (0.37) (for comparison, ChemIDplus alone achieved better precision with the same recall; Table 2). Since the removal of PubChem from the combined dictionary improved both the recall and the precision, the combined dictionary without PubChem was used for further analysis. Notably, the curated dictionary with disambiguation rules applied has much higher precision (0.67) than the combined dictionary reported by Kolarik *et al*. [111] (0.13), with a difference in recall of 9 percentage points. The combined dictionary without PubChem contains 1 692 020 terms belonging to 278 577 concepts. Of these concepts, 266 705 have a CAS number and 34 146 have an InChI string. The curated combined dictionary (PubChem excluded) with disambiguation rules applied had the highest *F*-score (*F*=0.50) at a reasonable

precision (0.67), closely followed by the curated version of ChemIDplus with disambiguation rules applied (*F*=0.49, precision=0.71). Overall, the recall was best for the TRIV class of entities (Supplementary Material 6), with ChemIDplus as the best performing dictionary (recall 0.82) and the combined dictionary (PubChem excluded) as a close number two (0.80). The PART class of entities had the lowest recall of all classes (0.00) with the combined dictionary (PubChem excluded) and PubChem as the best performing dictionary (0.04). The PART class is, however, more relevant when the corpus is going to be used for machine learning purposes since parts of chemical names are not expected to be found in dictionaries. This class was, therefore, left out of the error analysis in Section 3.3.

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

Dictionary	Unprocessed			Filtered			Curated			Disambiguation			Kolarik		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
ChEBI	0.21	0.28	0.24	0.58	0.28	0.38	0.63	0.28	0.39	0.71	0.25	0.37	0.13	0.27	0.18
ChemIDplus	0.27	0.41	0.33	0.43	0.4	0.41	0.6	0.4	0.48	0.71	0.37	0.49	–	–	–
DrugBank	0.4	0.22	0.28	0.5	0.22	0.31	0.7	0.21	0.32	0.77	0.19	0.3	0.33	0.13	0.19
HMDB	0.21	0.22	0.21	0.57	0.21	0.31	0.66	0.21	0.32	0.71	0.18	0.29	0.21	0.16	0.18
KEGGc	0.43	0.25	0.32	0.58	0.25	0.35	0.7	0.25	0.37	0.72	0.23	0.35	0.3	0.24	0.27
KEGGd	0.63	0.16	0.26	0.73	0.16	0.26	0.76	0.16	0.26	0.78	0.16	0.27	0.59	0.12	0.2
MeSHchem	0.7	0.23	0.35	0.7	0.23	0.35	0.74	0.23	0.35	0.75	0.22	0.34	0.34	0.27	0.3
MeSHsupp	0.75	0.08	0.14	0.75	0.08	0.14	0.82	0.08	0.15	0.83	0.07	0.13	0.15	0.1	0.12
PubChem	0.24	0.47	0.32	0.39	0.47	0.43	0.58	0.47	0.52	0.73	0.35	0.47	0.15	0.33	0.21
UMLSchem	0.43	0.32	0.37	0.62	0.32	0.42	0.74	0.32	0.45	0.78	0.29	0.42	–	–	–
Combined (PubChem included)	0.18	0.49	0.26	0.36	0.49	0.42	0.51	0.49	0.5	0.62	0.39	0.48	0.13	0.49	0.21
Combined (PubChem excluded)	0.2	0.47	0.28	0.39	0.46	0.42	0.55	0.46	0.5	0.67	0.4	0.5	–	–	–

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

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

To investigate the effect of a general normalization procedure on chemical terms, we ran an analysis using the normalization program *norm* that comes with the LVG normalizer [143]. The LVG normalizer operates after the tokenization has taken place but before disambiguation of terms. The normalization procedure constitutes lower casing each token, converting each token to its base form, ignoring punctuation and sorting the tokens in a multi-token term into alphabetic order. The analysis run resulted in one percentage point lower precision and one percentage point higher recall. The additional terms resulting in the higher recall for the combined dictionary, however, all corresponded to family names being mapped to a single chemical in the dictionary (e.g. diphenols mapped to diphenol), which for the purpose of term identification is to be considered an error. The lower precision was caused by the removal of punctuation, a very important feature of chemical terms, which introduces unnecessary homonyms in the dictionary [e.g. '(-)-Catechol' (CAS 18829-70-4) becomes the same as 'Catechol' (CAS 120-80-9)]. To further illustrate the importance of punctuation in chemical term identification, we ran an analysis using the original tokenizer in Peregrine (Section 2.5). This run resulted in a precision of 0.42 and a recall of 0.40, the much lower precision mainly arising from erroneous partial mapping of terms. In contrast, the original tokenizer in Peregrine has produced good results (precision 0.75, recall 0.76) for a combined dictionary of gene names on the BioCreAtIvE 2 test set [139]. We used an updated version of the combined dictionary of gene names on the same BioCreAtIvE 2 test set with the two different tokenizers, resulting in a precision of 0.74 and recall of 0.81 (*F*=0.77) for the original tokenizer and a precision of 0.76 and a recall of 0.79 (*F*=0.77) for the modified tokenizer. Judged by these results, punctuation is less important for gene names than for chemical names.

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

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

Entity class	Combined	OSCAR3	OSCAR3_dict
IUPAC (391)	0.21	0.82	0.08
PART (92)	0.04	0.84	0.1
SUM (49)	0.29	0.82	0
TRIV (414)	0.8	0.79	0.5
ABB (161)	0.22	0.84	0.08
FAM (99)	0.19	0.84	0.44

Error analysis

We performed a manual error analysis for the combined curated dictionary with disambiguation rules applied and the results from OSCAR3 and OSCAR3_dict. A random set of maximum 25 false negatives from each class (Table 4) and a random set of 50 false positives (Table 5) were analyzed for each approach. We defined six error categories for the false negatives: *partial match* (e.g. only 'azaline' in 'azaline B' was recognized); *annotation error* (e.g. only part of the chemical name has been marked in the text: 'thiophen' in 'thiophene'); *not in dictionary*; *removed by disambiguation* (e.g. single letter 'T'); *removed by manual check of highly frequent terms* (e.g. 'acid'); and *tokenization error* [e.g. 'Ca(2+)' will not be found in the sentence '… free calcium concentration ([Ca(2+)]i) of human peripheral blood lymphocytes…' due to the positioning of the 'i' that does not allow the surrounding brackets to be removed from the entity]. For the false positives, we defined four error categories: *partial match*; *annotation error*; *out of corpus scope* (e.g. larger drug molecules such as protein drugs); *not a chemical* (e.g. '*n*=34' was tokenized and mapped to 'N 34', which is a synonym for Calcium Carbonate). The major reason that entities were not found (i.e. were false negatives) was that they simply were not in the combined curated dictionary or the dictionary in OSCAR3_dict, or for OSCAR3, were not recognized by the NER algorithm (Table 4). For the combined curated dictionary, this holds true for all classes except ABB, for which a larger part was removed during the disambiguation step. This is not surprising since abbreviations are notoriously ambiguous and difficult to resolve. For OSCAR3, the exceptions are instead the IUPAC class, where a majority of the false negatives were only partially found and the SUM class for which the tokenizer performed poorly. For the false positives, it was clear that the corpus is not optimal for a dictionary that aims at both small molecules and drugs, since larger drug molecules have not been annotated in the corpus. This was true for 42% of the entities in the random set for the combined dictionary approach, 32% of the entities in the random set for the OSCAR3_dict and 26% of the entities in the random set for OSCAR3 (Table 5). Another major source for the false positives using all approaches was partial matches of longer

chemical names. For OSCAR3, it can be noted that it recognized a higher percentage of non-CM than the combined dictionary and OSCAR3_dict.

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

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

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

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

Discussion

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

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

precision of 0.67 in a different light. If these false positives would be excluded from the analysis, the combined dictionary would have a precision of 0.90.

According to our study, a recall of 0.49 (at a precision of 0.51) would be the highest achievable recall for a pure dictionary approach to term recognition and mapping. This is the recall reached by the curated combined dictionary (PubChem included) without disambiguation rules applied. Kolarik *et al*. [111] reached the same recall at a precision of 0.13. If higher recall is desired, an approach such as has been implemented in OSCAR3, i.e. a combined NER approach using machine learning together with a dictionary, would be the better choice. This approach has, however, the disadvantages of lower precision (at least on the corpus used in this study) and an incomplete mapping of entities to external data sources. The precision of OSCAR3 on the corpus (0.45) is lower than what has been reported by Corbett *et al*. [144] on a non-public PubMed corpus (0.75), but the recall (0.82) is better (Corbett *et al*. reported a recall of 0.74). Notably, many (44%) of the false positives arising from OSCAR3 fell under the non-CM error category. These were mainly abbreviations of entities such as 'CNS' for 'Central Nervous System' or 'AD' for 'Alzheimer Disease' or text structures that resemble CM such as '11*a*–*c*' or 'IC(50)'. The false positives arising from non-chemical abbreviations could possibly be removed with the use of the disambiguation rules described in this study. If the false positives that were due to corpus mismatch and annotation errors are removed from the calculation, the precision is still lower (57.3%) but at least closer to the one earlier reported. The difference in precision and recall can be due to differences in the annotation scheme of CM underlying the training corpus used in OSCAR3 and the corpus by Kolarik *et al*. [111]. The dictionary in OSCAR3 had a lower recall than the combined dictionary [precision 0.68 (0.84 when corrected for corpus mismatch and annotation errors) and recall 0.25 versus precision 0.67 (0.90 when corrected for corpus mismatch and annotation errors) and recall 0.40], which suggests that the dictionary in OSCAR3 would benefit from a combined dictionary approach. However, embedding the combined dictionary from this study in OSCAR3 is out of the scope of this article and we suggest this for future research.

In our study and in the study by Kolarik *et al*. [111], an important class with low recall was IUPAC. The main reason for not finding these entries was that they simply were not present in the combined dictionary, even though IUPAC-like names had been added when available. Clearly, dictionary-based term identification is not capable of identifying multiple-term systematic names to a satisfactory extent since not enough of these types of names are available in current resources. If only a synonym for an entity is missing, this might be solved by term variant generation but if the whole entity is missing from the dictionary it can only be solved by adding the entity to the dictionary. Spelling errors might be helped by fuzzy matching [145-148], with a possible cost to precision. In contrast, machine-learning or rule-based systems have reported good performance for the recognition of multiple-term systematic names [e.g. Klinger *et al*. reported an *F*-score of 0.82 on a PubMed corpus for their method based on conditional random fields (CRFs) and CRFs was also used in a high proportion of entries in the latest BioCreative evaluation [149], OSCAR3 had a recall of 0.82 for the IUPAC class of entities on the corpus used in this study, Corbett and Copestake an *F*-score of 0.83 for a system of cascaded classifiers on a PubMed corpus and Wren [150] a recall of 0.93 with an average precision of 0.83 (depending upon the cutoff score used) for a first order Markov model on a PubMed corpus], but then the problem remains of mapping a term to its referent data source.

The lower recall of 0.40 for the combined dictionary with the disambiguation rules applied compared to without disambiguation is foremost due to the problem associated with the disambiguation of abbreviations and summary structures. Yu *et al*. [151] divided the problem of disambiguating abbreviations into two types. First, abbreviations may be disambiguated ('defined') near their occurrence in the text. The second type of abbreviation appears without the intended full form nearby. This second type of abbreviation is more prevalent and harder to disambiguate [151, 152]. Abbreviations and summary structures of chemicals are of the second type, in the sense that they are used in abstracts to a large extent without the long form of the term, which will cause these

entities to be removed since there is not enough extra information to make sure that they actually represent a CM. Using full text articles instead of abstracts might be an answer but unfortunately there has been a report of high (75%) occurrence of abbreviations without their long forms also in full text articles [151]. To resolve this, another way of taking the context into account is needed, using for example document labeling. If a document is labeled, a term could be assigned directly if it was not an in-dictionary homonym.

Conclusions

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

Supplementary information:

Supplementary data are available at *Bioinformatics* online:
http://bioinformatics.oxfordjournals.org/content/25/22/2983/suppl/DC1

Chapter 4

Comparing automatic and manual chemical term curation

Kristina M Hettne[1,2] , Antony J Williams[3] , Erik M van Mulligen[1] , Jos Kleinjans[2] , Valery Tkachenko[3] and Jan A Kors[1]

1 Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands
2 Department of Health Risk Analysis and Toxicology, Maastricht University, Maastricht, The Netherlands
3 Royal Society of Chemistry, 904 Tamaras Circle, Wake Forest, NC-27587, USA

Published as: **Automatic vs. manual curation of a multi-source chemical dictionary: the impact on text mining** in *Journal of Cheminformatics* 2010, **2**:3
and
Correction: Automatic vs. manual curation of a multi-source chemical dictionary: the impact on text mining in Journal of Cheminformatics 2010, 2:4

Abstract

Background

Previously, we developed a joint chemical dictionary (Jochem) for the identification of small molecules and drugs in text based on a number of publicly available databases and tested it on an annotated corpus. To achieve an acceptable recall and precision we used a number of automatic and semi-automatic processing steps together with disambiguation rules. However, it remained to be investigated which impact an extensive manual curation of a multi-source chemical dictionary would have on chemical term identification in text. ChemSpider is a chemical database that has undergone extensive manual curation aimed at establishing valid chemical name-to-structure relationships.

Results

We acquired the component of ChemSpider containing only manually curated names and synonyms. Rule-based term filtering, semi-automatic manual curation, and disambiguation rules were applied. We tested the dictionary from ChemSpider on an annotated corpus and compared the results with those for the Jochem dictionary. The ChemSpider dictionary of ca. 80 k names was only a 1/3 to a 1/4 the size of Jochem at around 300 k. The ChemSpider dictionary had a precision of 0.43 and a recall of 0.19 before the application of filtering and disambiguation and a precision of 0.87 and a recall of 0.19 after filtering and disambiguation. The Jochem dictionary had a precision of 0.20 and a recall of 0.47 before the application of filtering and disambiguation and a precision of 0.67 and a recall of 0.40 after filtering and disambiguation.

Conclusions

We conclude the following: (1) The ChemSpider dictionary achieved the best precision but the Jochem dictionary had a higher recall and the best F-score; (2) Rule-based filtering and disambiguation is necessary to achieve a high precision for both the automatically generated and the manually curated dictionary. ChemSpider is available as a web service at http://www.chemspider.com/ and the Jochem dictionary is freely available as an XML file in Simple Knowledge Organization System format on the web at http://www.biosemantics.org/Jochem.

Background

Finding chemical terms in free text is essential for text mining aimed at exploring how chemical structures link to biological processes [117]. However, the techniques behind current text mining applications have mainly focused on the ability of the system to correctly identify gene and protein names in text, while less effort has been spent on the correct identification of chemical names [2,3][91, 112]. This is however about to change as more and more chemical resources are becoming freely available [114-116]. For example, resources such as DrugBank [122] and the Unified Medical Language System metathesaurus (UMLS) [93] have been applied for the identification of drug names in text [123, 124] (for a recent review of literature mining in support of drug discovery see Agarwal and Searls [125]). Briefly, the challenges of chemical name identification differ from the ones in the genomics field in the sense that the exact placement of tokens such as commas, spaces, hyphens, and parentheses plays a much larger role. Chemical named entity recognition (NER) in general has been reviewed by Banville [117] and methods for confidence-based chemical NER have been evaluated by Corbett and Copestake [118].

In this paper we focus on the task of term *identification*, which goes beyond NER to also include term *mapping*, i.e. the linking of terms to reference data sources. In the case of chemicals, they can also be identified by a specific structure representation such as a connection table, an InChI string or a simplified molecular input line entry specification (SMILES). To achieve this, a dictionary with database links, or structures, is essential. Naturally, the usefulness of the dictionary approach depends on the coverage of terms in the dictionary for the particular domain and how well the terms are suited for natural language processing. Previously, we developed a combined dictionary named Jochem for the identification of small molecules and drugs in text based on a number of publicly available databases and tested it on an annotated corpus [110]. To achieve an acceptable precision (0.67) and recall (0.40) we used a number of automatic and semi-automatic processing steps together with disambiguation rules. However, it remained to be investigated which impact an extensive manual curation of a multi-source chemical dictionary would have on chemical term identification in text. We expect that a higher precision can be reached with a manually curated dictionary.

Around 8% of the chemicals in Jochem contain structure information in the form of InChI strings. It should be noted that we did not validate the correctness of the association between the chemical names and the chemical structures/compounds as that was not the focus of the work. The challenges of chemical NER are clearly not limited only to the identification and extraction of a particular chemical name but also the association of the chemical name with an appropriate chemical structure or compound. ChemSpider [153] is an online database of chemical compounds and associated data and was developed with the intention of building a structure-centric database for the chemistry community. The chemicals contained within the database are sourced from over 200 different data sources including chemical vendors, government databases, commercial databases, open notebook science projects, blogs and personal chemistry collections deposited by members of the community. During the process of integrating and associating data from various sources the ChemSpider development team has identified a multitude of issues in regards to the quality of chemical structure representations. These include varying levels of accuracy in stereochemistry, the mis-association of chemical names with chemical entity and a myriad of other issues whereby chemical names are associated with incorrect chemical structures. The challenge here is one of assertion - what is a "correct" chemical structure and who asserts that it has a specific representation? While the chemical structure of benzene can be represented as either a series of alternating single and double bonds or as in a Kekule form, the connection table of atoms and bonds as captured in an electronic format remains consistent. In terms of compounds of biological interest the structure representation for a particular drug is based on the collective wisdom of the company registering the compound, the patent representation and a multitude of databases containing associated

information. The challenges of both conventions and assertions are taken into account when creating a validated dictionary of chemical names and associated structure representations.

As a result of the challenges associated with poor quality chemical name-structure relationships ChemSpider was developed to include a curation platform whereby chemists could participate directly in the validation of the relationships. A web-based interface to approve, delete and add chemical names to chemical entities was delivered and a multi-level curator role was established so that when members of the community made suggested changes to the relationships master curators would then further investigate and approve their work. ChemSpider was released to the community in March 2007 and many tens of thousands of curation actions have provided a highly curated dictionary.

The objective of this study is to determine the impact of manual curation of chemical name-structure relationships on the precision and recall of chemical term identification.

Results

The ChemSpider dictionary was filtered according to a set of pre-processing steps and tested on an annotated corpus (see Methods for details on the pre-processing steps and the corpus). Before pre-processing, the ChemSpider dictionary contained 157,173 terms belonging to 84,065 entities and after pre-processing 160,898 terms belonging to 84,059 entities. The processed version of Jochem contains 1,692,020 terms belonging to 278,577 entities. Dictionary term strings that matched the start and end positions of the chemical term strings in the corpus constituted true positives (TP), term strings that were not marked as chemical term strings in the corpus but still matched a dictionary term string were false positives (FP), and chemical term strings in the corpus that were not matched were false negatives (FN). Recall (R), precision (P), and F-score were computed in the usual way:

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

Table 1 shows the effect of pre-processing and disambiguation on precision and recall for the dictionaries. It is clear that the pre-processing steps and the disambiguation rules have a strong positive influence on the precision of both dictionaries. The ChemSpider dictionary had higher precision (0.87) and lower recall (0.19) compared to the Jochem dictionary (precision 0.67 and recall 0.40). The Jochem dictionary had the highest F-score (0.50). A combination of both dictionaries showed changes of less than 1 percentage point in recall and precision values (results not shown). The combination was created by matching concepts on CAS numbers and/or InChI strings, resulting in a merged dictionary with 317, 275 concepts. The overlap between the dictionaries was calculated to 45,361 concepts.

Overall, the recall was best for the TRIV class of entities (Table 2), with Jochem as the best performing dictionary (recall 0.80). The PART class of entities had the lowest

Table 1. Precision (P), recall (R) and F-score (F) of the dictionaries on the annotated corpus.

Dictionary	Unprocessed			Filtered			Frequent terms correction			Disambiguation		
	P	R	F	P	R	F	P	R	F	P	R	F
ChemSpider	0.43	0.19	0.26	0.81	0.19	0.31	0.85	0.19	0.31	0.87	0.19	0.31
Jochem	0.2	0.47	0.28	0.39	0.46	0.42	0.55	0.46	0.5	0.67	0.4	0.5

recall of all classes (0.00, ChemSpider dictionary). The PART class is however more relevant when the corpus is going to be used for machine learning purposes since parts of chemical names are not expected to be found in dictionaries. This class was therefore left out of the error analysis below.

Table 2. Recall values for the entity classes per dictionary.

Entity class	ChemSpider	Jochem
IUPAC (391)	0.08	0.21
PART (92)	0	0.04
SUM (49)	0.25	0.29
TRIV (414)	0.45	0.8
ABB (161)	0.01	0.22
FAM (99)	0.02	0.19

IUPAC: multiword systematic names, PART: partial chemical names, SUM: sum formulas, TRIV: trivial names (including single word IUPAC names), ABB: abbreviations, FAM: chemical family names.

Error analysis

We performed a manual error analysis for the dictionaries with disambiguation rules applied (see Methods). The major reason that entities were not found (i.e., were false negatives) was that they simply were not in the dictionaries (Table 3). For the Jochem dictionary, this holds true for all classes except ABB for which most belong to the category "removed by disambiguation". The major source of false positives for both dictionaries was partial matches of longer chemical names (Table 4). Notably, ChemSpider only had one entity out of corpus scope and no entities that were non-

Table 3. Error analysis of a random sample of max 25 false negatives from each class for ChemSpider (CS) and Jochem (CL).

Error type	TRIV		SUM		IUPAC		FAM		ABB	
	CS	CL	CS	CL	CS	CL	CS	CL	CS	CL
Partial match	0	3	0	0	0	0	0	0	0	0
Annotation error	0	2	0	0	0	1	0	0	0	0
Not in dictionary	25	15	22	16	25	24	25	24	25	8
Removed by disambiguation	0	5	0	7	0	0	0	1	0	12
Removed by manual check of highly frequent terms	0	0	0	1	0	0	0	0	0	2
Tokenization error	0	0	3	1	0	0	0	0	0	3

chemicals.

Discussion

The Jochem dictionary had the highest recall and the best F-score, but a lower precision than the ChemSpider dictionary. The precision of 0.87 (at a recall of 0.19) for the ChemSpider dictionary is the best reported for a chemical dictionary on the corpus used in this study. From the analysis of the false positives it was obvious that the ChemSpider dictionary was less out of the scope of the corpus and contained less non-chemical names than Jochem. As mentioned in previous work [110], the false positives in the categories *entity out of corpus scope* and *annotation error* might possibly be excluded from the analysis because these errors cannot be attributed to the dictionaries. When the false positives from these categories were excluded, Jochem had a precision of 0.82 and ChemSpider a precision of 0.91.

Table 4. Error analysis of the false positives (percentage) for ChemSpider and Jochem.

Error type	False positives	
	ChemSpider	**Jochem**
Partial match	21 (64%)	96 (41%)
Annotation error	11 (33%)	29 (13%)
Out of corpus scope	1 (3%)	79 (34%)
Not a chemical	0	28 (12%)

Worth noticing is that the precision for the manually curated dictionary from ChemSpider on the corpus without the use of the pre-processing steps was about half compared to the processed version. A reason for this might be that the dictionary was not curated with text-mining purposes in mind. For example, synonyms such as "As" for "Arsenic" might be correct but will give rise to many false positives when the dictionary is used for text mining. However, it should be noted that the ChemSpider team used their own curated dictionaries as the basis of their semantic markup approaches on the ChemMantis [154] platform. Their entity extraction approach accounted for direct identification of elements and included a list of stop words to allow for improved precision.

The recall for the Chemspider dictionary is substantially lower than that of Jochem in all categories except for the SUM class. It is to be expected that the ChemSpider dictionary scores lower for the FAM class since ChemSpider is a structure-centered database, but the relatively low recall for the IUPAC, TRIV and ABB classes were surprising. We therefore performed a search in the online version of ChemSpider (August 8, 2009) for the false negatives in these classes. Indeed, an additional 3 of the random 25 IUPAC false negative, 20 of the 25 random TRIV false negatives, and 10 of the 25 random ABB false negatives were found in the online version of ChemSpider. These differences might be explained by the update speed of the online ChemSpider database as hundreds of thousands of chemical entities can be added within a week. As of September 2009 there are eight million chemical entities waiting to be deduplicated into the ChemSpider database and there has been an increase of almost 10% in the unique number of chemical entities since this manuscript was started. There are presently over 23 million unique chemicals in the database.

Since despite the increased volume of ChemSpider only three of the 25 random IUPAC false negatives were found in the online version of ChemSpider, we performed a structural evaluation of the remaining 22 false negatives. This deep analysis of the structures of the false negatives from the IUPAC class highlights three different issues: firstly the annotation of the corpus, secondly the sometimes inconsistent or incorrect way scientists write chemical names in articles, and thirdly the ChemSpider database coverage.

The annotation issues follow. Two chemicals were annotated as IUPAC in the corpus but did not respond to unique structures (e.g. hexa-acetyl was annotated as IUPAC in the sentence "...it formed a **hexa-acetyl** derivative..."). We argue that these chemicals should be annotated as PART instead. Two cases were not chemical names but internal abbreviations in the abstract (e.g (S)-(-)-3-PPP). We argue that these should belong to the ABB class instead. These four cases reflect the relatively low annotator agreement on the corpus (80%) [111]. One annotation error concerns two chemicals after each other that were annotated as one in the sentence "On interaction with anhydrous **potassium acetate 14-bromcarminomycinone** (III) yield 14-acetoxycarminomycinone (IV)". Five annotation "errors" were family names (e.g. 1-(carboxyalkyl)hydroxypyridinones). These cases are however not annotation errors according to the class definition of the FAMILY class in Kolarik et al. [111], where "Substances used as bases for building various derivatives and analogs were tagged as IUPAC, not as FAMILY (e.g. 1,4-dihydronaphthoquinones)", but they are not expected to be found in ChemSpider since ChemSpider focuses on single compounds.

The way chemical names are written in articles concern the following cases: two were too generic to correspond to unique structures (e.g. 5-O-tetradecanoyl-2,3-

dideoxy-L-threo-hexono-1,4-lactone), and six were non-systematic names for which no structure could be drawn (e.g. N-(trifluoroacetyl)-14-phenyl-14-selenaadriamycin). Since ChemSpider strives to include only valid structures, these names are not expected to be found using ChemSpider. The poor quality of chemical names in common usage was discussed by Bretcher [155] already in the year 1999 and is 10 years later still an issue. Although many chemistry journals nowadays have rules about the naming of compounds and demands on the addition of structure information, this information has not always been updated for older issues, and unfortunately few MEDLINE abstracts contain structure information.

Database coverage applies to the following cases: three chemicals were present in the database as structures but lacked the specific synonym used in the abstract, and one was not present at all in the database at the time of this study (8-(methylthio)-1,2,3,4,5,6-hexahydro-2,6-methano-3-benzazocine) but has since been added to the database. These cases therefore fit into the *not in dictionary* category.

The fact that many of the random false negatives were found in the online ChemSpider database put the low recall of the manually curated ChemSpider dictionary in a different light. The ongoing online community-based annotation of chemical names in ChemSpider will ensure an increase in recall of the dictionary while hopefully maintaining the precision, and surely the important link to chemical structure. Jochem requires a more thorough accuracy check of text-mining results due to the lower precision compared to the ChemSpider dictionary but will retrieve more entities. On the other hand, in contrast to ChemSpider, Jochem is downloadable in its whole and can be used as a basis for the creation of a manually curated chemical dictionary for text mining. Structure information can be added for the entities lacking this information once it is available in the underlying databases.

Partial matches of compounds were an important issue for both dictionaries and something that might be solved by detection of chemical name boundaries before matching. However, the false positives in this category did not decrease when a system that uses this type of information (OSCAR3, available at http://sourceforge.net/projects/oscar3-chem) was tested [110] and further testing of different algorithms for chemical name boundary detection in combination with dictionary look-up is needed.

Conclusions

We conclude the following: (1) The Jochem dictionary had the highest recall (0.40) and the best F-score (0.50), but a lower precision (0.67) than the ChemSpider dictionary; the ChemSpider dictionary achieved the best precision (0.87) but at a cost of lower recall (0.19) than the Jochem dictionary; It should be noted that the ChemSpider dictionary of ca. 80 k names was only a 1/3 to a 1/4 the size of Jochem at around 300 k and this would be expected to dramatically impact recall. (2) Rule-based filtering and disambiguation is necessary to achieve a high precision for both the automatically generated and the manually curated dictionary.

Experimental

Dictionary pre-processing

The combined chemical dictionary Jochem has been described elsewhere [110]. Briefly, it is based on the following resources: the chemical part of the Unified Medical Language System metathesaurus (UMLS) [93], the chemical part of the Medical Subject Headings (MeSH) [130], the ChEBI ontology [120], DrugBank [122], KEGG drug [156], KEGG compound [134], the human metabolome database (HMDB) [135], and ChemIDplus [157]. Data from the fields used for entry term, synonyms, summary structure, and database identifiers were used to build the Jochem dictionary. CAS registry numbers [158] and Beilstein reference numbers [159] were not used for text mining due to their presumed ambiguity with other number types in text. CAS numbers do have a specific format that should help identify them in text and might be included as synonyms in

future releases of the dictionary. Entries were merged if they had the same CAS number, database identifier (cross-reference), or InChI string. No manual curation was performed to ensure the correctness of merged entities. A set of rules was used to rewrite and suppress terms in the dictionary and a manual check for highly frequent terms was performed [110]. Briefly, we *removed* a term if (1) the whole term after tokenization and removal of stop words is a single character, or is an arabic or roman number (e.g. "T" as an abbreviation for "Tritium"); (2) the term contained any of the following features: a dosage in percent, gram, microgram or milliliter, "not otherwise specified", "not specified", or "unspecified", "NOS" at the end of a term and preceded by a comma, or "NOS" within parentheses or brackets at the end of a term and preceded by a space, "other" at the beginning of a term and followed by a space character or at the end of a term and preceded by a space character, "deprecated", "unknown", "obsolete", "miscellaneous", or "no" at the beginning of a term and followed by a space character (e.g. "unspecified phosphate of chloroquine diphosphate" as synonym for "chloroquine diphosphate"); (3) the term corresponded to a general English term in the top 500 most frequent terms found in a set of 100,000 randomly selected MEDLINE abstracts indexed with the Jochem dictionary. We *added* (1) the syntactic inversion (e.g. "acid, gamma-vinyl-gamma-aminobutyric" is rewritten to "gamma-vinyl-gamma-aminobutyric acid"); (2) the stripped possessive version (e.g. "Ringer's lactate" rewritten to "Ringer lactate"); (3) the long form and short form version of a term (e.g. "Hydrogen chloride (HCL)" is split into "Hydrogen chloride" and "HCL"). Since a rewritten term will be added to the dictionary without removing the original term, an increase in synonyms after using the rewrite rules will take place.

We acquired a dictionary subset from the chemical database ChemSpider (February 12, 2009), containing only manually annotated names and synonyms. Before manual curation, robots had been used to ensure that there were no inappropriate correspondences between the chemical names and the chemical structures. For example, it is rather common in the public databases to have the chemical names of salts despite the fact that the chemical itself may be a neutral compound. A series of processing runs to clean up mis-associations in the following manner improved the validity of names associated with structures: 1) for names containing chloride, bromide, iodide, and fluoride check the molecular formulae for the presence of the associated halogens in the molecular formula and treat as necessary; 2) for names containing nitrite, nitrate, sulfate/sulphate, and sulfite/sulphite, check molecular formulae for presence of nitrogen or sulphur and remove names as necessary; 3) for hydrate/dihydrate, check for presence of one or more waters of hydration and remove names as appropriate; 4) convert names to chemical structures using commercial software tools and check for consistency and flag as checked by robots. This is a different level of curation than checked by humans. The manually annotated names are those approved primarily by users of ChemSpider and then further validated by master curators. The result is a highly curated database of chemical structures with their associated manually curated identifiers. These identifiers are not limited to systematic names and trade names but also include CAS registry numbers, EINECS or ELINCS numbers [160] and Beilstein reference numbers. In order to make a fair comparison with the Jochem dictionary, we applied the same filtering rules and manual check for highly frequent terms to the ChemSpider dictionary as were previously applied to the Jochem dictionary. This time, the manual check for highly frequent terms was based on a MEDLINE indexation using the ChemSpider dictionary.

Term identification

We used our concept recognition software Peregrine [139] to index a corpus of annotated chemical abstracts from Kolarik et al. [111] http://www.scai.fraunhofer.de/chem-corpora.html . The Peregrine system translates the terms in the dictionary into sequences of tokens. When such a sequence of tokens is found in a document, the term, and thus the chemical associated with that term, is recognized. Some tokens are ignored, since these are considered to be non-informative ('of', 'the', 'and', 'in'). The tokenizer in Peregrine considers everything that is not a letter or a digit to be a word delimiter.

Similar to Hettne et al. [110], we made the following adjustments to the tokenizer: full stops, commas, plus signs, hyphens, single quotation marks and all types of parentheses ((, {, []) were excluded from the word delimiter list. After tokenization, the tokens were stripped of trailing full stops, commas and non-matching parentheses. Parentheses were also removed if they surrounded the whole token. In addition, a list of common suffixes was used to remove these suffixes at the end of tokens [59]. We used Peregrine with the following settings: case-insensitive, word-order sensitive and largest match.

The annotated corpus consists of 100 MEDLINE abstracts with 1206 annotated chemical occurrences divided into the following groups: multiword systematic names (IUPAC, 391 occurrences), partial chemical names (PART, 92 occurrences), sum formulas (SUM, 49 occurrences), trivial names (including single word IUPAC names) (TRIV, 414 occurrences), abbreviations (ABB, 161 occurrences), and chemical family names (FAM, 99 occurrences). Larger drug molecules such as protein drugs had not been annotated in the corpus [111]. The creators used a simple system for detecting IUPAC names [77] to select abstracts containing at least one found entity. Next to abstracts selected with this procedure, they selected abstracts containing problematical cases as well as abstracts containing no entities. The inter-annotator F1 was 80% when recognizing the boundaries without considering the different classes.

We indexed the corpus using three versions of the ChemSpider dictionary: unprocessed, filtered (after application of the filtering rules), and frequent terms correction (after the check for frequent English terms). To compare the effect of disambiguation rules during the indexing process we used the same rules as in Hettne et al. [13]. That is, we first determine whether a term is a dictionary homonym, i.e., if it refers to more than one entity in the dictionary. If the term is a dictionary homonym, but it is the preferred term of that entity, it is further handled as if it is not a dictionary homonym. If the term is not a dictionary homonym it still needs further processing since it can have many meanings in text. Therefore, terms that are shorter than five characters or do not contain a number are also considered potential homonyms, and require extra information to be assigned. A (potential) homonym is only kept if (1) another synonym of the entity is found in the same piece of text; (2) a keyword (i.e., a word or "token" that occurs in any of the long-form names of the small molecule, and appears less than 1000 times in the dictionary as a whole) is found in the same piece of text. The results from the ChemSpider dictionary were compared to the results previously reported for the Jochem dictionary.

Error analysis

A random set of maximum 25 false negatives from each class of entities in the corpus, the 232 false positives for Jochem, and the 33 false positives for the ChemSpider dictionary were analyzed. For comparison, we used the same error categories for the false negatives and false positives as in Hettne et al. [110]. For the false negatives, these were: *partial match* (e.g. only "beta-cyclodextrin" in "hydroxypropyl beta-cyclodextrin" was recognized); *annotation error* (e.g. only part of the chemical name has been marked by the annotators in the text: "thiophen" in the sentence"... Gewald *thiophen***e** synthesis was...", or a whole entity has been overlooked by the annotators); *not in dictionary*; *removed by disambiguation* (e.g. single letter "T"); *removed by manual check of highly frequent terms* (e.g. "Me"); and *tokenization error* (e.g. "Ca(2+)" will not be found in the sentence "...free calcium concentration ([Ca(2+)]i) of human peripheral blood lymphocytes..." due to the positioning of the "i" that does not allow the surrounding brackets to be removed from the entity). The error categories for the false positives were: *partial match*; *annotation error*; *out of corpus scope* (e.g. larger drug molecules such as protein drugs); *not a chemical* (e.g. the term "metabolite").

Chapter 5

Literature-aided interpretation of gene expression data

Rob Jelier[1] ,Jelle J. Goeman[2], Kristina M. Hettne[3,4], Martijn J. Schuemie[3], Johan T. den Dunnen[5], Peter A.C. 't Hoen[5]

[1]EMBL-CRG Systems Biology Research Unit, Barcelona, Spain
[2]Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands
[3]Department of Medical Informatics, Erasmus Medical Centre, Rotterdam, The Netherlands
[4]Department of Health Risk Analysis and Toxicology, Maastricht University, Maastricht, The netherlands
[5]Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands

Published as: **Literature-aided interpretation of gene expression data with the weighted global test**. *Briefings in bioinformatics* 2011.

Abstract

Most methods for the interpretation of gene expression profiling experiments rely on the categorization of genes, as provided by the Gene Ontology (GO) and pathway databases. Due to the manual curation process, such databases are never up-to-date and tend to be limited in focus and coverage. Automated literature mining tools provide an attractive, alternative approach. We review how they can be employed for the interpretation of gene expression profiling experiments. We illustrate that their comprehensive scope aids the interpretation of data from domains poorly covered by GO or alternative databases, and allows for the linking of gene expression with diseases, drugs, tissues and other types of concepts. A framework for proper statistical evaluation of the associations between gene expression values and literature concepts was lacking and is now implemented in a weighted extension of global test. The weights are the literature association scores and reflect the importance of a gene for the concept of interest. In a direct comparison with classical GO-based gene sets, we show that use of literature-based associations results in the identification of much more specific GO categories. We demonstrate the possibilities for linking of gene expression data to patient survival in breast cancer and the action and metabolism of drugs. Coupling with online literature mining tools ensures transparency and allows further study of the identified associations. Literature mining tools are therefore powerful additions to the toolbox for the interpretation of high-throughput genomics data.

Background

Gene expression profiling has become an important technology in modern molecular biology, but the interpretation of gene expression data is still a challenging task. Many gene expression studies reveal expression changes in large numbers of genes. The characterization of these changes, for instance in terms of involved biological processes, remains difficult and requires an overview of the very large amount of information on gene function currently available. Gene annotation databases are commonly used in characterization efforts. They group genes according to a shared feature, such as involvement in the same biological process. KEGG [161], Biocarta and NetPath [162] are among the best known catalogues for metabolic and signal transduction pathways. Alternative gene annotation schemes based on protein function, localization and expression regulation are available from Gene Ontology (GO) [163] and the molecular signature database (MSigDB) [20].

The gene annotation databases are intensively used to identify functional categories, represented by gene sets, which show an association with the gene expression data. A wide variety of methods have been proposed for this task (see ref. [164] for an overview). In brief, three different types of statistical tests can be discriminated: (i) tests for the overrepresentation of a gene set in a list of differentially expressed genes using a hypergeometric or equivalent test (see ref. [165] for an overview); (ii) methods that use the *p*-values of all the genes [166, 167]. Well known is the gene set enrichment analysis (GSEA), that uses ranked *p*-values and tests whether the ranks of genes in a gene set differ from a uniform distribution [20, 168]; (iii) regression analyses that use the actual expression levels of the genes in the gene set and test whether these are associated with the studied phenotype, an example is the global test [9, 169, 170].

A serious argument can be made in favor of this last type of tests. The resulting *p*-value has a clear interpretation in the context of the experiment (the probability there are no differentially expressed genes in the gene set), confounders can be included in the model and the test procedure can be generalized from a test for a single gene (a gene set with size one) to a gene set containing all genes. For a further discussion of methodological issues we refer to the review by Goeman and Bühlmann [171].

Results and discussion

Automated gene annotation

The construction of gene annotation databases is mostly a manual process in which genes are annotated based on information in scientific publications [172]. Due to its labor-intensive nature, manual annotation efforts struggle to keep up-to-date [173] and focus on a limited subject area. Also, these databases provide a black and white view: a gene is either part of a category or not. This implies that some inclusion criterion must be used during the annotation, and that all genes are of similar importance to what the gene set represents. However, this may not accurately reflect biology and is not very flexible.

Automated text mining can complement manual approaches, as the automation can provide a broader scope, as well as that it more easily provides up-to-date information and adaptability. The field of text mining has grown rapidly in recent years, with several applications for gene expression data analysis. A selection of web-based tools is given in Table 1. The majority of tools work with gene lists. They can retrieve concepts or terms strongly associated to the selected genes, [174] and/or cluster the genes to retrieve functionally coherent subclusters [36, 37, 39, 64, 81-83, 175-181]. The main differences between methods are on the following three aspects: First, which information is retrieved from the texts. One approach is to rely on the words of the texts directly [81, 180], with some approaches using the automatic combination of related words through a factor analysis to reduce the dimensionality [176, 182]. Another approach relies on a thesaurus and a tagging engine to identify thesaurus entries in texts [36, 39, 64]. Second, methods vary in how texts are linked to genes. Some tools rely on

thesaurus based approaches to identify references to genes in texts [37, 64], whereas others rely on automated Pubmed queries [179, 180] or manual gene to document links such as provided by NCBI's Entrez gene [181, 182]. Third, the way the associations between terms and genes are calculated. Some approaches focus on direct co-occurrences between genes or concepts in documents [39], whereas other approaches allow indirect relations, e.g. two genes regularly co-occur with the same term, to play a role in a variety of ways [82, 183]. Apart from the tools that focus on retrieving gene relations from a list of genes, some methods retrieve functional associations shared between gene lists of different experiments [49, 66]. Finally, text mining can be used in combination with other resources in an integrated framework to retrieve functional associations between the genes [184].

Table 1. Useful websites for literature-aided interpretation of gene expression profiling data

Name	Description	Website
Anni	Versatile text mining tool. Exploration of associations used in the literature weighted globaltest.	www.biosemantics.org/anni
Babelomics	Platform for the analysis of transcriptomics, proteomics and genomic data with functional profiling.	babelomics.bioinfo.cipf.es
Biocarta	Pathway database.	www.biocarta.com
ConceptGen	An enrichment testing and concept mapping tool that includes MeSH-based gene sets	conceptgen.ncibi.org
CoPub	A text-mining based enrichment testing tool	services.nbic.nl/cgi-bin/copub3/CoPub.pl
GenCLiP	Clustering of gene lists by literature profiling and constructing gene co-occurrence networks	www.genclip.com
Gene Ontology	Controlled vocabulary for the functonal annotation of genes.	www.geneontology.org
Gene2MeSH	Contains co-occurrences between genes and MeSH terms	gene2mesh.ncibi.org
Genelist Analyzer	Statistical evaluation of overrepresentation of literature concepts in gene lists	workerbee.igb.uiuc.edu:8080/BeeSpace/Search.jsp
Global test	Homepage of the global test R package.	www.bioconductor.org/packages/release/bioc/html/globaltest.html
Hanalyzer	Gene network visualization and reasoning tool based on literature, ontology and database mining	hanalyzer.sourceforge.net
Literature weighted global test	Described in this paper	biosemantics.org/weightedglobaltest
KEGG	Metabolic and regulatory pathway database	www.genome.jp/kegg
MILANO	Automated searches in Medline for co-occurrence with gene-based search terms	milano.md.huji.ac.il
MSigDB	Molecular signatures database, gene sets that accompany the GSEA test	www.broadinstitute.org/gsea/msigdb
NetPath	Curated signal transduction pathways in humans	www.netpath.org
Pubgene	Contains module that displays literature co-occurrence networks	www.pubgene.org

Literature-based annotation and statistical testing

The analyses performed by the mentioned text mining-based approaches are of an exploratory nature, and do not provide a statistical evaluation for the identified associations in the context of the performed experiment. However, text mining algorithms can readily be combined with the three previously mentioned classes of statistical approaches for evaluating gene annotation categories. A class one approach could simply entail the creation of gene sets, for instance by applying a threshold on the literature derived association scores between genes and biomedical concepts. Sartor *et al***.** [67] provide literature-based gene sets in their tool ConceptGen, which uses Gene2MeSH (http://gene2mesh.ncibi.org) to identify gene and MeSH term pairs with a significantly higher number of co-occurrences than expected by chance. Frijters *et al***.** [69] and Leong and Kipling [70] calculate biomedical term over-representation for a set of regulated genes in a similar fashion to standard class one over-representation tools.

Several text-mining approaches have been published that resemble the earlier mentioned class two, GSEA-like approach. Kueffner *et al***.** [185] integrate the rank of the genes after sorting on *p*-value with an analysis of the literature. However, their approach is based on factorization which complicates the interpretation of their results, and does not include formally testing retrieved associations. Minguez *et al***.** [72] test if a ranked list of genes shows a significant correlation with the genes' associations to a biomedical term. These associations are based on the literature and reflect the extent to which rate a gene and a biomedical term occur together in documents exceeds the rate expected by chance.

However to our knowledge, no class three, or regression analysis based text mining tool has been published. Here we introduce the literature-weighted global test to use text mining-derived associations in combination with a regression based analysis of gene expression changes.

The literature-weighted global test

The literature-weighted global test is able to identify biomedical concepts associated with gene expression changes in genome-wide expression studies. The approach integrates previously developed text mining approaches [49, 83, 186] with the global test [9, 21, 169], a statistical framework to evaluate if a set of genes shows significant changes in gene expression.

In our framework, the sources of textual information are abstracts from MEDLINE, a bibliographical database. We use a thesaurus to identify textual references to biomedical concepts in the texts. Concepts have a definition, a list of synonymous terms and can be linked to, for instance, online databases. In the thesaurus, concepts are grouped by semantic categories such as 'gene', 'drug' or 'neoplastic process' and this grouping can be used to select interesting sets of biomedical concepts. After the identification of concepts in texts, we let concepts be represented by the set of documents in which they are mentioned. Subsequently, we use so-called concept profiles to characterize the textual information associated to concepts. A concept profile is a list of concepts with for every concept a weight to indicate its importance.

The vector product of two concept profiles is a measure for the strength of the association between two concepts. Before, association scores between concept profiles have successfully been used to infer functional associations between genes [82, 83] and between genes and GO codes [186], to infer novel genes associated with the nucleolus [51], and to identify new uses for drugs and other substances in the treatment of diseases [187]. We have developed Anni [64] to provide a versatile and user-friendly tool to work with concept profiles (www.biosemantics.org/anni). The tool can be used for a wide variety of queries, such as finding functional associations between genes or retrieving all the genes associated to a disease, and can also serve as a literature-based knowledge discovery tool.

The global test brings a whole framework for statistical testing, including analytical graphs and the ability to test for several types of response variables, such as two state, multi state and continuous variables, as well as survival. We propose to incorporate text-mining derived information in the test, by weighing the participation of genes in the test based on the match of their concept profiles with the concept profile of a biomedical concept (see the Supplementary Data for implementation details).

Figure 1 illustrates how the different resources are used and interact. The input for the global test is a data set of appropriately normalized gene expression measurements and the definition of the experimental variable (e.g. survival time of subjects). The literature-derived association scores used to weight the participation of genes in the test are provided in matrices that can be downloaded from our regularly updated website (http://biosemantics.org/weightedglobaltest). The literature-weighted global test calculates a test statistic and a *p*-value for every biomedical concept tested. Diagnostic plots are available, for instance to study the contribution of individual genes to the test statistic. The literature evidence underlying the inferred associations between a gene and a biomedical concept can be studied with our online tool Anni [64].

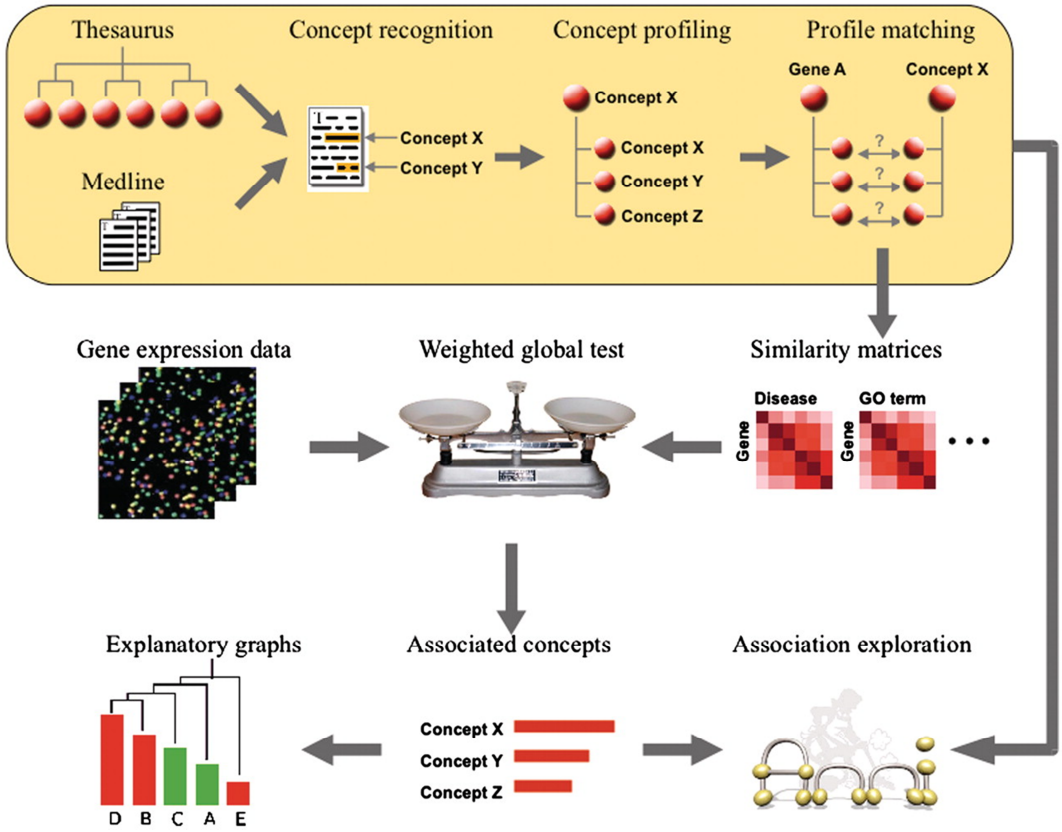

Figure 1. Overview of the literature-weighted global test framework (Description in paragraph 'The literature-weighted global test'). Concepts are characterized by concept profiles, reflecting the literature context in which a concept is mentioned. Association scores reflect the overlap in concept profiles are used to calculate association score between genes and other concepts. These association scores serve as the weights for the literature weighted global test (represented by the balance). The literature-weighted global test calculates a test-static and estimates a *p*-value for every concept evaluated and provides diagnostic plots to study the contribution of individual genes to the test statistic. Literature evidence underlying the inferred associations between a gene and a biomedical concept can be studied with the online tool Anni (**www.biosemantics.org/anni**)

Below we will illustrate the approach, and compare it to a standard GO analysis, by analyzing three data set s in different biomedical domains: (i) cardiac arrhythmia, a biomedical domain that is poorly covered by GO; (ii) breast cancer metastasis, where we demonstrate the use of patient survival data; (iii) peroxisome proliferator alpha function, where gene expression profiles are linked to drug metabolism.

Evaluation

Cardiac development and TBX3

Cardiac development and function are domains poorly covered by GO or other gene annotation databases. The literature-weighted global test can be of assistance in this type of domains as other classes of concepts can be evaluated. Here, we analyze a comparison between normal mouse atrial working myocardium and atria in which the transcription factor TBX3 was ectopically expressed. TBX3 expression is usually confined to the cardiac conduction system and represses properties of the working myocardium, such as fast conduction and contraction, and high level of metabolic activity [188-190]. Ectopic expression causes the spread of the conduction myocardium properties, such as pacemaker activity, slow conduction and contraction and reduced metabolic activity. In these hearts frequent spontaneous contractions and arrhythmias are observed.

We tested the semantic category 'pathologic function' for association with gene expression changes (Table 2). Fifteen of the twenty-five most significant pathological functions were found to be associated with disturbances of cardiac conduction or ion channel activity, in line with the observed phenotypes of the hearts of the mutant mice.

Table 2. The top 25 most significant concepts retrieved by the literature weighted global test for the category 'Pathological Function' on the TBX3 data set

Rank	Concept		*p*-value
1	Chronic dilatation	*	1.91E-05
2	Glycogen depletion		2.07E-05
3	Neonatal bradycardia	*	3.10E-05
4	Global developmental delay		3.72E-05
5	Electrolyte imbalance	*	4.45E-05
6	Fatty infiltration		4.64E-05
7	Abnormal cardiac conduction	*	5.01E-05
8	Cardiac arrhythmia	*	5.14E-05
9	Ventricular Couplet	*	5.75E-05
10	Ventricular arrhythmia	*	5.85E-05
11	Sinus bradycardia	*	6.29E-05
12	Cardiac Arrest	*	6.43E-05
13	Neonatal hypoxia		6.52E-05
14	Upper motor neurone lesion		6.97E-05
15	Sudden cardiac death	*	7.03E-05
16	Neurogenic muscular atrophy		7.42E-05
17	Tetanic uterine contractions		7.53E-05
18	Sudden death	*	7.79E-05
19	Tachycardia, Ventricular	*	8.08E-05
20	Left coronary artery occlusion		8.33E-05
21	Progressive atrophy		8.35E-05
22	Ventricular fibrillation and flutter	*	8.59E-05
23	Ectopic atrial pacemaker	*	8.66E-05
24	Diffuse atrophy		8.77E-05
25	Premature ventricular contraction	*	8.95E-05

Concepts marked with an asterisk are associated with disturbances of cardiac conduction or ion channel activity. The *p*-values have been corrected for multiple testing according to Holm's method.

GO-terms can be represented by concept profiles. This enables a direct comparison of the literature-weighted global test and the standard global test based on the GO consortium gene sets, shown in Table 3 for the GO branch Biological Processes (see Supplementary Tables S1 and S2 for other branches). The literature-weighted global test retrieves more significant GO terms related to cardiac development and cardiac conduction than standard GO analysis (7 versus 2 concepts in the top 25 most significant concepts; Table 3). A review of the top 100 concepts shows that the literature-weighted global test retrieves a higher fraction of relevant terms than the standard GO analysis (0.35 versus 0.26, respectively; Supplementary Table S3). The top concept from the literature-weighted global test is 'action potential propagation' and reflects the observed action propagation defect and spontaneous contractions in the mutant mice. To gain insight into the workings of the test, the R-package includes several graphs. For example, Figure 2 shows the genes contributing most to the finding of this concept, among them the gap junction proteins Gja1 and Scn5a, which are under direct control of TBX3 [189, 191-193], as well as several differentially expressed sodium and potassium channels. Using the standard GO annotations, action potential propagation is only connected to five genes on the microarray. The standard global test only retrieves the generic heart related terms

'heart process' and 'heart contraction', which were represented by an identical set of 52 genes.

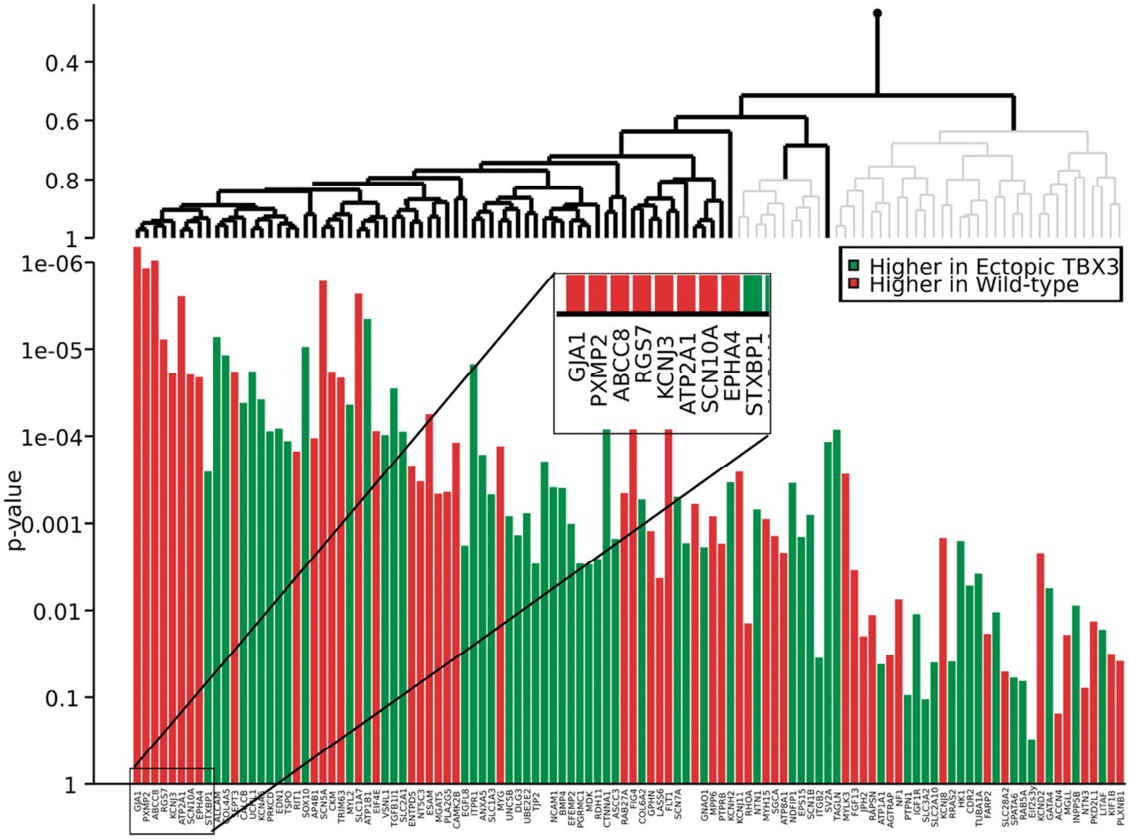

Figure 2. Features plot of differential gene expression between TBX3 overexpressing and wild-type mice based on the importance weights for the concept 'Action Potential Propagation'. The bottom panel gives unadjusted *p*-values for differential expression of each selected gene, colored for the direction of association (green = upregulated in TBX3, red = downregulated in TBX3). The top panel gives a hierarchical clustering (average linkage) of the genes based on absolute correlation distance between their expression values. Imposed on this graph are the results of the inheritance multiple testing procedure of J.J. Goeman and L. Finos (submitted for publication), which is based on the same clustering graph and on the importance weights for the concept 'Action Potential Propagation'. Significant branches and leaves (alpha = 0.05) are shown in black, non-significant branches in gray. To facilitate presentation, all branches that are completely non-significant have been pruned from this picture, removing 1953 out of 2024 genes that have low importance weight and/or low differential expression.

Table 3. The top 25 most significant Biological Processes GO concepts for the standard and literature-weighted global test on the TBX3 data set

	Standard global test		**Literature-weighted global test**	
Rank	GO-concept	*p*-value	GO-concept	*p*-value
1	Apoptotic program	2.69E-05	Action potential propagation *	2.00E-05
2	Cellular component disassembly	5.84E-05	Seed development	4.41E-05
3	Monovalent inorganic cation transport	5.98E-05	Xylanase regulator	4.65E-05
4	Cellular macromolecule catabolic process	6.44E-05	Root morphogenesis	5.54E-05
5	Macromolecule catabolic process	7.08E-05	Aspartate metabolism	5.62E-05
6	Regulation of catabolic process	8.06E-05	Activ. of prog. cell death	6.19E-05
7	Nucleus organization	8.33E-05	Lipoprotein toxin	7.02E-05
8	Biopolymer catabolic process	9.80E-05	Reg. of programmed cell death	7.47E-05
9	Energy deriv. by organic comp. oxidation	9.81E-05	Suppression of hr	7.81E-05
10	DNA catabolic process	1.02E-04	Potassium conductance *	7.92E-05
11	Catabolic process	1.13E-04	Neural crest cell development	8.01E-05
12	Lipid catabolic process	1.17E-04	Cation transport	8.11E-05
13	Regulation of lipid metabolic process	1.26E-04	Muscle hyperplasia	8.19E-05
14	Heart process *	1.28E-04	l-glutamate transport	9.57E-05
15	Heart contraction *	1.28E-04	Reg. of cardiac contraction *	1.01E-04
16	DNA fragmentation, apoptosis	1.30E-04	Activation of atpase activity	1.04E-04
17	Cell struc. disassembly, apoptosis	1.30E-04	TCE metabolism	1.17E-04
18	Apoptotic nuclear changes	1.31E-04	Retrograde axonal transport	1.29E-04
19	Neg. reg. of multicell. organismal process	1.36E-04	Diaphragm contraction	1.35E-04
20	Purine nucleotide metabolic process	1.36E-04	Membrane hyperpolarization *	1.56E-04
21	Regulation of TGF-? receptor pathway	1.74E-04	Adherens junction assembly *	1.62E-04
22	Cation transport	1.96E-04	Generation of action potential *	1.63E-04
23	Nucleoside triphosph. metabolic process	2.03E-04	Potassium ion conductance	1.74E-04
24	Regulation of cell communication	2.04E-04	GlUcose catabolism	1.82E-04
25	Purine nucleotide biosynthetic process	2.11E-04	Muscle hypertrophy	1.86E-0

The *p*-values have been corrected for multiple testing according to Holm's method. Concepts that are related to cardiac conduction and ion channel activity are indicated with an asterisk.

The literature-weighted global test (Venn diagrams in Figure 3) finds more significant terms at the 5% confidence level than the standard test (3.6, 1.8, 2.9-times more for Biological Process, Molecular Function, Cellular Component). The extra concepts scored as significant by the literature-weighted global test are typically more specific than those found with the standard global test. This is reflected by the number of genes annotated in GO per Biological Process category: a median of 11 genes for those specific for the literature-weighted global test versus a median of 123 for the standard global test.

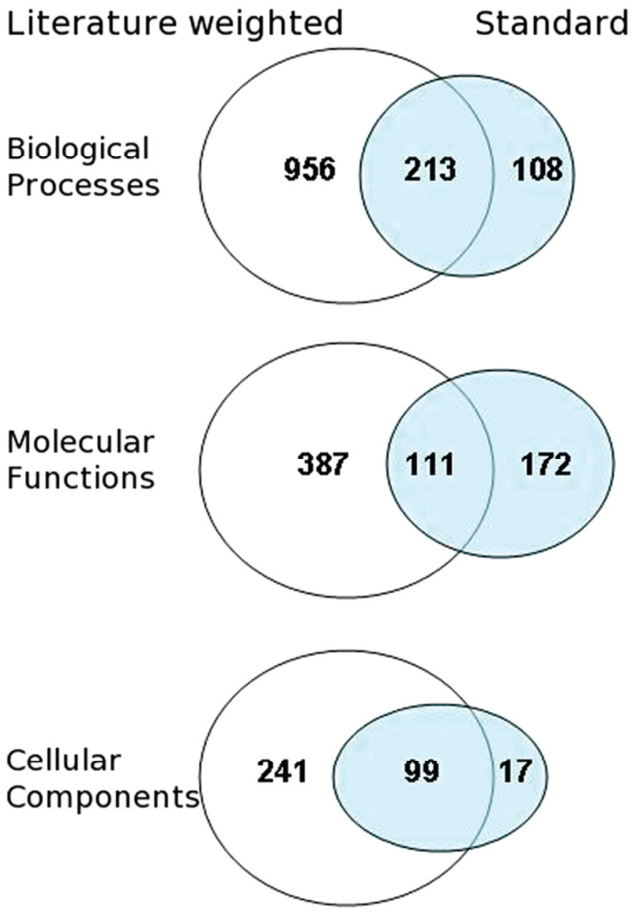

Figure 3. Venn diagrams showing the number of significant GO concepts for both the literature-weighted global test and the standard global test for the TBX3 data set. The three GO branches are Biological processes, Molecular function and Cellular compartment.

The top 25 list of the literature-weighted global test contains some apparently false positive hits. In Table 3, we see 'seed development', 'xylanase regulator' and 'root morphogenesis', which are plant-related concepts. Interestingly, these associations can be traced back to the functions in plants of homologues with the same name and molecular function as the genes differentially expressed in mice. For example, diacylglycerol acetyl transferase 2 (Dgat2) contributes to 'seed development', and is an enzyme with a conserved function in a wide range of organisms, expressed in the heart but also involved in the development of seeds [194]. On one hand, this example illustrates the power of the approach to infer relationships across species. On the other hand, it suggests that further improvements in the removal of concepts irrelevant to the domain under study should be considered.

Analysis of patient survival and gene expression in breast cancer

Van de Vijver et al. [195] investigated the association between breast cancer tissue expression profiles and patient survival and identified prognostic gene signatures. The global test is unique among the gene annotation-based methods, in that it is able to analyze survival time as a response. Biological processes previously associated with survival [196] mainly relate to cell division, and microtubule organization. The genes in these GO categories probably play a role in the development of metastases [197], an important predicting factor for survival. Accordingly, frequently prescribed drugs such as taxanes inhibit cell division and bind microtubules [198, 199]. We analyzed this 'classic' data set with the literature-weighted global test and compared the results to the global test with standard GO term assignments. Similar to the results for the TBX3 data set, we reviewed the top 100 GO terms and found a higher fraction of relevant terms for the

literature-weighted global test compared to the standard GO analysis (0.68 versus 0.48, respectively) (Supplementary Table S4). Table 4 displays the top 25 GO-terms associated with patient survival. Again, the literature-weighted global test produces more specific results than the standard global test, presenting concepts such as 'Telomerase inhibitor activity', 'polycomb group protein complex' and 'thymidine kinase activity' known to be highly relevant for cell division and cancer progression. Although these are genuine GO-categories, they contain less than 10 annotated genes and are therefore unlikely to be found with standard GO-based testing methods. The genes contributing most to the finding of these concepts are TSPYL5, EZH2 and TK1, respectively. EZH2 and TK1 were previously also associated with metastasis indicating that the molecular processes in which they participate are more generally associated with tumor metastasis and survival [200, 201]. TSPYL5 is part of Van de Vijver's 70-gene signature to predict survival in breast cancer, but not much is known about the function of the gene. We associated TSPYL5 with 'metastasis' in Anni and found two papers [202, 203], one of them describing its inhibitory activity on growth of tumor cells and its epigenetic silencing in gastric cancer and gliomas [202]. In addition, TSPYL5 was associated with metastasis through the concepts 'histone deacetylation' and 'telomerase activity'. Evaluations of the Cellular Component and Molecular Function GO branches for this data set s are given in Supplementary Tables S5 and S6.

Table 4. The top 25 most significant Biological Processes GO concepts for the normal and literature-weighted global test for the Van de Vijver data set

	Standard global test			**Literature-weighted global test**		
Rank	GO-concept		*p*-value	GO-concept		*p*-value
1	DNA replication	*	1.28E-05	Chloroplast fission		1.40E-05
2	Protein complex localization		2.05E-05	Meiotic cell cycle regulator	*	1.47E-05
3	Chromosome segregation	*	3.75E-05	Cytokinetic process	*	1.64E-05
4	Protein–DNA complex assembly	*	3.94E-05	Bouquet formation	*	2.08E-05
5	Cytokinesis	*	4.55E-05	Meiotic recombination checkpoint	*	2.19E-05
6	Microtubule cytoskeleton organization	*	5.72E-05	Heterochromatic silencing telomere	*	2.61E-05
7	Establishment of organelle localization		6.85E-05	Stimulation of atpase activity		2.62E-05
8	DNA metabolic process		8.13E-05	Septin ring assembly		2.63E-05
9	Response to DNA damage stimulus	*	8.17E-05	Pyrimidine salvage	*	2.65E-05
10	Mitotic sister chromatid segregation	*	8.50E-05	Sister chromatid cohesion	*	2.97E-05
11	Sister chromatid segregation	*	8.50E-05	Hypusine biosynthesis		3.04E-05
12	Cell cycle	*	8.68E-05	Phosphatidylcholine Biosynthesis		3.05E-05
13	Cell division	*	9.00E-05	Regulation of dna replication	*	3.06E-05
14	Mitotic cell cycle	*	1.19E-04	Deoxycytidine metabolism	*	3.28E-05
15	Cellular macromolecular complex org.		1.47E-04	Telomere clustering	*	3.36E-05
16	M phase	*	1.52E-04	Cell tip growth		3.47E-05
17	Microtubule-based process	*	1.64E-04	Leaf morphogenesis		3.69E-05
18	M phase of mitotic cell cycle	*	1.77E-04	Telomerase inhibitor activity	*	3.71E-05
19	Nuclear division	*	1.77E-04	Heterocycle biosynthesis		3.73E-05
20	Mitosis	*	1.77E-04	Mesendoderm development		3.76E-05
21	Organelle fission		1.77E-04	Activation telomere maintenance	*	3.80E-05
22	Cell cycle phase	*	1.79E-04	TMP biosynthesis		4.02E-05
23	Cell cycle process	*	1.90E-04	Centriole replication	*	4.29E-05
24	Chromosome organization	*	2.07E-04	Double strand break repair	*	4.31E-05
25	Meiosis	*	2.09E-04	Diakinesis		4.42E-05

The *P*-values have been corrected for multiple testing according to Holm's method. Concepts indicated with an asterisk are related to disturbances in cell division and proliferation, or other cancer-related processes.

PPARalpha-mediated effects of dietary lipids on intestinal barrier gene expression

The literature-weighted global test can link gene expression data with (drug) metabolism as our thesaurus contains a large compendium of drugs and other small molecules. We demonstrate this possibility on a data set from de Vogel-van den Bosch *et al***.** [204] that evaluates the effect of a synthetic Peroxisome Proliferator Activated Receptor alpha (PPARalpha) stimulator, the fibrate WY14643, on the gene expression in the small intestine of both wild-type and PPARalpha-null mice. PPARalpha is a nuclear receptor highly expressed in enterocytes, and is thought to play a role in the reaction of the intestine to fatty acids. The PPARA study reported the following manually annotated processes as responsive to PPARalpha stimulation: fatty acid oxidation, cholesterol flux, glucose transport, amino acid metabolism, intestinal motility and oxidative stress. We replicated this manual literature study in an automatic way by using the global test and weighted global test (Supplementary Table S7). With a standard global test on GO biological processes, the most significant processes found were related to fatty acid metabolism. However, GO categories corresponding to the other manual categories were not significant at the 5% level. With the weighted global test we identified significant GO categories corresponding to 5 of the manually grouped categories: 'fatty acid omega oxidation', 'regulation of cholesterol transport', 'glucose transport', 'amino acid metabolism' and 'glutathione metabolism pathway. The GO vocabulary does not contain a concept for intestinal motility. Similar to the previous examples, the literature-weighted global test produces more specific and less redundant results than the standard global test. For example, when the standard global test provided general concepts such as 'lipid metabolic process', 'fatty acid metabolic process, 'carboxylic acid metabolic process' and 'organic acid metabolic process; the literature-weighted global test provided specific concepts such as 'lauric acid metabolism', 'arachidonic acid metabolism' and 'leukotriene metabolism'.

Subsequently, we used the literature-weighted global test to evaluate the concept profiles of drugs. Known PPAR alpha agonists such as clofibrate, gemfibrozil, bezafibrate and fenofibrate were found to be significantly associated with the gene expression differences between PPARalpha-null and wild-type mice (Supplementary Table S8). The most significant drug was benazepril, which is used to treat high blood pressure and inhibits angiotensin-converting enzyme (ACE). Indeed, ACE is differentially expressed and amongst the top contributing genes. The gene with the highest contribution, CYP4A11, showed a large change in gene expression and is directly involved in blood pressure regulation [205]. Also some of the most significant concepts found with the literature-weighted global test on GO biological processes were associated with high blood pressure. We hypothesize that PPAR alpha stimulation also regulates perfusion of the intestine.

Methods

Preparation of the thesaurus and concept tagging

The basis of our analysis is a corpus of 12,648,116 MEDLINE abstracts from January 1, 1980 until July 7, 2009. We used titles, Medical Subject Headings (MeSH) headings, and abstracts. Stop words were removed and words were stemmed to their uninflected form by the lexical variant generator (LVG) normalizer [143]. Subsequently concepts were identified in the texts with the use of a thesaurus. A concept is a defined abstract ideas or entities, and the thesaurus provides a list of terms that are used to refer to the concept in texts. The thesaurus was composed of three parts: the 2008AB version of the Unified Medical Language System (UMLS) [93], a gene thesaurus derived from multiple databases and a chemical thesaurus derived from multiple databases [110]. The gene thesaurus was a combination of gene names from Entrez-Gene, Online Mendelian Inheritance in Man (OMIM) [206], UniProt [207], and the Human Gene Nomenclature Database (HGNC) [208] for *H. sapiens, M. musculus, R. norvegicus, E. coli, S. cerevisiae, D. melanogaster, D. rerio, C. elegans*, and *G. gallus*. Homologous genes between the

species were mapped to each other using NCBI's HomoloGene database [209]. In order to exclude irrelevant concepts, a list of UMLS semantic categories that can confer relevant biological information about genes was made (see [83]) and all concepts of other semantic categories were not further considered. Following Hettne *et al.*[210] the UMLS thesaurus was also adapted for efficient natural language processing, avoiding overly ambiguous or duplicate terms, and terms that are very unlikely to be found in natural text. The gene thesaurus was expanded by rewrite rules to improve recall by taking into account common spelling variations (see [139]). For instance, numbers were replaced with roman numerals and vice versa, and hyphens before numbers at the end of gene symbols were inserted or removed (e.g. "WAF1" was rewritten as "WAF-1" and added as a synonym). After all processing the UMLS part of the thesaurus consisted of 635104 concepts, including 111770 genes and 277300 chemicals.

Concept profiles

References to concepts in texts are identified with our indexing engine Peregrine [139]. To construct a concept profile, a set of Medline records is associated to a concept. For concepts in general this set is composed of the texts in which the concept is mentioned. For genes only a subset of the medline is observed to limit the impact of ambiguous terms and distant homologs. The subset is defined by the Pubmed query: (protein OR gene) AND (mammal OR melanogaster OR gallus OR elegans OR rerio OR cerevisiae OR coli). GO terms are sometimes given as words or phrases that are infrequently found in the normal texts. To still provide broad coverage of GO terms we add the Medline records that were used as evidence for annotating genes with this GO term. For efficiency in the computation of the concept profiles the number of records used for a profile was limited to 10 000. If the set is large the 10 000 most recent records are taken. For every concept in our thesaurus that was associated to at least 5 records, we characterized the texts with a concept profile. A concept profile of a concept *i*, for instance a gene, is an *M*-dimensional vector $w_i=(w_{i1},w_{i2},\ldots,w_{iM})$ where *M* is the number of concepts in the thesaurus. The weight w_{ij} for a concept *j* in this profile indicates the strength of its association to the concept *i*. The weights in a concept profile for concept *i* are derived from the set of texts in which concept *i* occurs, D_i, which is a subset of the total set of texts *D*. To obtain the weight w_{ij} we apply the symmetric uncertainty coefficient $U(X_i,Y_j)$ [211] as suggested and evaluated earlier [63]:

$$W_{ij} = U(X_i, Y_j) = \frac{H(Y_j)+H(X_i)-H(X_i,Y_j)}{\frac{1}{2}(H(X_i)+H(Y_j))} \qquad (1)$$

Here the variable X_i defines whether a text is in D_i, and Y_j gives the occurrence frequency of concept *j*. The Shannon entropies *H* are defined as

$$H(Z) = -\sum_{r=1}^{n} p(z_r)\ln p(z_r)$$

.

Literature-derived association measure

The literature-derived association measure reflects the similarity of the context in which concepts are mentioned in Medline abstracts. Highly related concepts are not necessarily restricted to those occurring in the same abstracts (explicit link), but 2 can also co-occur with common third concepts in Medline abstracts (implicit link). Since genes are concepts themselves, the literature-derived association measure can be expressed as the similarity of the concept profiles of a gene and another concept (GO-term, disease, tissue, drug etc.). The similarity was evaluated by calculating the cosine of the angle (equivalent to the uncentered Pearson correlation) of the two concept profiles. We also evaluated the inner product as an alternative to the cosine and got highly similar results. To avoid taking large numbers of associations with very low scores into account, all cosines < 1e-4 (approximately 80% of all possible associations) were set to zero. This threshold was empirically determined to cause only slight changes in the results compared to the unthresholded test and significantly decrease the computational load. Increasing the

threshold significantly changed the results. The resulting scores were used to weigh the importance of the genes when testing for the association between the observed conditions and a concept using the weighted global test.

The weighted global test

The global test [9] is a method to test for association of the gene expression levels of a set of genes with a response variable. In its original form, the test treats all genes in the set equally, giving each gene equal weight in the test. We replace the global test by a weighted variant that uses the literature-derived association measures defined in the previous section as weights in the test procedure. The mathematical theory required for the construction of the weighted global test is provided by [169]. The global test is defined in terms of a regression model for prediction of the response variable from the gene expression measurements. The type of regression model depends on the response variable: e.g. logistic regression for a two-class response or the Cox proportional hazards model for a survival response. In such a regression model the distribution of the response of subject *i* depends on the gene expression measurements $x_{i1},\ldots,x_{ip}$ through a linear predictor

$$\eta_i = \sum_{j=1}^{p} \chi_{ij\beta_j} \qquad (2)$$

where each gene $j=1,\ldots,p$ in the gene set of p genes has a regression coefficient β_j. The null hypothesis of the global test is

$$H_0: \beta_1 = \ldots = \beta_p = 0, \qquad (3)$$

i.e. the hypothesis that no gene in the gene set is associated with the response. To test this hypothesis even in the situation that the number of genes in the gene set is larger than the sample size *n*, the global test treats the regression coefficients as random effects, assuming that the coefficient vector $\beta = (\beta_1, \ldots, \beta_p)^\tau$ has a distribution with mean zero and covariance matrix $\tau^2\Sigma$, for some pre-specified positive definite $p \times p$ matrix Σ, in previous applications $\Sigma=I$, and an unkown prior variance parameter τ^2. The global test proceeds to test the equivalent null hypothesis H_0: $\tau^2=0$, equivalent to the null hypothesis (3), with a score test. The precise mathematical form of the test statistic of this score test depends on the specific model (Cox, logistic) used, but always involves a quadratic form in the residuals of the fitted model under the null hypothesis. If $\Sigma=I$, this quadratic form involves the matrix $X\Sigma X^T$, where $X=\{x_{ij}\}$ is the n×p matrix of gene expression measurements. If a diagonal Σ is used, the quadratic form involves the matrix $X\Sigma X^T$ instead, and the diagonal elements of Σ function as importance weights for the corresponding genes. The weighting in the resulting test is calibrated in such a way that giving a gene an importance weight of *k* is identical in its effect to duplicating that gene in the gene set *k* times. By its construction, the global test is optimal on average in a neighborhood of the null hypothesis, the shape of which depends on the choice of Σ [169]. Where in its basic form with $\Sigma=I$, the global test weighs all genes in a gene set (e.g. GO category) equally and does not use the expression of genes not contained in the gene set, the global test framework was adapted for the current study to work with gene-specific weights that reflect the strength of the association of a gene with a concept, as defined in the previous paragraph. For each concept, the vector of association scores for each gene was retrieved, and these scores were incorporated as importance weights into the test for that concept using a diagonal matrix Σ with the literature-based score of gene *i* for the tested concept as the *i*th diagonal element. This leads to a weighted test in a very natural way, and the derivation of the resulting test is straightforward [169]. For testing each concept, this test uses all genes that have a non-zero literature-derived score for that concept, but focuses its power on alternatives that

have large regression coefficients for genes with high importance weight with the response. It should be remarked that the null hypotheses of the weighted test is still given by (3), asserting that no gene with a positive importance weight is associated with the response. With a large enough sample size, therefore, a test could return a significant result even when a single gene with a very small importance weight is associated with the response. However, the construction of the test is such that the test has very little power for detecting such an alternative, and much more for detecting an alternative for which many genes with a large importance weight have an association with the response [169]. The ranking of different concepts with different importance weight vectors for the same genes can be expected to reflect these power properties quite accurately. Testing of multiple concepts requires correction for multiple testing. As the tests for different concepts are generally correlated, we suggest using a method of multiple testing control that is valid under all dependency structures, e.g. the method of Holm [212]. Holm's method was used in the applications in this paper. In some cases it may occur that there is no gene with a high, or specific, association score to the tested concept. In this case only weak, potentially spurious, associations will determine the outcome. It is therefore important to observe the used association scores in a test and to verify that the retrieved concepts represent valid associations.

Processing of Datasets

Ectopic expression of TBX3

In this microarray study the gene expression patterns in the heart of two groups of mice were compared (Hoogaars *et al.* manuscript in preparation). One group consists of control mice, the other group ectopically expresses a TBX3 transgene. TBX3 is normally only active in the conduction myocardium. In the transgenic mice, TBX3 is also active in the working myocardium. RNA from left atria of six TBX3 transgenic mice and six control mice (male, 6 weeks) was hybridized to Illumina MouseRef-6 BeadChips. Unprocessed intensity values were averaged per bead type and normalized using VSN in R [213]. Probesets were annotated with EntrezGene IDs using the annotation file provided by Illumina. To summarize the data at the EntrezGene ID level, data from probesets with the same EntrezGeneID were averaged.

Patient survival and gene expression in breast cancer

This dataset [195] contains gene expression profiles of 295 breast cancer tissue samples taken from patients. The orignal aim of the study was to identify genes and pathways important for metastasis and survival. For the current study, we used the normalized dataset filtered for invariant genes [195]. The survival time of the patients was used as the response variable in the global test, and the global test was run with the Cox proportional hazards model for uncensored survival [214].

PPARalpha-mediated effects of dietary lipids on intestinal barrier gene expression

De Vogel-van den Bosch *et al.* [204] used the synthetic PPARalpha agonist WY14643 as a reference when examining the effects of acute nutritional activation of PPARalpha on expression of genes encoding intestinal barrier proteins. Male, 4 months old Wild-type (129S1/SvImJ) and PPARalpha -/- mice (129S4/SvJae) were exposed to dietary fatty acids with WY14643 as a reference during an exposure time of 6 hours. The exposure time of 6 hours was determined with the goal to elucidate the specific, direct contribution of PPARalpha in regulating the expression of transport and phase I/II metabolism genes in the small intestine. Here we use the PPARalpha activation by WY14643. RNA was hybridized on an Affymetrix GeneChip Mouse Genome 430A. We downloaded the raw data from the Gene Expression Omnibus (GSE9533) and normalized the data using RMA normalization [215]. Probesets were annotated with EntrezGene IDs using the bioconductor moe430a.db package. To summarize the data at the EntrezGene ID level, read-outs from probesets with the same EntrezGeneID were averaged.

Concluding remarks

A standard functional analysis of gene expression data involves the testing of gene sets associated with biological processes. Though powerful, these methods mostly rely on manual annotation efforts, which are highly focused and struggle to be complete and up-to-date. Tools based on literature mining can be continuously updated and provide an essentially comprehensive scope. We combined literature mining tools with a thorough statistical framework and show that our approach compares favorably to that of a classic GO analysis, retrieving more, more relevant and more specific GO terms than an analysis based on standard GO annotations.
As expected given the ambiguous nature of gene names, a notorious problem in text mining [216], the automated literature mining comes at the cost of increased number of false positives. It is therefore an important feature of our approach that it is transparent and that results can readily be traced back to the literature that underlies a result. We nevertheless believe that the higher information content of the concepts retrieved by the literature-weighted global test, not only including GO terms but also other types of concepts such as diseases, organ structures and drugs, makes the test more suitable for the interpretation of gene expression data than other available gene set testing algorithms.

Availibility

The weighted global test is now an integral part of the R-package global test and can be obtained from www.bioconductor.org. The matrices containing the literature-derived association scores for genes with biological processes, molecular functions, cellular components, diseases, pathologic functions, tissues and drugs and a file for the mapping between EntrezGene IDs and concept identifiers can be obtained from http://biosemantics.org/weightedglobaltest. Example R-code is available from the same website. Further investigation of the concept profiles and concepts underlying the identified associations can be performed with Anni (www.biosemantics.org/anni) [64].

Acknowledgements

We would like to thank Drs W.M.C. Hoogaars, V.M. Christoffels and G. Hooiveld for useful discussions and critical reading of the article.

Supplementary Data

Supplementary Table S1: Comparison standard vs. literature-weighted global test on the TBX3 dataset of top 25 for the GO category Cellular Component. Highlighted concepts are associated with disturbances of cardiac conduction or ion channel activity.

Standard global test - Cellular Component

GO-category	p-value (Holm)
cell junction	2.34E-07
membrane fraction	3.93E-06
insoluble fraction	5.09E-06
cell fraction	5.95E-06
cell-cell junction	1.50E-05
apical plasma membrane	2.38E-05
extrinsic to membrane	3.56E-05
plasma membrane part	4.14E-05
heterotrimeric G-protein complex	5.49E-05
extrinsic to plasma membrane	5.63E-05
basolateral plasma membrane	8.64E-05
gap junction	1.36E-04
apical part of cell	1.47E-04
envelope	1.63E-04
membrane-bounded vesicle	1.64E-04
organelle envelope	1.66E-04
cytoplasmic membrane-bounded vesicle	1.67E-04
connexon complex	1.70E-04
cytoplasmic part	1.75E-04
plasma membrane	2.21E-04
sarcolemma	2.37E-04
cytoplasm	3.31E-04
cytoplasmic vesicle	3.38E-04
membrane	3.43E-04
vesicle	3.60E-04

Literature-weighted global test - Cellular Component

Concept	p-value (Holm)
photoreceptor connecting cilium	2.17E-06
beta-catenin destruction complex	3.07E-06
axonemal dynein heavy chain	4.62E-06
voltage-gated potassium channel complex	5.83E-06
Radial spoke	1.66E-05
Synaptic cleft	1.72E-05
Fascia adherens	2.99E-05
Neuromuscular Junction	3.67E-05
stereocilium membrane	3.88E-05
Intercalated disc	4.81E-05
Septate desmosome	5.05E-05
Secretory Vesicles	6.09E-05
pigment granule	8.35E-05
Myofibrils	8.40E-05
myosin thick filament	8.85E-05
transverse tubule of muscle cell	9.42E-05
Sarcomeres	1.00E-04
Muscle Fibers	1.05E-04
Gap Junctions	1.06E-04
I band	1.17E-04
Sarcolemma	1.21E-04
Sarcoplasm	1.27E-04
glycogen granule	1.31E-04
hyphal cell wall	1.43E-04
terminal cisterna	1.50E-04

Supplementary Table S2: Comparison standard vs. literature-weighted global test on the TBX3 dataset of top 25 for the GO category Molecular Function. Highlighted concepts are associated with disturbances of cardiac conduction or ion channel activity.

Standard global test - Molecular Function

GO-category	p-value (Holm)
identical protein binding	8.07E-06
aminoacylase activity	2.30E-05
aspartoacylase activity	3.64E-05
voltage-gated ion channel activity	5.00E-05
voltage-gated channel activity	5.00E-05
cation channel activity	5.00E-05
protein homodimerization activity	5.61E-05
channel activity	5.91E-05
passive transmembrane transporter activity	5.91E-05
gated channel activity	6.21E-05
alkali metal ion binding	7.76E-05
ion channel activity	8.20E-05
substrate specific channel activity	9.09E-05
cation transmembrane transporter activity	1.15E-04
metal ion transmembrane transporter activity	1.30E-04
potassium ion binding	1.71E-04
hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, in linear amides	2.04E-04
ion transmembrane transporter activity	2.60E-04
voltage-gated potassium channel activity	2.60E-04
voltage-gated cation channel activity	3.27E-04
transcription activator activity	3.27E-04
substrate-specific transmembrane transporter activity	3.30E-04
potassium channel activity	3.45E-04
solute:cation symporter activity	3.54E-04
transferase activity, transferring acyl groups other than amino-acyl groups	3.64E-04

Literature-weighted global test - Molecular Function

Concept	p-value (Holm)
transmembrane receptor protein tyrosine kinase activity	1.05E-05
myosin binding	2.33E-05
high-density lipoprotein binding	4.53E-05
ion channels	4.79E-05
lipoprotein toxin	4.90E-05
aspartoacylase activity	5.20E-05
transmembrane receptor protein tyrosine kinase activity	5.46E-05
malate oxidase activity	5.75E-05
phosphatidylinositol-4,5-bisphosphate hydrolysis	6.55E-05
channel activity	6.84E-05
carboxypeptidase activity	8.15E-05
alpha-toxin activity	9.77E-05
plasma membrane ca-atpase	1.06E-04
transporter activity	1.12E-04
potassium transporter	1.17E-04
coupled atpase activity	1.25E-04
alpha adrenoceptor activity	1.33E-04
creatine phosphokinase activity	1.50E-04
ion channel activity	1.60E-04
2-keto-4-hydroxyglutarate aldolase activity	1.71E-04
MUSK	1.83E-04
GTPase activity	3.13E-04
intercellular channel activity	3.25E-04
carrier activity	3.29E-04
connexin activity	3.42E-04

Supplementary Table S3: Evaluation of True Positive and False Positive rates for the TBX3 dataset. We have evaluated the top100 concepts in Biological Processes for and classified each hit as "relevant" (True Positive), "cannot decide", or "irrelevant or too unspecific" (False Positive). Precision was defined as TP/(TP+FP).

Standard global test: Top 100 Biological Processes

FDR	Term	Relevant	Cannot decide	Irrelevant or too unspecific
2.59E-05	apoptotic program		*	
5.84E-05	cellular component disassembly			*
5.98E-05	monovalent inorganic cation transport		*	
6.44E-05	cellular macromolecule catabolic process			*
7.08E-05	macromolecule catabolic process			*
8.06E-05	regulation of catabolic process			*
8.33E-05	nucleus organization		*	
9.80E-05	biopolymer catabolic process			*
9.81E-05	energy derivation by oxidation of organic compounds		*	
0.000102389	DNA catabolic process		*	
0.000112822	catabolic process			*
0.00011688	lipid catabolic process		*	
0.000126322	regulation of lipid metabolic process		*	
0.000127868	heart process	*		
0.000127868	heart contraction	*		
0.000130193	DNA fragmentation during apoptosis		*	
0.000130395	cell structure disassembly during apoptosis		*	
0.000131071	apoptotic nuclear changes		*	
0.000135795	negative regulation of multicellular organismal process			*
0.000136171	purine nucleotide metabolic process		*	
0.000173956	regulation of transforming growth factor beta receptor signaling pathway		*	
0.000195997	cation transport		*	
0.00020294	nucleoside triphosphate metabolic process		*	
0.0002043	regulation of cell communication		*	
0.00021056	purine nucleotide biosynthetic process		*	
0.000231734	sodium ion transport	*		
0.000233833	metal ion transport		*	
0.000235091	nucleoside phosphate metabolic process		*	
0.000235091	nucleotide metabolic process		*	
0.000240485	nucleoside triphosphate biosynthetic process		*	
0.000241692	purine nucleoside triphosphate metabolic process		*	
0.0002427	negative regulation of lipid metabolic process		*	
0.00025404	regulation of lipid catabolic process		*	
0.000256569	regulation of signal transduction		*	
0.000268039	nucleobase, nucleoside and nucleotide metabolic process		*	
0.000277131	regulation of heart contraction	*		
0.000289285	purine nucleoside triphosphate biosynthetic process		*	
0.000313885	negative regulation of lipid catabolic process		*	
0.0003145	MAPKKK cascade		*	
0.000337488	nucleotide biosynthetic process		*	
0.000345639	cell maturation		*	
0.000358886	positive regulation of signal transduction		*	
0.000364194	negative regulation of metabolic process		*	
0.00039376	purine ribonucleotide metabolic process		*	
0.000401146	positive regulation of cell communication		*	
0.00040171	negative regulation of catabolic process		*	
0.000413606	glycogen metabolic process		*	
0.000413606	cellular glucan metabolic process		*	
0.000413606	glucan metabolic process		*	
0.000416534	ribonucleoside triphosphate metabolic process		*	
0.000416974	purine ribonucleoside triphosphate metabolic process		*	
0.000418691	purine ribonucleotide biosynthetic process		*	
0.000418762	ATP metabolic process		*	
0.000422239	cellular catabolic process			*
0.000425033	ribonucleoside triphosphate biosynthetic process		*	
0.000425033	purine ribonucleoside triphosphate biosynthetic process		*	
0.000437684	ATP biosynthetic process		*	
0.000445663	myofibril assembly		*	
0.000445663	striated muscle cell development		*	
0.000452871	lipid localization		*	
0.000452871	lipid storage		*	
0.000458461	carbohydrate metabolic process		*	
0.000459704	energy reserve metabolic process		*	
0.000484803	potassium ion transport	*		
0.000486834	intracellular signaling cascade			*
0.000585736	negative regulation of tissue remodeling		*	
0.000619881	regulation of molecular function			*
0.000692142	negative regulation of bone remodeling			*
0.00075131	ion transport	*		
0.000773594	cellular amino acid and derivative metabolic process		*	
0.000797961	actomyosin structure organization		*	
0.000798572	ribonucleotide metabolic process		*	
0.000803118	negative regulation of ossification			*
0.000812495	developmental maturation		*	
0.000820615	negative regulation of transforming growth factor beta receptor signaling		*	
0.000828736	striated muscle cell differentiation		*	
0.00083741	negative regulation of macromolecule biosynthetic process		*	
0.000848679	glycogen biosynthetic process		*	
0.000848679	glucan biosynthetic process		*	
0.000851448	macromolecule localization			*
0.000862435	muscle cell development		*	
0.000907107	ribonucleotide biosynthetic process		*	
0.000914239	temperature homeostasis			*
0.000927143	cardiac myofibril assembly		*	
0.000964815	circulatory system process		*	
0.000964815	blood circulation		*	
0.00097011	organic acid metabolic process		*	
0.000978724	ubiquitin-dependent protein catabolic process		*	
0.00099378	modification-dependent protein catabolic process		*	
0.00099378	modification-dependent macromolecule catabolic process		*	
0.001005404	carboxylic acid metabolic process		*	
0.001030953	heart looping		*	
0.001050168	regulation of catalytic activity		*	
0.001118148	amino acid derivative metabolic process		*	
0.00118922	cellular biopolymer catabolic process		*	
0.00118922	cellular protein catabolic process		*	
0.001191852	proteolysis involved in cellular protein catabolic process		*	
0.001271085	hair follicle development			*
0.001271085	molting cycle process			*
0.001271085	hair cycle process			*

Category	Counts
Relevant	6
Cannot decide	77
Completely irrelevant or too unspecific	17
Precision	0.26087

Supplementary Table S3 (continued)

Literature-weighted global test: Top 100 Biological Processes

Holm	alias	Relevant	Cannot decide	Irrelevant or too unspecific
2.00E-05	action potential propagation	*		
4.41E-05	seed development			*
4.65E-05	xylanase regulator			*
5.54E-05	root morphogenesis			*
5.62E-05	aspartate metabolism		*	
6.19E-05	activation of programmed cell death		*	
7.02E-05	lipoprotein toxin		*	
7.47E-05	regulation of programmed cell death		*	
7.81E-05	suppression of hr			*
7.92E-05	potassium conductance	*		
8.01E-05	neural crest cell development		*	
8.11E-05	cation transport		*	
8.19E-05	muscle hyperplasia		*	
9.57E-05	l-glutamate transport		*	
0.000101247	regulation of cardiac contraction	*		
0.000104353	activation of atpase activity		*	
0.000116799	TCE metabolism		*	
0.000129173	retrograde axonal transport		*	
0.000134822	diaphragm contraction			*
0.000155574	membrane hyperpolarization	*		
0.000162067	adherens junction assembly		*	
0.000162513	generation of action potential	*		
0.000174251	potassium ion conductance	*		
0.000182159	glucose catabolism		*	
0.000186443	Muscle hypertrophy		*	
0.000188041	regulation of synaptic transmission		*	
0.000188603	cellular homeostasis			*
0.000190431	acidic amino acid transport		*	
0.000201504	glutamic acid biosynthesis		*	
0.000205595	glycogen catabolism		*	
0.000206213	auxin homeostasis			*
0.000212417	interferon-gamma biosynthesis		*	
0.000215876	mixed acid fermentation			*
0.000221733	glutamate reuptake		*	
0.000223355	skeletal morphogenesis			*
0.000228344	sensory transduction			*
0.000230734	regulation of adipocyte differentiation			*
0.000244387	anaerobic glycolysis		*	
0.000247234	aspartate biosynthesis		*	
0.000251486	potassium ion transport	*		
0.000261529	membrane depolarization	*		
0.000263536	inhibition of tumor necrosis factor production		*	
0.000274445	mesenchymal cell differentiation		*	
0.00027756	inhibition of action potential	*		
0.000280817	neural crest formation		*	
0.000280963	stimulation of pigmentation			*
0.000283896	muscle plasticity		*	
0.000284306	carbohydrate biosynthetic process		*	
0.000285175	stimulation of smooth muscle contraction		*	
0.000286197	system process			*
0.000287606	Protein Biosynthesis			*
0.000292041	uterus development			*
0.000297614	glutamate secretion		*	
0.000314413	fin regeneration			*
0.000316549	ATP regeneration		*	
0.000317232	Myocardial Contraction	*		
0.000320751	muscle adaptation		*	
0.000322126	regulation of protein catabolism		*	
0.000332165	alanine biosynthesis		*	
0.000341639	dorsal closure			*
0.00034634	antimicrobial peptide production			*
0.000352508	regulation of insulin secretion		*	
0.000354725	NADH metabolism		*	
0.000368929	endothelial cell differentiation		*	
0.000377304	l-aspartate transport		*	
0.000390507	stimulation of appetite			*
0.000409561	induction of programmed cell death		*	
0.000416336	sertoli cell proliferation			*
0.000421768	aspartate transport		*	
0.000439735	amino acid metabolism		*	
0.000455327	Intercellular Communication		*	
0.000470039	cell volume homeostasis		*	
0.000472791	inhibition of fatty acid beta-oxidation		*	
0.000486666	mRNA stabilization		*	
0.000513673	gonad development			*
0.000518471	hydrogen transport		*	
0.00052069	glutamate recycling		*	
0.00052156	Translation, Genetic		*	
0.000529036	heart morphogenesis	*		
0.000535767	stimulation of gtpase activity		*	
0.000539228	microtubule-based process		*	
0.000548692	down regulation of apoptosis		*	
0.000557397	smooth muscle atrophy		*	
0.000559856	Cardiac Chronotropy	*		
0.000560129	directional locomotion			*
0.000560511	Muscle Contraction	*		
0.000564765	negative regulation of translation		*	
0.000566133	membrane lipid breakdown		*	
0.000571453	regulation of action potential	*		
0.000576658	skeletal muscle hypertrophy		*	
0.000581913	activation of cell adhesion		*	
0.000587818	spinal cord development			*
0.000593018	inhibition of insulin secretion		*	
0.000600671	regulation of guanylate cyclase activity		*	
0.000600724	smooth muscle adaptation		*	
0.000607332	regulation of adenylate cyclase activity		*	
0.000617621	response to hypoxia		*	
0.000624919	antennal development			*
0.000638862	glutamatergic synaptic transmission		*	
0.000642544	response to freezing			*

Category	Counts
Relevant	14
Cannot decide	61
Completely irrelevant	25
Precision	0.358974

Supplementary Table S4: Evaluation of True Positive and False Positive rates for the Van de Vijver dataset. We have evaluated the top100 concepts in Biological Processes for and classified each hit as "relevant" (True Positive), "cannot decide", or "irrelevant or too unspecific" (False Positive). Precision was defined as TP/(TP+FP).

Standard global test: Top 100 Biological Processes

Pvalue	Gene Set	relevant	cannot decide	irrelevant or too unspecific
1.28E-05	DNA replication	*		
2.05E-05	protein complex localization			*
3.75E-05	chromosome segregation	*		
3.94E-05	protein-DNA complex assembly	*		
4.55E-05	cytokinesis	*		
5.72E-05	microtubule cytoskeleton organization	*		
6.85E-05	establishment of organelle localization			*
8.13E-05	DNA metabolic process		*	
8.17E-05	response to DNA damage stimulus	*		
8.50E-05	mitotic sister chromatid segregation	*		
8.50E-05	sister chromatid segregation	*		
8.68E-05	cell cycle	*		
9.00E-05	cell division	*		
0.000118958	mitotic cell cycle	*		
0.000147144	cellular macromolecular complex subunit organization		*	
0.000151745	M phase	*		
0.000164247	microtubule-based process	*		
0.000177427	M phase of mitotic cell cycle	*		
0.000177427	nuclear division	*		
0.000177427	mitosis	*		
0.000177427	organelle fission		*	
0.000179208	cell cycle phase	*		
0.000189937	cell cycle process	*		
0.000206901	chromosome organization	*		
0.000208806	meiosis			*
0.000208806	meiotic cell cycle			*
0.000208806	M phase of meiotic cell cycle			*
0.000209821	regulation of DNA metabolic process		*	
0.000250365	nucleosome assembly			*
0.000251791	modification-dependent protein catabolic process			*
0.000251791	modification-dependent macromolecule catabolic process			*
0.000251791	proteolysis involved in cellular protein catabolic process			*
0.0002524	spindle organization	*		
0.000252489	protein catabolic process			*
0.000266266	ubiquitin-dependent protein catabolic process		*	
0.000277797	cellular response to stress		*	
0.000302082	DNA-dependent DNA replication	*		
0.000312571	cellular biopolymer catabolic process			*
0.000312571	cellular protein catabolic process			*
0.000314051	organelle organization			*
0.000335328	cellular response to stimulus			*
0.00033654	cellular biopolymer biosynthetic process			*
0.00033654	biopolymer biosynthetic process			*
0.000367383	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	*		
0.000367839	cellular macromolecular complex assembly			*
0.000390641	DNA packaging	*		
0.000409651	dephosphorylation			*
0.000413569	protein amino acid dephosphorylation			*
0.000426948	DNA recombination	*		
0.000428794	macromolecule catabolic process			*
0.000429878	cellular response to DNA damage stimulus	*		
0.000460831	cellular macromolecule catabolic process			*
0.000484719	cellular macromolecule metabolic process			*
0.000497446	cellular biopolymer metabolic process			*
0.000502957	primary metabolic process			*
0.000510297	biopolymer metabolic process			*
0.000521725	biopolymer catabolic process			*
0.00053611	cellular macromolecule biosynthetic process			*
0.000541834	tRNA processing		*	
0.000557438	DNA repair	*		
0.000631164	macromolecule biosynthetic process			*
0.000638285	cellular metabolic process			*
0.000643606	macromolecule metabolic process			*
0.000650257	cellular component organization			*
0.00065568	macromolecular complex subunit organization			*
0.000692862	translation			*
0.000724103	biosynthetic process			*
0.000725043	mitotic chromosome condensation	*		
0.000772283	metabolic process			*
0.00078229	cellular biosynthetic process			*
0.0007976	nucleosome organization	*		
0.000805513	cytoskeleton organization	*		
0.000809126	cellular catabolic process			*
0.000840188	RNA processing			*
0.000867074	gene expression			*
0.000927096	protein metabolic process			*
0.000928606	cellular protein metabolic process			*
0.000940732	protein ubiquitination		*	
0.000940732	protein modification by small protein conjugation		*	
0.000946491	T cell homeostasis		*	
0.000970614	cyclin catabolic process	*		
0.000970614	positive regulation of exit from mitosis	*		
0.000971347	RNA modification			*
0.001015191	female pregnancy			*
0.00108299	chromatin assembly or disassembly	*		
0.001127238	DNA unwinding during replication	*		
0.001127238	DNA geometric change	*		
0.001127238	DNA duplex unwinding	*		
0.00113852	chromatin assembly	*		
0.001149994	regulation of DNA replication	*		
0.001163612	cellular aldehyde metabolic process		*	
0.001177291	regulation of cell cycle	*		
0.00118864	positive regulation of mitotic metaphase/anaphase transitio	*		
0.00120459	RNA metabolic process		*	
0.001283676	regulation of metabolic process		*	
0.001390441	regulation of exit from mitosis	*		
0.001390441	exit from mitosis	*		
0.001405485	proteolysis		*	
0.001415773	deoxyribonucleotide biosynthetic process		*	
0.001417207	regulation of macromolecule metabolic process		*	

Category	Counts
relevant	40
cannot decide	17
irrelevant or too unspecific	43
precision	0.481928

Supplementary Table S4 (continued)

Literature-weighted global test: Top 100 Biological Processes

Holm	alias	relevant	cannot decide	irrelevant or too unspecific
1.40E-05	chloroplast fission			*
1.47E-05	meiotic cell cycle regulator			*
1.64E-05	cytokinetic process	*		
2.08E-05	bouquet formation			*
2.19E-05	meiotic recombination checkpoint			*
2.61E-05	heterochromatic silencing at telomere	*		
2.62E-05	stimulation of atpase activity		*	
2.63E-05	septin ring assembly		*	
2.65E-05	pyrimidine salvage	*		
2.97E-05	sister chromatid cohesion	*		
3.04E-05	hypusine biosynthesis		*	
3.05E-05	Phosphatidylcholine Biosynthesis		*	
3.06E-05	regulation of dna replication	*		
3.28E-05	deoxycytidine metabolism		*	
3.36E-05	telomere clustering	*		
3.47E-05	cell tip growth			*
3.69E-05	leaf morphogenesis			*
3.71E-05	telomerase inhibitor activity	*		
3.73E-05	heterocycle biosynthesis		*	
3.76E-05	mesendoderm development		*	
3.80E-05	activation of telomere maintenance	*		
4.02E-05	TMP biosynthesis		*	
4.29E-05	centriole replication	*		
4.31E-05	double strand break repair	*		
4.42E-05	Diakinesis			*
4.47E-05	activation of meiosis			*
4.47E-05	contractile ring formation cytokinesis	*		
4.57E-05	establishment of sister chromatid cohesion	*		
4.67E-05	upregulation of meiosis			*
5.11E-05	DNA underwinding	*		
5.28E-05	ribosome assembly		*	
5.36E-05	centrosome localization	*		
5.42E-05	regulation of chromosome segregation	*		
5.49E-05	inhibition of nuclear division	*		
5.54E-05	cytokinesis checkpoint	*		
5.56E-05	Telomere Maintenance	*		
5.71E-05	regulation of cytokinesis	*		
5.82E-05	ascus development			*
5.88E-05	DNA Ligation	*		
6.05E-05	nuclease inhibitor		*	
6.08E-05	Telomere Capping	*		
6.20E-05	TTP biosynthesis	*		
6.28E-05	guanosine salvage	*		
6.37E-05	Nucleotide Biosynthesis	*		
6.99E-05	microsporogenesis			*
7.14E-05	spindle disassembly	*		
7.29E-05	propionate biosynthesis		*	
7.34E-05	synaptonemal complex assembly			*
7.34E-05	Chromosome Pairing	*		
7.58E-05	chiasma formation	*		
7.78E-05	plasmid maintenance			*
8.25E-05	mitotic sister chromatid separation	*		
8.33E-05	ribosome biogenesis and assembly		*	
8.44E-05	meiotic prophase I			*
8.63E-05	replication fork arrest	*		
8.80E-05	inosine biosynthesis		*	
8.83E-05	meiosis I			*
8.89E-05	mitotic chromosome condensation	*		
9.00E-05	pyrimidine metabolism pathway	*		
9.16E-05	epidermal cell division		*	
9.29E-05	postreplication repair	*		
9.43E-05	replication priming	*		
9.50E-05	Microtubule Bundle	*		
9.77E-05	DNA metabolism	*		
9.79E-05	rotokinesis	*		
9.99E-05	asymmetric protein localization		*	
0.00010034	regulation of mitosis	*		
0.00010086	replication fork maintenance	*		
0.00010123	pyrimidine base catabolism	*		
0.00010139	chromosome transmission	*		
0.00010429	maintenance of genome integrity	*		
0.00010442	tRNA splicing		*	
0.0001068	response to oleate		*	
0.0001077	Nucleotide Metabolism		*	
0.00010968	meiotic chromosome segregation			*
0.00011153	response to hydrostatic pressure			*
0.00011313	leptotene			*
0.00011447	mitotic spindle elongation	*		
0.00011533	meiosis II			*
0.00011565	regulation of photosynthesis			*
0.00011581	Cytokinesis	*		
0.00011881	Gluconeogenesis Inhibition		*	
0.0001195	diplotene			*
0.00011972	post-embryonic development		*	
0.00012394	regulation of meiosis			*
0.00012498	centrosome cycle	*		
0.00012605	inhibition of cell division	*		
0.00012626	zygotene			*
0.00012659	endomitotic cell cycle	*		
0.00012816	mitotic spindle formation	*		
0.00013039	chromate transport		*	
0.00013051	microtubule nucleation	*		
0.00013323	Meiosis			*
0.00013517	DNA unwinding	*		
0.00013611	chromosome separation	*		
0.00013699	nucleoside biosynthesis		*	
0.00013703	mitotic sister chromatid segregation	*		
0.00013717	nuclear division	*		
0.00013757	spindle orientation checkpoint	*		
0.00013768	budding			*

Category	Counts
relevant	53
cannot decide	22
irrelevant or too unspecific	25
precision	0.679487

SupplementaryTable S5: Comparison standard vs. literature-weighted global test on the Van de Vijver dataset of top 25 for the GO category Cellular Component. Highlighted concepts are related to disturbances in cell division and proliferation, or other cancer-related processes.

Standard global test - Cellular Component

GO-category	p-value(Holm)
chromosome	6.01E-06
nuclear lumen	6.61E-06
nuclear chromosome	7.70E-06
nucleoplasm	8.74E-06
nuclear part	9.07E-06
chaperonin-containing T-complex	9.78E-06
intracellular organelle lumen	9.91E-06
organelle lumen	1.33E-05
chromosomal part	1.48E-05
membrane-enclosed lumen	1.49E-05
non-membrane-bounded organelle	1.59E-05
intracellular non-membrane-bounded organelle	1.59E-05
condensed chromosome	1.98E-05
microtubule cytoskeleton	2.21E-05
nucleus	2.32E-05
microtubule	2.50E-05
nuclear periphery	2.80E-05
chromosome, centromeric region	2.89E-05
centrosome	3.30E-05
kinetochore	3.77E-05
cytosol	3.97E-05
nuclear chromosome part	3.97E-05
condensed chromosome, centromeric region	4.44E-05
condensed chromosome kinetochore	4.86E-05
microtubule organizing center	4.87E-05

Literature-weighted global test - Cellular Component

Concept	p-value (Holm)
septin ring	3.49E-06
SMC complex	6.84E-06
pol-prim	7.13E-06
Mesosome	8.06E-06
interphase microtubule organizing center	8.41E-06
stromule	8.52E-06
Chiasma	8.60E-06
mismatch repair complex	8.61E-06
replisome	9.20E-06
processivity clamp	9.42E-06
primosome	1.05E-05
cohesin complex	1.07E-05
Polysomal ribosome	1.10E-05
Nuclear chromosome	1.15E-05
polycomb group protein complex	1.18E-05
Interphase chromosome	1.30E-05
pre-IC	1.34E-05
extrachromosomal DNA	1.38E-05
resolvesome	1.41E-05
chromosome scaffold	1.50E-05
Kinetoplast	1.55E-05
monopolin complex	1.57E-05
inner centromere core complex	1.57E-05
ribosome subunits, large, eukaryotic	1.62E-05
Chromatids	1.82E-05

Supplementary Table S6: Comparison standard vs. literature-weighted global test on the Van de Vijver dataset of top 25 for the GO category Molecular Function. Highlighted concepts are related to disturbances in celldivision and proliferation, or other cancer-related processes.

Standard global test - Molecular Function

GO-category	p-value(Holm)
chaperone binding	6.39E-06
adenyl ribonucleotide binding	2.79E-05
ATP binding	2.79E-05
acid-amino acid ligase activity	3.32E-05
helicase activity	3.58E-05
ATPase activity	3.92E-05
endonuclease activity	4.06E-05
adenyl nucleotide binding	4.33E-05
small conjugating protein ligase activity	4.49E-05
protein tyrosine phosphatase activity	5.37E-05
phosphoprotein phosphatase activity	6.38E-05
ATPase activity, coupled	6.44E-05
nucleoside kinase activity	7.38E-05
endoribonuclease activity	7.53E-05
nucleoside-triphosphatase activity	9.46E-05
phosphatase activity	9.77E-05
ribonucleotide binding	9.79E-05
purine ribonucleotide binding	9.79E-05
endoribonuclease activity, producing 5'-phosphomonoesters	1.05E-04
endonuclease activity, active with either ribo- or deoxyribonucleic acids and producing 5'-phosphomonoesters	1.05E-04
DNA helicase activity	1.07E-04
ligase activity	1.10E-04
ubiquitin-protein ligase activity	1.11E-04
ligase activity, forming carbon-nitrogen bonds	1.12E-04
pyrophosphatase activity	1.29E-04

Literature-weighted global test - Molecular Function

Concept	p-value (Holm)
deoxycytidine kinase activity	2.45E-06
nucleoside binding	5.04E-06
translation regulator activity	7.82E-06
cellulose binding	1.09E-05
TMPK activity	1.14E-05
deoxyguanosine kinase activity	1.14E-05
dTTPase activity	1.43E-05
thymidylate synthase activity	1.56E-05
photoreactivating enzyme activity	2.07E-05
telomerase inhibitor activity	2.60E-05
mispair binding	2.63E-05
processivity clamp	2.95E-05
dual-specificity protein phosphatase	2.95E-05
phosphothreonine binding	3.22E-05
telomere binding	3.28E-05
phosphocholine cytidylyltransferase activity	3.41E-05
thymidine kinase activity	3.49E-05
neuropeptide receptor activity	3.62E-05
myelin basic protein kinase activity	3.98E-05
[phosphotyrosine]protein phosphatase activity	4.10E-05
deoxyribonuclease activity	4.27E-05
ARS binding	4.32E-05
nucleoside phosphotransferase activity	5.10E-05
deoxyribonucleic acid repair enzyme	5.14E-05
[phosphorylase] phosphatase activity	5.57E-05

Supplementary Table S7: Comparison standard vs. literature-weighted global test on the PPARalpha dataset of top 25 for the GO category Biological processes.

Standard global test - Biological Processes

GO category	p-value (Holm)
response to external stimulus	1.60E-03
hormone metabolic process	1.68E-03
regulation of lipid metabolic process	1.74E-03
regulation of hormone levels	2.69E-03
membrane lipid biosynthetic process	2.87E-03
oxidation reduction	3.07E-03
lipid biosynthetic process	3.31E-03
fatty acid metabolic process	3.61E-03
lipid metabolic process	4.42E-03
carboxylic acid metabolic process	4.58E-03
organic acid metabolic process	4.61E-03
monocarboxylic acid metabolic process	4.92E-03
cellular lipid metabolic process	5.01E-03
cellular carbohydrate metabolic process	5.28E-03
response to wounding	6.79E-03
positive regulation of immune system process	8.06E-03
organic acid biosynthetic process	9.45E-03
carboxylic acid biosynthetic process	9.45E-03
fatty acid biosynthetic process	9.65E-03
regulation of myeloid leukocyte differentiation	1.00E-02
hexose metabolic process	1.05E-02
regulation of immune system process	1.09E-02
long-chain fatty acid transport	1.11E-02
monosaccharide metabolic process	1.19E-02
regulation of fatty acid metabolic process	1.21E-02

Literature-weighted global test - Biological Processes

GO category	p-value (Holm)
blood pressure homeostasis	2.33E-05
blood pressure regulation	2.92E-05
tubuloglomerular feedback	3.13E-05
alkane biosynthesis	3.62E-05
regulation of natriuresis	4.17E-05
pressure natriuresis	4.34E-05
tissue death	5.23E-05
adenosine 3'-phosphate 5'-phosphosulfate transport	5.92E-05
lauric acid metabolism	5.96E-05
hepoxilin biosynthesis	6.13E-05
alkane breakdown	6.70E-05
Heterotrophy	6.76E-05
nicotine metabolism	7.64E-05
arachidonic acid metabolism	7.90E-05
acrylonitrile metabolism	8.62E-05
yolk production	9.12E-05
leukotriene metabolism	9.34E-05
icosanoid biosynthetic process	9.87E-05
cyclooxygenase pathway	1.09E-04
TCE metabolism	1.11E-04
cyanide biosynthesis	1.15E-04
phenanthrene metabolism	1.15E-04
lipoxygenase pathway	1.18E-04
dichloromethane metabolism	1.28E-04
taxol metabolism	1.32E-04

Supplementary Table S8: The *p*-value and rank for known PPAR alpha activators after evaluation of the semantic category drugs.

PPAR agonist	p-value (Holm)	Rank
clofibrate	2.63E-04	210
gemfibrozil	9.79E-04	333
bezafibrate	1.14E-03	352
fenofibrate	1.41E-03	378

Chapter 6

Next-generation text-mining mediated chemical-response specific gene sets for interpretation of gene expression data

Kristina M. Hettne[*,†,‡], André Boorsma[§], Dorien A. M. van Dartel[¶,#], Jelle J.Goeman[||], Esther de Jong[¶,|||], Aldert H. Piersma[¶,|||], Rob H. Stierum[§], Jos C. Kleinjans[*], Jan A. Kors[†]

[*] Department of Toxicogenomics, Maastricht University, Maastricht, The Netherlands
[†] Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, the Netherlands
[‡] Department of Human Genetics, Leiden University Medical Center, Leiden, the Netherlands
[§] Microbiology and Systems Biology, TNO, Zeist, The Netherlands
[¶] Laboratory for Health Protection Research, National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands
[#] Human and Animal Physiology, Wageningen University, Wageningen, The Netherlands
[||] Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, the Netherlands
[|||] Institute for Risk Assessment Sciences, Utrecht University, Utrecht, The Netherlands

Submitted.

Abstract

Background
Availability of chemical response-specific gene sets for pharmacological and/or toxic effect prediction for compounds is limited. The biomedical literature contains a lot of information about small molecules that is not currently stored in gene expression databases, and this information could be used to build chemical response-specific signatures.

Methods
Using text-mining (TM), we were able to create 30,211 chemical response-specific gene sets. We compared TM-derived gene sets against gene sets derived from the curated Comparative Toxicogenomics Database (CTD), using three different gene set analysis tools in three case studies (identification of chemical treatment, pharmacological mechanism elucidation, and compound toxicity profile comparison).

Results
First, we tested the differential expression of the TM-derived gene sets and the CTD-derived gene sets in data sets of five chemicals (from experimental models). The TM-derived gene sets matching the chemical treatment were significantly altered (false discovery rate -corrected p-values < 0.05) in three data sets and in five data sets the CTD-derived gene sets matching the treatments were significantly altered. Second, we tested the differential expression of gene sets for six fibrates in a peroxisome proliferator-activated receptor alpha knock-out dataset. Six TM-derived fibrate gene sets were significantly altered and four CTD-derived fibrate gene sets were significantly altered. Finally, we tested the differential expression of 319 TM-derived gene sets for environmental toxins in three gene expression data sets of triazoles: 33 gene sets were significantly altered. Twenty-one of these toxins had a similar toxicity pattern as the triazoles. Using TM-derived gene sets for embryonic structures, we confirmed embryotoxic effects seen in the whole embryo culture, and discriminated triazoles from other chemicals in a principal component analysis.

Conclusions
Gene set analysis with TM-derived chemical response-specific gene sets is a scalable method for identifying similarities in gene responses to other chemicals, from which one may infer potential mode of action and/or toxic effect.

Background

Connectivity mapping is a promising method, based on gene-expression similarity, that can be applied to toxicogenomic data to understand mechanisms of toxicity (for a recent review see [217]). In essence, toxicological properties of a chemical entity are discerned by connecting a query gene signature generated as a result of exposure of a biological system (whole animal, tissue or cell line) to the chemical, to other, already known, chemical entities via a database of chemical compound reference gene expression signatures (pioneered in [15]. In contrast to gene expression data repositories such as Chemical Effects in Biological Systems Knowledge Base (http://www.niehs.nih.gov/research/resources/databases/cebs/) that contains data from different laboratories running different experimental platforms on different biological samples, the data in a reference gene expression signature database comes from systematic screening of chemical compounds against specific cell lines simulating biological conditions. Unfortunately, building such a reference gene expression signature database with many different cell types and compound concentrations represented is not easily feasible due to high costs and long development time [80]. For example, in the case of the Connectivity Map (cMap) (http://www.broadinstitute.org/cmap/), which is the largest public reference gene expression signature database (7,000 expression profiles representing 1,309 compounds), not all small molecules were tested in every cell model, and not all were tested across the same spectrum of concentrations. Moreover, how to best interpret the result of a query is still an open question.

As an alternative to such reference databases of gene expression signatures, genes could be annotated to a chemical response through text mining (TM) techniques. The biomedical literature contains a lot of information about small molecules that is not currently stored in gene expression databases, and this information could be used to build chemical response-specific signatures. For example, searching the biomedical literature database PubMed (http://www.ncbi.nlm.nih.gov/pubmed/) with the search string "79983-71-4 OR hexaconazole" results in 69 entries, while no information about the chemical can be found in the CEBS Knowledge Base or cMap (search performed Nov 21, 2011). In addition, TM-derived chemical response-specific signatures would not be specific to a certain biological system or compound concentration, and might therefore provide a less biased view on compound action than standard connectivity maps. Given a gene expression experiment where a biological system has been exposed to a chemical, these TM-generated chemical response-specific signatures could then be tested against the gene expression data set using gene set analysis (GSA) methods [218]. The test would produce a ranking of chemicals in a way similar to the results from a connectivity-mapping query, with the difference that the chemicals are represented by gene sets derived from the literature instead of gene expression signatures from a reference database.

Gene set analysis (GSA) has fast become one of the standard methods in bioinformatics, and there are many tools available (for a review of tools, see [164]). Most GSA tools provide gene sets based on the Gene Ontology (GO) [163], with only a few providing additional sources of gene sets such as metabolic pathways, protein domains, disease associations, tissue expression, transcription factors sequence motifs, miRNA sequences, drug-gene associations [164] and toxicologically relevant gene sets (Boorsma *et al*., submitted). GSA with literature-derived chemical response-specific gene sets has been used before to relate chemical structures to gene expression patterns in microarray experiments. Minguez *et al*. [72] used their tool MarmiteScan to associate chemicals with the characteristics of acute myeloid leukemia cell differentiation. However, there is no information about the size and scope of the chemical dictionary they used to mine the literature and their gene sets are not separately available, thus forcing the researcher to use their GSA method. There is also no possibility to test a sub-set of the gene sets, for example only those that are relevant for evaluation of developmental toxicity. In contrast, we provide chemical response-specific gene sets that can be used with any GSA tool that allows for user-supplied gene sets. Patel and Butte [80] used the

hypergeometric test to associate gene sets derived from curated chemical-gene interactions from > 4,000 chemicals in the Comparative Toxicogenomics Database (CTD) [219] with six gene expression data sets selected based on their diversity with respect to species, chemical exposure and cell type. Manual curation of chemical-gene interactions from publications is a very time-consuming process producing high-quality information but with limited coverage, reflected by the number of chemicals that Patel and Butte could create gene sets for (1338 chemicals). The chemical dictionary used in the CTD limits the number of possible chemicals that can be annotated with genes in the CTD. CTD's chemical dictionary is a modified subset of descriptors from the "Chemicals and Drugs" category and Supplementary Concept Records from the U.S. National Library of Medicine (NLM) Medical Subject Headings, a hierarchical vocabulary used to index PubMed articles containing > 100 000 chemicals. We aim to increase the number of chemicals that can be annotated with genes by using TM instead of manual curation, and by the use of a larger chemical dictionary based on an aggregation of a number of chemical dictionaries. Jelier *et al*. [220] used the weighted global test, with weights based on drug-gene associations generated from text-mining (TM), to associate known peroxisome proliferator-activated receptor alpha (PPARalpha) agonists with the gene expression differences between PPARalpha-null and wild-type mice exposed to the fibrate WY14643. Although Jelier *et al*. demonstrated the usefulness of the TM technology for gene set testing, its reliability still remains to be investigated by evaluating other chemical classes than drugs and considering a wider range of gene expression data sets using other GSA methods; this validation is one of the aims of the present study.

According to a predefined procedure which avoids researcher bias we generated gene sets using TM technology, and test these gene sets using three different GSA tools: ToxProfiler (Boorsma, Hoogstrate, Hettne, Someren, and Stierum, submitted; http://www.toxprofiler.org), the weighted global test (http://www.bioconductor.org/packages/release/bioc/html/globaltest.html), and GeneCodis (http://genecodis.dacya.ucm.es). Hereby we would like to show that the usefulness of TM-based gene sets is not tied to a specific GSA method. To show the versatility of data for which the technology is applicable, we test our TM-based gene sets with the GSA tools in three different case studies. The first study focuses on identification of the chemical treatment, the second on pharmacological mechanism elucidation, and the third on compound toxicity profile comparison. We name these case studies 1, 2 and 3 in the rest of this paper.

Case study 1 comprises the same gene expression experiments used by Patel and Butte [80], as mentioned earlier. We aim to compare the performance of CTD-based gene sets with TM-based gene sets in predicting the particular chemical treatment response. In case study 2 we analyze the gene expression experiment used by Jelier and coworkers [220], as mentioned earlier. For this case study, we aim to predict the PPARalpha agonism characteristic of fibrates. Fibrates have a profound pharmacological response with well-defined molecular toxicological properties, and constitute a clear test case in which both methods (CTD, TM) together with different gene-set testing approaches (ToxProfiler, weighed global test, GeneCodis) should perform well. To compare our approach to connectivity mapping, we compare the results from the GSA tools with the results from the cMap. Since case study 2 is focused on associating drugs with a gene expression profile and cMap contains mainly drugs, this case study was considered an appropriate choice for which to include a comparison with cMap. In case study 3, we demonstrate the power of the TM-based gene sets to link chemicals with similar gene expression response, where the manually curated gene sets from the CTD fail due to lacking chemical-gene annotations. We do this by analyzing a recently published *in vitro* gene expression data set on three chemicals belonging to the triazole class of developmental toxins [5] with the aim to find chemicals with known, or unknown (to our knowledge), links to the class. Using TM, we make chemical response-specific gene sets for the chemicals contained in the ToxRefDB_DevTox database (http://www.epa.gov/ncct/toxrefdb/) [221] and link these gene sets to the data set for the triazoles. For the same gene expression data set we show that TM-derived gene sets also can be used for other purposes than chemical similarity matching. As an example,

gene sets associated with embryonic structures are used to discriminate triazoles from other developmental toxicants, and from non-developmental toxicants.

Methods

Associating chemicals with genes in the literature

TM for chemical-gene associations

The TM-based gene sets were compiled based on a literature-derived score for the chemical-gene association. The association score is calculated by matching the concept profile of a gene with the concept profile of a chemical. The technology behind the creation and matching of concept profiles with the purpose to analyze gene expression data has been described elsewhere [64]. In short, software that performs thesaurus-based indexing is used to identify biomedical concepts in documents. Concept occurrence statistics are then used to build a concept profile, in reality a vector, containing all concepts related to the main concept and weighted by their importance. To calculate the association score between two concept profiles, the cosine of the angle between the two concept vectors is calculated. The literature corpus used in this study consists of titles, Medical Subject Headings (MeSH) headings, and abstracts from the PubMed database (http://www.ncbi.nlm.nih.gov/pubmed). A total of 13,834,150 PubMed IDs from January 1, 1980 to January 6, 2011 were used (PubMed IDs referring to experiments used in the case studies in this work were excluded).

The thesaurus is composed of four parts: the 2010AB version of Unified Medical Language System (UMLS) (http://www.nlm.nih.gov/research/umls/), a gene thesaurus derived from multiple databases [139]), a chemical thesaurus derived from multiple databases [110], and a toxicology thesaurus derived from the International Union of Pure and Applied Chemistry (IUPAC) glossary of terms used in toxicology (http://sis.nlm.nih.gov/enviro/iupacglossary/frontmatter.html). The thesaurus creation and testing procedure has been described in detail elsewhere (Singh, Hettne, van Schouwen, Kors, van Mulligen, and Roos, submitted).

For concept profile matching, the top 200 concepts in a concept profile were used. To make the gene sets we used a concept profile matching score cut-off of 1e-04 [220] in combination with a maximum of 1000 gene associations per chemical. To allow for comparison between the CTD-based and TM-based gene sets, only chemicals with Chemical Abstract Service (CAS) numbers were included. Using these restrictions, we were able to create 30,211 chemical response-specific gene sets.

Processing the CTD for chemical-gene associations

The CTD includes manually curated cross-species interactions between chemicals, genes, and diseases. We downloaded the chemical–gene interaction database from the CTD on January 6, 2011. The database contained 266,266 interactions in total. After filtering for *H. sapiens* and *M. musculus* (the two species included in the gene expression experiments used in this paper), 185,792 interactions remained. The nature of the interaction was not taken into account. All different gene interactions with a specific chemical were summarized into one chemical response-specific gene set. During this step, all interactions based on any of the gene expression experiments used in the case studies in this work were removed. From the single chemical-gene associations in the CTD, we created gene sets with at least five genes (similar to Patel and Butte [80]) per chemical. For every chemical-gene association, CTD provides the CAS number for the chemical (if available), the Entrez Gene ID for the gene, and the PubMed ID and organism for which the association was reported. When filtering for chemicals with a CAS number and restricting the organisms to human and mouse we were able to make a total of 1,189 (*H. sapiens*) and 588 (*M. musculus*) gene sets. Sometimes in the CTD, *H. sapiens* Entrez Gene IDs are incorrectly annotated to *M. musculus*. These *H. sapiens* Entrez Gene IDs were mapped to *M. musculus* Entrez Gene IDs using the Homologene database (http://www.ncbi.nlm.nih.gov/homologene).

Tools selection

We selected different GSA tools that all allow for user-supplied gene sets but implement different statistical tests: ToxProfiler, implementing the unpaired t-test to score the difference between the genes from a pre-defined gene set and the remainder of the genes [222], the weighted version of the global test [9] where TM-derived matching scores are used to weigh the contribution of each gene in a gene set to the test [220], and GeneCodis [223], where biological annotations, or relationships among annotations based on co-occurrence patterns, are tested for over-representation in a list of differentially expressed genes with respect to a reference gene list using the hypergeometric test or the chi-square test. ToxProfiler and the weighted global test do not require a pre-selection of differentially expressed genes while GeneCodis does require such a list. For all tools, p-values were corrected for multiple testing using the False Discovery Rate (FDR) [224], and corrected p-values < 0.05 were considered significant.

In case study 2 we use the text-mining tool Anni (http://www.biosemantics.org/anni) [64] to explore the relevance of the top significant results on a biological process level. Anni is based on the same concept profile technology that was used to generate the TM-based gene sets, and provides direct links to the literature for the produced annotations.

Gene expression microarray data sets selection and pre-processing

Gene expression profiling of Bisphenol A effects on human Ishikawa cells (GEO accession number: GSE17624, set short name: BPA)

The aim of the BPA study was to provide a comprehensive evaluation of changes in gene expression during treatment with Bisphenol A in vitro, and the study was performed using five doses at three different time points with four replicates each [225]. RNA was hybridized on an Affymetrix Human Genome U133 Plus 2.0 Array. For this study, we used the high dosed (10 nM) cells at 48 hours (four treated samples and four control samples).

Gene expression profiling of 17 beta-Estradiol effects on human MCF7 breast cancer cells (GEO accession number: GSE11352, set short name: ESThsa)

The original aim of the ESThsa study was to identify 17 beta-Estradiol responsive genes in the estrogen-receptor positive breast cancer cell line, MCF7 [226]. MCF7 cells were exposed to 10 nM Estradiol (or vehicle only) at 12, 24, and 48 hours. Each time point was performed in triplicate. RNA was hybridized on an Affymetrix Human Genome U133 Plus 2.0 Array. For this study, we used only the samples from 24 hours with their corresponding controls (three treated samples and three control samples).

Gene expression profiling of 17 beta-Estradiol effects on mouse thymus (GEO accession number: GSE2889, set short name: ESTmmu)

The aim of the ESTmmu study was to compare the effects of Estradiol and its analog Genistein on mouse thymus [227]. Control samples were from the mice that were untreated (day 0). Two treatments (Estradiol injection and Genistein diet) and three time points were studied. Two replicated samples at each time point and each treatment were collected. RNA was hybridized on an Affymetrix Mouse Expression 430A Array. For this study, we used only the samples from the mice treated with Estradiol (day 2) and their corresponding controls (two treated samples and two control samples).

Gene expression profiling of 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD) effects on mouse liver (GEO accession number: GSE10082, set short name: TCDD)

The aim of the TCDD study was to map the complete spectrum of aryl hydrocarbon receptor (Ahr) dependent genes in male adult liver by contrasting mRNA profiles of Ahr-null mice (Ahr–/–) with those in mice with wild-type Ahr (Ahr+/+) [228]. Transcript

profiles were determined both in untreated mice and in mice treated 19 h earlier with 1000 µg/kg TCDD. RNA was hybridized on an Affymetrix Mouse Genome 430 2.0 Array. For this study, we used only the data from the wild-type mice (six treated samples and five control samples).

Gene expression profiling of 1alpha,25-Dihydroxyvitamin D3 (VitD3) effects on bronchial smooth muscle cells (GEO accession number: GSE5145, set shortname: VitD3)

The aim of the VitD3 study was to study gene regulation in bronchial smooth muscle cells following VitD3 stimulation [229]. The cells were from the same patient in all hybridization. Cells were treated for 24 hours with 100 nM of VitD3 or with the same concentration of vehicle (ethanol at 0.05%). The experiment was done in triplicates and a total of six samples (three treated and three control) were analyzed. RNA was hybridized on an Affymetrix Human Genome U133 Plus 2.0 Array.

Gene expression profiling of zinc sulfate (ZnSO4) effects on human bronchial epithelial cells (GEO accession number: GSE2111, set short name: ZnSO4)

The aim of the ZnSO4 study was to discriminate vanadium (VOSO4) from ZnSO4 using gene profiling [230]. Human bronchial epithelial cells were treated with vehicle (control), VOSO4 (50 uM) or ZnSO4 (50 uM) for four hours (four replicates each). RNA was hybridized on an Affymetrix Human Genome U133A Array. For this study, we used only the data from the four samples treated with ZnSO4 and their corresponding four control samples.

Gene expression profiling of WY14643 effects on mouse small intestine (GEO accession number: GSE9533, set short name: PPARA)

The aim of the PPARA study was to examine the effects of acute nutritional activation of PPARalpha on expression of genes encoding intestinal barrier proteins [204]. Male, four months old Wild-type (129S1/SvImJ) and PPARalpha -/- mice (129S4/SvJae) were exposed to dietary fatty acids with WY14643 as a reference during an exposure time of six hours, after which the small intestines were removed. RNA was hybridized on an Affymetrix Mouse Genome 430 2.0 Array. Here we use the samples corresponding to the PPARalpha activation by WY14643 (four treated samples and four control samples).

The Affymetrix CEL files for all the different gene expression data sets listed above were preprocessed using GenePattern [231]. CEL files were normalized applying Robust Multichip Average (RMA), using the ExpressionFileCreator module. In addition, the MBNI Custom CDF, which contains updated probe set definitions for Entrez GeneIDs, was applied:
http://brainarray.mbni.med.umich.edu/Brainarray/Database/CustomCDF/11.0.1/entrezg.asp. After normalization, the data was floored by setting a threshold value of 50 for all probe sets. Log(2)ratios were created by dividing the median of the treatment values of each probe set by the median of the control values. Probe sets were discarded if both values were equal to 50.

No supplementary CEL files were available for the Zinc sulfate dataset (GEO: *GSE2111*) and the estradiol mouse dataset (GEO: *GSE2889),* instead the provided MAS5-calculated signal intensities were used to calculate the log(2)ratios. Affymetrix probeset identifiers were mapped to entrez gene identifiers using the appropriate affymetrix CDF. Log(2)ratios were created by dividing the median of the values of each probe set by the median of the time matched control where only probesets were selected that contained a "present" flag.

Gene expression profiling of triazole effects on mouse embryonic stem cell differentiation (ArrayExpress accession number: E-MTAB-300, set short name: triazoles)

In the original experiment, M. musculus embryonic stem cells were exposed to a range of developmental toxicants and non-developmental toxicants (carbamazepine, flusilazole, hexaconazole, methotrexate, methylmercury chloride, monobutyl phthalate, monoethylhexyl phthalate, monomethyl phthalate, nitrofen, saccharine, triadimefon, warfarin) with the aim to evaluate developmental toxicant identification using gene expression profiling in embryonic stem cell test (EST) differentiation cultures [5]. RNA was hybridized on an Affymetrix Mouse Genome 430 2.0 Array. For the analysis with the weighted global test, we used the data at the 24 hour timepoint for the three different thriazoles (flusilazole, hexaconazole, and triadimefon), one negative control (saccharine), and one unexposed control (dimethyl sulfoxide), each with eight replicates. The Affymetrix CEL files for the triazoles gene expression data set were normalized using the expresso package in R with the default settings. Due to the large number of replicates, no probe filtering was performed. Probesets were annotated with Entrez gene IDs using the bioconductor 4302.db package. To summarize the data at the Entrez gene ID level, read-outs from probesets with the same Entrez gene ID were averaged. For the PCA analysis, we used the data from all groups. The Affymetrix CEL files were normalized using Robust Multichip Average (RMA) normalization and probe to gene mapping was performed as described previously [232]. Probe sets for Affymetrix internal controls or probe sets that did not correspond to an Entrez gene ID were not used in further analyses.

Results

Case study 1

In this case study we aimed to compare the performance of CTD-based gene sets with

Table 1. Chemical treatment prediction ranks.
Chemical treatment prediction rank and false discovery rate-corrected p-value (within parentheses) per gene expression data set for the different gene set analysis tools. Significant results (false discovery rate-corrected p<0.05) are in bold. GeneCodis only reports results with a false discovery rate-corrected p-value smaller than 0.05. Missing values denote such non-significant results.

	ToxProfiler		**Weighted global test**		**GeneCodis single annotation**	
	CTD	**TM**	**CTD**	**TM**	**CTD**	**TM**
BPA	**9** (2.0e-07)	92 (1.0e-01)	**18** (3.9e-04)	**21** (6.2e-05)	**10** (9.1e-08)	-
ESThsa	38 (8.8e-02)	400 (2.4e-01)	**269** (3.2e-03)	**920** (5.6e-04)	**4** (2.4e-02)	-
ESTmmu	**18** (3.5e-05)	**189** (1.3e-04)	360 (5.7e-01)	482 (4.3e-01)	-	-
TCDD	**19** (8.0e-07)	**89** (1.2e-03)	**14** (2.9e-10)	**84** (5.7e-10)	**1** (5.3e-32)	**89** (1.7e-05)
VitD3	**9** (4.5e-02)	400 (3.0e-01)	**186** (2.4e-03)	**762** (6.0e-04)	**1** (2.0e-26)	**200** (6.0e-03)
ZnSO4	**9** (3.8e-05)	**6** (3.4e-04)	**3** (2.6e-06)	**4** (6.8e-05)	-	-

TM-based gene sets in predicting the particular treatment response for six gene

expression data sets. Patel and Butte [80] used the hypergeometric test to test their CTD-based gene sets on the same six gene expression data sets. By using the same lists of differentially expressed genes (kindly supplied by C.J. Patel and A.J. Butte, see [80] for details on the preparation of the lists) and selecting the hypergeometric test in GeneCodis, we allow for comparison with their results. When intersecting the CAS numbers for the CTD and TM gene sets, 1,179 H. sapiens and 585 M. musculus gene sets remained. The H. sapiens gene sets were tested with the GSA tools against the human gene expression data sets (BPA, ESThsa, VitD3 and ZnSO4) and the M. musculus gene sets were tested with the GSA tools against the mouse gene expression data sets (ESTmmu and TCDD). Ranking based on the FDR-corrected p-value for the gene sets from each method (CTD, TM) for each GSA tool was used as outcome measure. The CTD-based gene sets ranked consistently and considerably higher than the TM-based gene sets over all GSA tools, with one exception: the TM-based ZnSO4 gene set ranked higher on the ZnSO4 gene expression data set using ToxProfiler (Table 1). On average, the gene set representing the treatment was significantly altered in three experiments using the

TM-based gene sets and in five experiments using the CTD-based gene sets (Table 1). An exception is the weighted global test, for which both approaches scored significant in five experiments.

The statistical test behind ToxProfiler allows for comparison of gene sets over different gene expression data sets based on the t-values produced by the test [222]. Such a profile for a gene set is called a t-profile, and gives information about the specificity of the gene set (that is, if the gene is significantly differential expressed in more than one data set). We compared the t-profiles for the CTD-based and TM-based gene sets that corresponded to the chemical treatments for the data sets. The t-profiles were made on species level so that the gene sets containing mouse genes were compared across the mouse gene expression data sets and the gene sets containing human genes were compared across the human gene expression data sets. T-profiling of the CTD-based and TM-based gene sets corresponding to the chemical treatment for the different gene expression data sets showed similar t-profiles for three (Figure 1A and B, and Figure 2D) of the six gene sets. The other three gene sets (Figure 2A-C) had a dissimilar t-profile.

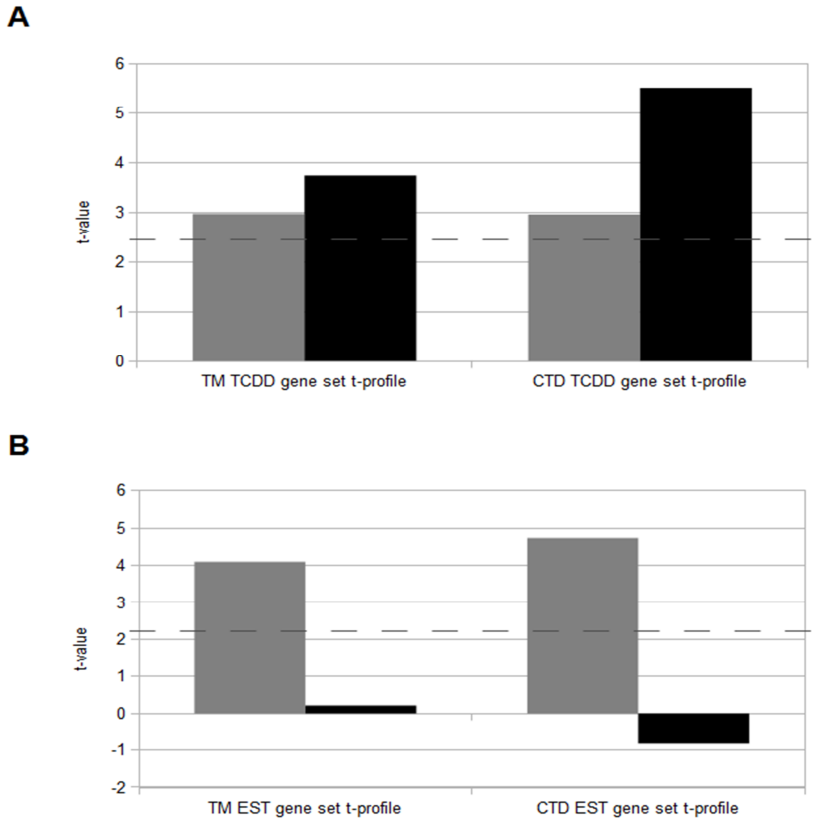

Figure 1. T-profiles of mouse gene sets.
T-profiles of the CTD-based and TM-based TCDD mouse gene sets (A) and Estradiol (EST) mouse gene sets (B) for the TCDD (black bar) and ESTmmu (dark gray bar) mouse gene expression data sets. The dotted line represents the border of significant t-values.

The CTD-based gene sets ranked highest when using the GeneCodis single annotation option (Table 1), but for some data sets the gene set representing the treatment did not score significant. For CTD, the Estradiol and ZnSO4 treatments were not significant. When opting for co-occurring annotations, GeneCodis additionally predicted the CTD-based Estradiol mouse gene set for the ESTmmu gene expression data set as significant with a rank of 14, together with the co-occurring gene sets TCDD and Tretionin. Zinc and zinc chloride CTD-based gene sets were reported as significant at rank 1 together for the ZnSO4 gene expression data set, but not ZnSO4. For TM, three additional gene sets were reported significant when opting for co-occurring annotations: the Estradiol human gene

set for the ESThsa gene expression data set with a rank of 164 together with nine other chemicals; the Estradiol mouse gene set for the ESTmmu gene expression data set with a rank of 41 in a cluster with 11 other chemicals; the ZnSO4 gene set for the ZnSO4 gene expression data set with a rank of 1 in a cluster together with 12 other chemicals. Among the other tools, the weighted global test had the highest number of significant scoring gene sets while ToxProfiler had the better ranking in most cases (Table 1).

In short, the CTD-based gene sets ranked higher than the TM-based gene sets for all gene expression data sets except the ZnSO4 data set, but t-profiling indicated a similar significance pattern for 50% of the TM-based and CTD-based gene sets.

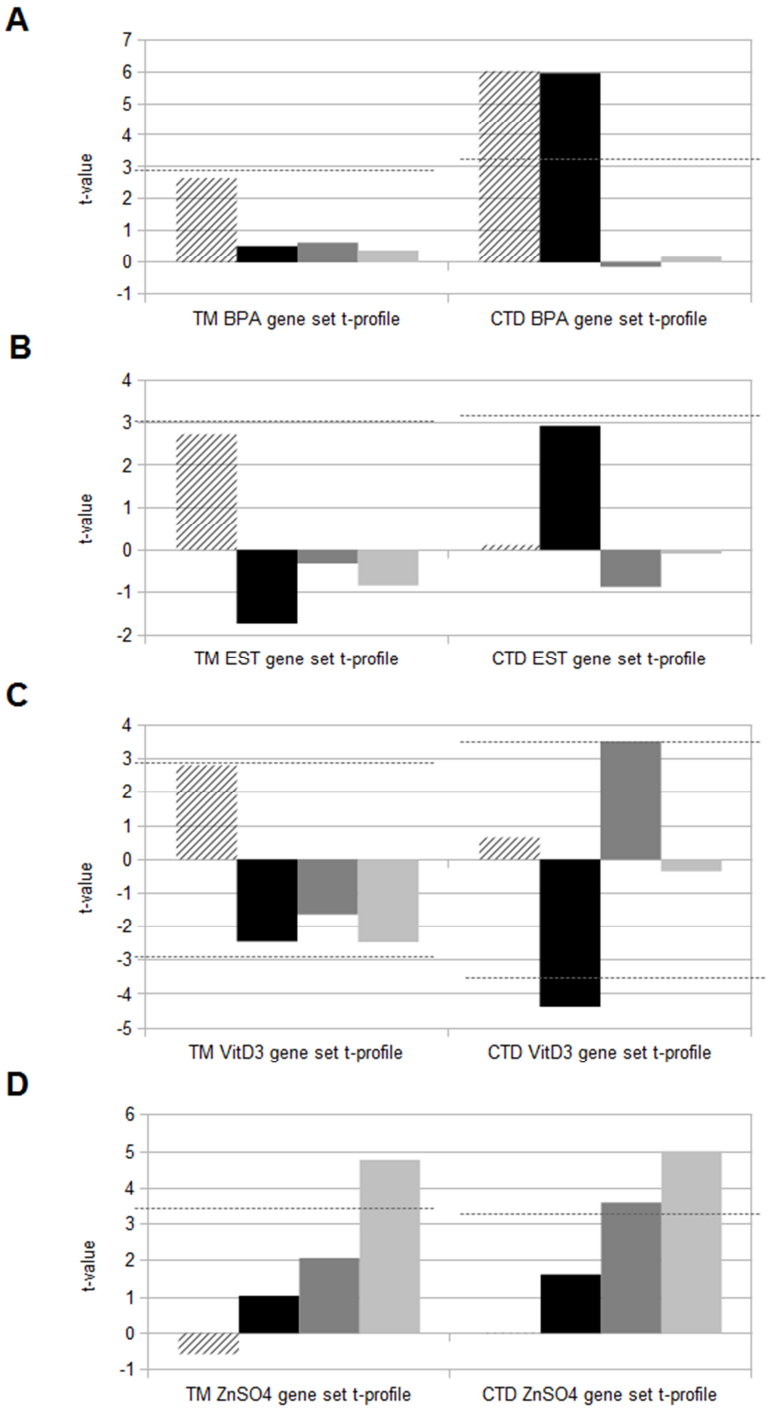

Figure 2. T-profiles of human gene sets.

T-profiles of the CTD-based and TM-based BPA human gene sets (A), the Estradiol (EST) human gene sets (B), the VitD3 human gene sets (C), and the ZnSO4 human gene sets (D) for the BPA (striped bar), ESThsa (black bar), VitD3 (dark gray bar) and ZnSO4 (light gray bar) human gene expression data sets. The dotted line represents the border of significant t-values.

Case study 2

In this case study we analyzed a PPARalpha-knock-out gene expression data set with the aim to predict the PPARalpha agonism characteristic of fibrates. Jelier *et al*. [220] used the weighted global test to associate drugs with the same PPARA gene expression data set. They reported significant results for the known PPARalpha agonists clofibrate, gemfibrozil, bezafibrate and fenofibrate. We aim to predict these fibrates and two more (ciprofibrate and WY14643), and to show that the significantly scoring chemicals relate to relevant biological processes.

When intersecting the CAS numbers for the CTD and TM gene sets, 585 *M. musculus* gene sets remained. The resulting 585 *M. musculus* CTD-based and TM-based gene sets were tested with the GSA tools against the PPARA gene expression data set, and compared to the results from the cMap. As mentioned before, ToxProfiler and the weighted global test do not require a list of differentially expressed genes, but GeneCodis does. To generate the list of differentially expressed genes needed for the analysis with GeneCodis, we used the topTable function from the bioconductor Limma package in R to obtain a ranked list of probes with the most evidence of differential expression between the knockout and wild-type samples. Probes with an adjusted *p*-value of less than 0.05 where kept (930 probes). Probesets were annotated with EntrezGene IDs using the bioconductor 4302.db package. To compare the results with cMap, we started with the same probesets and proceeded to make the query signature. cMap only accepts probes from the Affymetrix GeneChip Human Genome U133A Array. We therefore used the NetAffx tool supplied by Affymetrix (http://www.affymetrix.com/analysis/netaffx/index.affx) to map the probe IDs from our signature to Affymetrix GeneChip Human Genome U133A Array IDs.

Fibrate gene sets ranking and number of relevant biological processes annotated to the significant scoring chemical response-specific gene sets were used as outcome measures. All six TM-based fibrate gene sets ranked among the top-10 significant results in ToxProfiler and GeneCodis (using the single annotation option in GeneCodis) (Table 2). Three CTD-based fibrate gene sets ranked within the top-10 significant results in ToxProfiler and GeneCodis (using the single annotation option in GeneCodis) (Table 2). When opting for co-occurring annotations in GeneCodis, GeneCodis additionally predicted the CTD-based ciprofibrate gene set as significant with a rank of 554 in a cluster together with Acetaminophen, WY14643, TCDD, and Ethinyl Estradiol. All six fibrate gene sets scored significant for both CTD and TM using the weighted global test, but with lower average rank than for the other tools (Table 2). Jelier and co-workers [220] used the weighted global test to test four of the fibrates and reported ranks between 210 and 378. However, they used a different selection of chemicals (the semantic category "drugs" in the thesaurus) to test against the gene expression experiment. When testing the same selection of chemicals but with our concept profile matching association scores, the fibrates ranked between 183 and 304.

None of the fibrate signatures in cMap scored significant against the PPARA gene expression set signature.

In the PPARA study [204], a list of differentially expressed genes was manually annotated with the following categories of biological processes: fatty acid oxidation, cholesterol flux, glucose transport, amino acid metabolism, intestinal motility, and oxidative stress. To investigate if the significant chemicals from the GSA tools and cMap were annotated with these biological processes, we matched the concept profiles for the significantly scoring chemicals against the concept profiles of the following semantic categories in Anni: cell function, molecular function, molecular dysfunction, organ or tissue function (this combination of semantic categories covered all biological process categories from the PPARA study). If more than 100 chemicals had scored significant

using the GSA tools and cMap, the top-100 were used for the analysis. Venn diagrams of overlapping top-100 biological processes for the significant chemicals for each GSA tool and cMap showed higher overlap between CTD and TM, than between cMap and CTD, or cMap and TM (Figure 3). For ToxProfiler and cMap, no common concept could be found for CTD, TM and cMap, but 13 matched either TM or CTD (Figure 3A). For the weighted global test and cMap, four common concepts were found (Figure 3B), as for GeneCodis and cMap (Figure 3C). We then manually inspected the top 100 biological processes. Among the biological processes that were manually annotated to a list of differentially expressed genes in the PPARA study, fatty acid oxidation, cholesterol flux, glucose transport and oxidative stress were found among the top-100 biological processes for the significant chemicals for all GSA tools. For the significant chemicals from cMap, fatty acid oxidation and intestinal motility concepts where found.

Table 2. Fibrate drug prediction ranks.
Fibrate drug prediction ranks and false discovery rate-corrected p-values (within parentheses) for the PPARA gene expression data set for the different gene set analysis tools. Significant results (false discovery rate-corrected p<0.05) are in bold. GeneCodis only reports results with a false discovery rate-corrected p-value smaller than 0.05. Missing values denote such non-significant results. ToxProfiler does not report false discovery rate-corrected p-value less than 1.e-10. These are denoted as <1.e-10 in the table.

	ToxProfiler		**Weighted global test**		**GeneCodis single annotation**	
	CTD	**TM**	**CTD**	**TM**	**CTD**	**TM**
Clofibrate	**7** (6.1e-10)	**2** (<1e.-10)	**22** (1.0e-03)	**71** (1.0e-06)	**5** (1.1e-04)	**2** (7.6e-29)
Gemfibrozil	53 (2.0e-01)	**8** (<1e.-10)	**35** (2.0e-03)	**106** (5.4e-06)	-	**6** (3.5e-19)
Bezafibrate	54 (2.5e-01)	**7** (<1e.-10)	**42** (2.1e-03)	**105** (5.3e-06)	-	**5** (9.3e-22)
Fenofibrate	**3** (<1e.-10)	**10** (<1e.-10)	**27** (1.6e-03)	**115** (7.1e-06)	**3** (2.2e-07)	**9** (1.5e-18)
Ciprofibrate	**33** (4.2e-02)	**5** (<1e.-10)	**21** (8.7e-04)	**97** (2.6e-06)	-	**3** (3.1e-25)
WY14643	**2** (<1e.-10)	**4** (<1e.-10)	**12** (5.5e-05)	**118** (7.2e-06)	**1** (3.2e-79)	**4** (6.5e-22)

In short, The TM-based fibrate gene sets ranked similar to or better than the CTD-based fibrate gene sets using two of the GSA tools, and all TM-based fibrate gene sets were significant using all GSA tools. The top-scoring chemicals from the GSA analyses represented underlying biological processes relevant to the gene expression experiment, both for the CTD-based and TM-based gene sets. In contrast, none of the fibrate signatures in cMap scored significant against the PPARA gene expression set signature, and fewer relevant biological processes were found for the top-ranking chemicals from cMap.

Case study 3

The triazoles gene expression data set was analyzed with two aims: 1) to link chemicals with biological effects derived from -omics data similar to triazoles, and 2) to discriminate triazoles from other developmental toxicants (carbamazepine, methotrexate, methylmercury chloride, monobutyl phthalate, monoethylhexyl phthalate, nitrofen, warfarin), and from non-developmental toxicants (monomethyl phthalate and saccharine) using gene sets associated with embryonic structures.

Only the weighted global test was applied in this case study. The triazoles gene expression data set show very small variation in gene expression levels [5]. The weighted global test has a very strong null hypothesis, asserting that no gene with a positive importance weight is associated with the response, making it an appropriate test for analyzing such a data set.

For aim 1, a subset of the 30,211 TM gene sets derived from chemicals with CAS numbers was used. We only considered mouse genes, since the gene expression experiment was performed on mouse embryonic stem cells. Chemicals were selected based on the list of 384 compounds tested for developmental toxicity contained in the ToxRefDB_DevTox database. The number of similar morphological developmental toxicity endpoints *in vivo* in rabbit or mouse (as recorded in the ToxRefDB_DevTox database)

between the significantly scoring chemicals and the triazoles was used as outcome similarity measure. The morphological developmental toxicity endpoints were the following: cleft lip and/or cleft palate; variations or abnormalities of the limb, including scapula, clavical and pelvis; variations or abnormalities of the vertebral column, ribs or sternum; variations or abnormalities of the cranium; abnormalities of the metanephric kidney; and abnormalities of the ureter. The triazoles also caused general developmental toxicity endpoints: change in weight of fetus at near-term of pregnancy; histopathological, clinical, or unclassified abnormalities in fetus; preimplantation loss, postimplantation loss (resorptions) or fetal death impacting litter size; and pregnancy loss or maternal wastage, but these endpoints were not considered specific enough to be included in the similarity outcome measure.

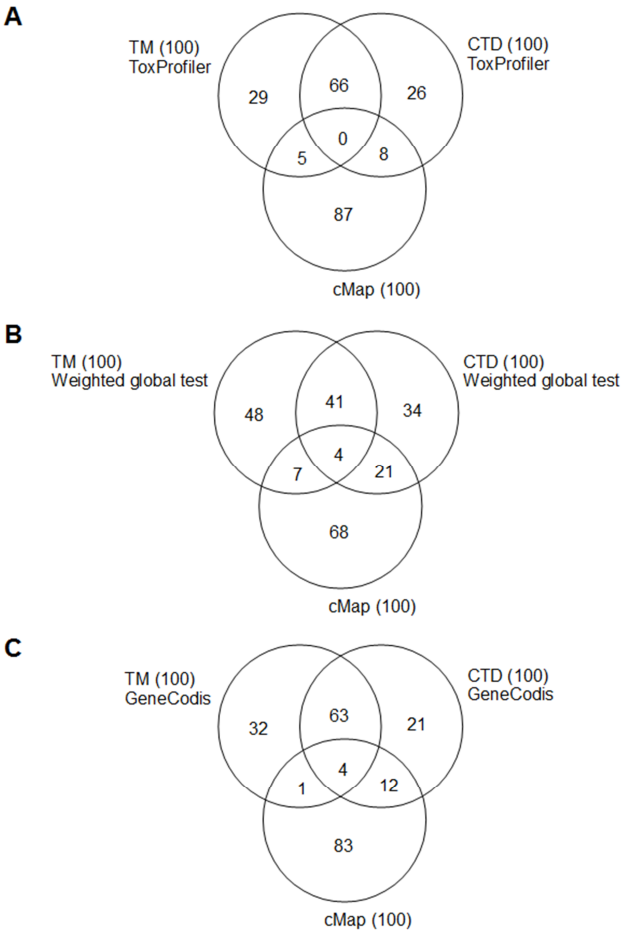

Figure 3. Venn diagrams showing the overlap in biological processes.
Venn diagrams showing the overlap in biological processes for the most significant (false discovery rate-corrected p-value<0.05) chemicals between ToxProfiler and cMap (A), between the weighted global test and cMap (B), and between GeneCodis and cMap (C).

After matching the CAS numbers of the chemicals in the ToxRefDB_DevTox to the CAS numbers of compounds with TM-based gene sets, 319 gene sets remained. Matching the CAS numbers of the chemicals in the ToxRefDB_DevTox to the CAS numbers for the CTD gene sets resulted in 30 gene sets. Of the triazoles used in this study, only triadimefon was included in the CTD gene sets. All triazoles used in this study were included in the TM gene sets. Therefore, we decided to continue the analysis using only the TM-based gene sets. Testing the 319 gene sets against the triazoles gene expression data set with the weighted global test resulted in 33 chemicals with significant changes in gene expression compared to untreated controls (Table 3). Of these 33 chemicals, 21

had a similar morphological developmental *in vivo* toxicity pattern as the triazoles according to the ToxRefDB_DevTox (at least one similar morphological developmental toxicity endpoint (see section Materials and Methods)). Six of these were triazoles: diniconazole, difenoconazole, febuconazole, triadimenol, myclobutanil, and cyproconazole. Of the remaining 12 significantly scoring chemicals, nine had a non-specific pattern of *in vivo* toxicity compared to the triazoles according to the ToxRefDB_DevTox database (at least one similar general developmental toxicity endpoint), and three are reported as non-toxic in the ToxRefDB_DevTox database. Testing the 319 gene sets against the saccharine gene expression data set with the weighted global test resulted in five chemicals with significant changes in gene expression (Primisulfuron, Monosodium methane arsenate, Benfluralin, Pyridaben, Cyprodinil), of which four are developmentally toxic.

Table 3. Number of similar toxicological endpoints.
Number of similar toxicological endpoints and false discovery rate-corrected p-values for chemicals predicted similar to the triazole gene expression data set using the weighted global test. Chemicals with a similar morphological developmental in vivo toxicity profile to triazoles are in bold.

Chemical name	CAS number	# similar general developmental toxicity endpoints	# similar morphological developmental toxicity endpoints	p-value
Metam-sodium	137-42-8	3	3	5.6e-03
Mepiquat chloride	24307-26-4	0	2	9.3e-03
Clomazone	81777-89-1	2	3	9.3e-03
6-Deisopropylatrazine	1007-28-9	0	2	9.3e-03
Zoxamide	156052-68-5	0	0	9.3e-03
Iprodione	36734-19-7	4	2	9.3e-03
Monosodium methane arsenate	2163-80-6	3	0	9.3e-03
Cyprodinil	121552-61-2	2	2	9.5e-03
Carfentrazone-ethyl	128639-02-1	0	1	1.3e-02
Primisulfuron-methyl	86209-51-0	1	3	1.3e-02
3-(3,5-Dichlorophenyl)-1,5-dimethyl-3-azabicyclo(3.1.0)hexane-2,4-dione	32809-16-8	0	0	1.7e-02
Bromadiolone	28772-56-7	1	0	1.7e-02
MGK 264	113-48-4	0	0	2.5e-02
Diniconazole	83657-24-3	3	3	2.5e-02
Difenoconazole	119446-68-3	3	1	2.5e-02
Fenpropathrin	39515-41-8	1	0	2.5e-02
Fenbuconazole	114369-43-6	4	1	2.5e-02
Bendiocarb	22781-23-3	2	0	2.5e-02
Dacthal	1861-32-1	1	0	2.5e-02
Fludioxonil	131341-86-1	0	2	2.5e-02
Mancozeb	7-1-8018	3	3	2.5e-02
Triadimenol	55219-65-3	0	3	2.5e-02
Myclobutanil	88671-89-0	2	1	2.5e-02
Cyproconazole	94361-06-5	2	2	2.9e-02
Fluroxypyr	69377-81-7	1	0	3.1e-02
Fluazifop-P-butyl	79241-46-6	2	4	3.5e-02

Ethametsulfuron methyl	97780-06-8	4	2	3.5e-02
Cyclanilide	113136-77-9	0	0	3.8e-02
EPTC	759-94-4	4	1	4.1e-02
Metaldehyde	108-62-3	1	0	4.1e-02
Cyhexatin	13121-70-5	3	0	4.1e-02
Pyrimethanil	53112-28-0	3	2	4.1e-02
Nitrapyrin	1929-82-4	1	1	4.7e-02

For aim 2, we created a total of 442 gene sets associated with embryonic structures and tested these against the triazoles gene expression data sets with the weighted global test. The embryonic structure concepts originated from the semantic category "Embryonic Structure" from the UMLS part in our thesaurus, and the gene sets were created in a similar fashion to the chemical response-specific gene sets (see the TM-based gene set creation section). To validate the relevance of the these gene sets, we compared the top-25 gene set obtained from the weighted global test to the effects of triazoles seen in rat postimplantation Whole Embryo Culture (WEC) [233]. The top-25 gene sets for embryonic structures resulting from the weighted global test correlated well with effects on the branchial arches, otic vesicles, posterior neuropore, heart, and somites as seen in the WEC (Table 5). Three gene sets could not be translated to

Table 4. The neural plate/tube gene set.
False discovery rate-corrected p-values for the genes from the neural plate/tube gene set that contributed most to the weighted global test.

Entrez Gene ID	Gene symbol	p-value
15394	HOXA1	5.2e-10
14472	GBX2	2.4e-07
64290	FOXB1	6.4e-05
94222	OLIG3	2.9e-04
218772	RARB	4.1e-05
20668	SOX13	1.5e-07
22413	WNT2	1.2e-05
320202	LEFTY2	1.5e-04
17292	MESP1	6.9e-05
14174	FGF3	1.1e-04
54352	IRX5	1.2e-05
18423	OTX1	3.1e-07
57028	PDXP	2.0e-06

changes seen in the WEC, since there is no scoring parameter for these changes in the WEC. Two of these gene sets concern the cloacal membrane. Even though there is no annotation for these in the WEC, the results correspond well with the *in vivo* data in the ToxRefDB_DevTox database (triazoles give rise to urogenital malformations). The third gene set (structure of embryo stage 6) has no direct correspondence in the WEC, but might be linked to the decrease in Total Morphological Score as seen in the WEC.

In order to discriminate triazoles from other developmental and non-developmental toxicants using PCA, the most significant genes in the gene set with the lowest p-value from the weighted global test were used in the PCA analysis. These genes were extracted by calling the leafNodes function in the global test package. The leafNodes function gives an efficient summary of the test result by extracting the most significant subset of genes within a gene set using the inheritance multiple testing procedure of J.J. Goeman and L. Finos (as explained in the manual for the weighted global test). The "neural plate/tube" embryonic structure gene set had the highest FDR-

corrected *p*-value (1.5e-07) of the 442 embryonic structure gene sets tested with the weighted global test against the triazoles gene expression data set. No embryonic structure gene sets came out significant after FDR correction for the saccharine gene expression data set. The "neural plate/tube" gene set contained 993 genes, and the 13 genes contributing most to the test (i.e. the leaf nodes) (Table 4) were used for PCA discrimination. Gene expression values of these selected genes were derived from the triazoles data set. These data were used to evaluate discrimination of triazoles from other developmental and non-developmental toxicants using PCA. Gene expression changes within these 13 genes indicated clear separation between the triazoles on one hand and the other compound-exposed cultures as well as the unexposed time-matched cultures on the other hand (Figure 4).
In short, 33 chemicals could be linked to the gene expression changes induced by triazoles. Toxins corresponding to 21 of these 33 had a similar morphological developmental *in vivo* toxicity pattern as the triazoles. Using TM-derived gene sets for embryonic structures, we confirmed the relation between gene expression patterns and embryotoxic effects seen in the Whole Embryo Culture, and discriminated triazoles from other chemicals in a principal component analysis.

Table 5. The top 25 embryonic structure gene sets.
False discovery rate-corrected p-values for the top 25 embryonic structure gene sets from the weighted global test on triazole gene expression data from the embryonic stem cell test, together with the effect seen in rat postimplantation Whole Embryo Culture (WEC).

Embryonic structure	p-value	WEC effect
neural plate and or tube	1.5e-07	Posterior neuropore open
second branchial arch structure	4.9e-06	Branchial bars deformed
fourth branchial arch structure	4.9e-06	Branchial bars deformed
entire fourth branchial arch	4.9e-06	Branchial bars deformed
structure of first pharyngeal pouch	6.3e-06	Otic vesicles deformed
entire first pharyngeal pouch	6.3e-06	Otic vesicles deformed
entire neural tube	6.4e-06	Posterior neuropore open
Branchial Region	7.3e-06	Branchial bars deformed
entire second branchial arch	7.3e-06	Branchial bars deformed
structure of tympanic annulus	7.3e-06	Otic vesicles deformed
entire tympanic annulus	7.3e-06	Otic vesicles deformed
Neuroectoderm	7.5e-06	Posterior neuropore open
entire cloacal membrane	9.8e-06	No corresponding scoring parameter available in the WEC.
branchial arch structure	1.1e-05	Branchial bars deformed
structure of cloacal membrane	1.1e-05	No corresponding scoring parameter available in the WEC.
entire ostium secundum	1.1e-05	Heart ventrally turned
structure of third aortic arch	1.1e-05	Heart ventrally turned
entire branchial arch	2.2e-05	Branchial bars deformed
Otic Vesicle	2.2e-05	Otic vesicles deformed
entire auditory vesicle	2.2e-05	Otic vesicles deformed
structure of embryo stage 6	2.2e-05	No corresponding scoring parameter available in the WEC.
structure of early somite stage	2.2e-05	Somites irregular
entire early somite stage	2.2e-05	Somites irregular
third branchial arch structure	2.2e-05	Branchial bars deformed
entire third branchial arch	2.2e-05	Branchial bars deformed

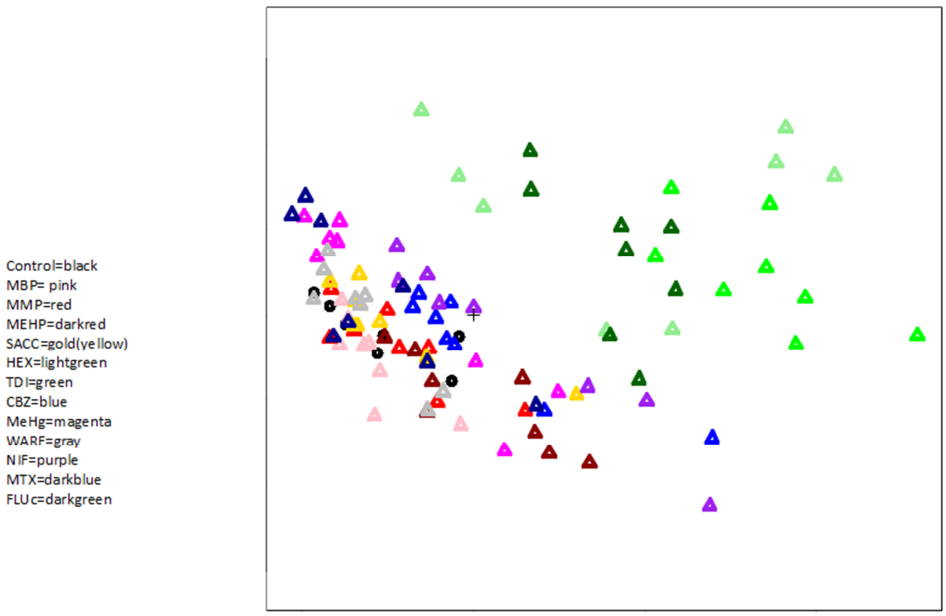

Figure 4. Discrimination of triazoles using PCA.
Discrimination of triazoles (hexaconazole (HEX), triadimefon (TDI), flusilazole (FLUc)) from other developmentally toxic compounds (monobutyl phthalate (MBP), monoethylhexyl phthalate (MEHP), carbamazepine (CBZ), methylmercury (MeHg), warfarin (WARF), nitrofen (NIF), methotrexate (MTX)), non-toxic compunds (monomethyl phthalate (MMP), saccharine (SACC)) and time-matched unexposed cultures (Control) using principal component analysis on the basis of the leafnodes from the neural plate/tube gene set. Variance of the first principal component (horizontal axis): 70%. Variance of the second principal component (vertical axis): 14%.

Discussion

In the present study we perform identification of the chemical treatment, pharmacological mechanism elucidation, and compound toxicity profile comparison by testing chemical response-specific TM-based gene sets with GSA methods against different gene expression data sets. Such experiments have been performed before but were limited to evaluation of one specific tool and used a limited number of literature-based gene sets [72, 80, 220]. In contrast to the tool-specific gene sets that were used for identification of the chemical treatment in Minguez et al., we provide the size and scope of our chemical response-specific gene sets, and make them available for download in a generic format (http://www.biosemantics.org/index.php?page=chemical-response-specific-gene-sets). We extend the evaluation of literature-based gene sets for identification of the chemical treatment performed by Patel and Butte to include TM. In doing this, we were able to create many more chemical response-specific gene sets (~30,200 gene sets for both human and mouse using TM compared to ~1,200 gene sets for human and ~600 gene sets for mouse using the CTD). While Jelier and co-workers tested TM-based gene sets for drugs on one gene expression data set using one GSA method, we tested TM-based gene sets on diverse sets of gene expression data with a variety of methods. We show that gene sets generated by TM are less compound-specific than gene sets based on manual curation efforts but perform equal to or better than manual curation efforts in elucidating the pharmacological mechanism of fibrates. In addition, we show that TM allows for toxicity profile comparison of compounds for which manual annotation efforts are lacking (triazoles). We also successfully use TM-based gene sets for embryonic structures to describe effects induced by the triazoles hexaconazole, triadimefon and flusilazole in the EST, and to discriminate the triazoles from other chemicals using PCA.

A possible limitation of our approach is that we do not take the nature of the relation (for example expression (negative or positive) or phosphorylation) between the genes and chemicals or genes and embryonic structures into account when creating the gene sets. In the CTD, such information has been manually curated for the chemical-gene interactions. For the gene sets generated by text mining, such relations could be mined from the literature by using an ontology of known biological relations. In the future, it could be worth investigating how the nature of the relation influences the results of the GSA.

Identification of chemical treatment

In this case study, we aimed to show that TM-based gene sets compare well to CTD-based gene sets in associating chemicals with gene expression data sets. The CTD-based gene sets ranked higher than the TM-based gene sets for all gene expression data sets except the ZnSO4 data set. This might be expected since the CTD-based gene sets are curated. However, t-profiling of the CTD-based and TM-based gene sets showed similar t-profiles for three of the six gene sets, indicating a similar gene set significance pattern for these gene sets despite the higher ranks of the CTD-based gene sets. The ranks for the CTD-based and TM-based gene sets differed between the different tools and between the experiments, indicating that in general, not one tool is to be preferred over the other when performing GSA analysis-based connectivity mapping. One needs to select the tool best suited for the experimental design. For example, ToxProfiler performs well with small number of replicates, while the weighted global test deals well with small changes in gene expression levels. The CTD-based gene sets had the highest significant ranks when using GeneCodis, which is in line with what Patel and Butte [80] reported for these experiments (using the hypergeometric test). The co-occurring annotations option in GeneCodis proved useful compared to the single annotation option in some cases.

Pharmacological mechanism elucidation

In this case study, we aimed to associate PPARalpha agonists (fibrates) with a PPARalpha knock-out gene expression data set, and to show that the significantly scoring chemicals relate to relevant biological processes. The TM-based fibrate gene sets ranked similar or better than the CTD-based fibrate gene sets using two of the GSA tools, and all TM-based fibrate gene sets were significant using all GSA tools. Clearly, for this experiment, TM presents a comparable alternative to manual inspection of the literature. This might be the case since there is pharmacological and toxicological consensus on the action of peroxisome proliferators, and as such the literature accumulated over the past is more clearly described. The CTD-based gene sets for gemfibrozil and bezafibrate did not score significant using ToxProfiler and GeneCodis. These fibrates also had the lowest number of genes annotated to them: only four for gemfibrozil and nine for bezafibrate could be mapped to the current gene expression experiment. The size of the gene set has less influence when using the weighted global test compared to the other tools, which might explain that these gene sets scored significant using this method. The TM fibrate gene sets ranked much lower when using the weighted global test compared to the other tools. However, rankings were improved compared to the results by Jelier and coworkers [220]. The non-significant results for the fibrates when using cMap indicates that GSA tools present a good alternative when performing connectivity mapping.

We annotated the top-scoring chemicals for CTD and TM in Anni with biological processes to investigate if these corresponded to the ones annotated by hand in the PPARA study. For all GSA tools, biological processes corresponding to all categories reported in the PPARA study were found, except intestinal motility concepts. Intestinal motility concepts could be annotated to the significantly scoring signatures in cMap, while only one more (fatty acid oxidation) of the other five biological processes could be found using this method. These results show that the top scoring chemicals from the GSA analyses represent underlying biological processes relevant to the gene expression experiment, both for the CTD-based and TM-based gene sets.

Compound toxicity profile

The triazoles gene expression data set was analyzed with the aim to predict chemicals with a toxicity profile similar to triazoles and to discriminate triazoles from other developmental toxicants, and non-toxic compounds. Many (64%) of the predicted chemicals had an *in vivo* toxicity pattern corresponding to that of the triazoles. The highest ranking compound with a toxicology pattern different from the triazoles according to ToxRefDB_DevTox was Monosodium methane. Three of the eight genes that contributed most to the significance of this compound where annotated with the concept "axial skeletal structure" in Anni. Triazoles cause malformations in this structure according to the ToxRefDB_DevTox. Even though Monosodium methane does not cause malformations in the axial skeletal structure of rats or rabbits according to the ToxRefDB_DevTox, our results suggest that the compound should be tested for *in vivo* toxicity also in mouse. On the level of risk assessment we are interested in the situation in man, and the more information about the toxic effects of compounds in different systems the better. The highest ranking non-toxic compound (Zoxamide) scored significant mainly because of the presence of the gene CYP51A1 in the gene set. The fungicidal mode of action of triazoles is based on the inhibition of this gene [234]. We noted five significantly scoring chemicals for the saccharine gene expression experiment of which four were developmental toxicants, which confirms that gene set testing with chemical response-specific gene sets should be considered as hypothesis generation and that every significantly scoring compound need to be investigated further. This is something that also applies to standard connectivity mapping. Using the weighted global test, further investigation would constitute inspection of the leaf nodes in the gene sets, since these genes contribute most to the test. Anni can be used to infer the function of the leaf nodes, give information on how these are connected to each other, and via direct links to the literature more thorough information can easily be found.

De Jong and co-workers [233] showed that the embryonic stem cell test (EST) is able to give a relatively good potency ranking compared with the *in vivo* developmental toxicity potency of triazoles, but pointed out that the system gives little information on the type of effects that can occur after exposure to the chemicals. We show that by associating embryonic structure gene sets with a triazole gene expression profile obtained through the EST, effect information becomes readily available. Also, the genes in the embryonic structure gene set that contributed most to the weighted global test could separate triazoles from other chemicals in a PCA. Genes that effectively can separate between chemical classes are usually searched for by applying a statistical test on differential gene expression (see for example [235]). Here, we show that such genes can also be found by applying the weighted global test with relevant gene sets (in this case embryonic structure gene sets).

Conclusions

In conclusion, GSA with TM-derived chemical response-specific gene sets is a scalable method for identifying similarities in gene responses to other chemicals, from which one may infer potential mode of action and/or toxic effect.

Acknowledgements

We would like to acknowledge Herman van Haagen and Erik Schultes for advice on the cut-off for the number of concepts in a concept profile used for matching.

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.

References

1. Russell WMS, Burch RL: **The principles of humane experimental technique**: London Methuen; 1959.
2. Hamadeh HK, Afshari CA: **Toxicogenomics: Principles and Applications**: Wiley-Liss; 2004.
3. Waring JF, Ciurlionis R, Jolly RA, Heindel M, Ulrich RG: **Microarray analysis of hepatotoxins in vitro reveals a correlation between gene expression profiles and mechanisms of toxicity**. *Toxicology letters* 2001, **120**(1-3):359-368.
4. Ellinger-Ziegelbauer H, Stuart B, Wahle B, Bomann W, Ahr HJ: **Comparison of the expression profiles induced by genotoxic and nongenotoxic carcinogens in rat liver**. *Mutation research* 2005, **575**(1-2):61-84.
5. van Dartel DA, Pennings JL, Robinson JF, Kleinjans JC, Piersma AH: **Discriminating classes of developmental toxicants using gene expression profiling in the embryonic stem cell test**. *Toxicology letters* 2011, **201**(2):143-151.
6. van Delft JH, van Agen E, van Breda SG, Herwijnen MH, Staal YC, Kleinjans JC: **Comparison of supervised clustering methods to discriminate genotoxic from non-genotoxic carcinogens by gene expression profiling**. *Mutation research* 2005, **575**(1-2):17-33.
7. Bushel PR, Hamadeh HK, Bennett L, Green J, Ableson A, Misener S, Afshari CA, Paules RS: **Computational selection of distinct class- and subclass-specific gene expression signatures**. *Journal of biomedical informatics* 2002, **35**(3):160-170.
8. Minami K, Saito T, Narahara M, Tomita H, Kato H, Sugiyama H, Katoh M, Nakajima M, Yokoi T: **Relationship between hepatic gene expression profiles and hepatotoxicity in five typical hepatotoxicant-administered rats**. *Toxicol Sci* 2005, **87**(1):296-305.
9. Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC: **A global test for groups of genes: testing association with a clinical outcome**. *Bioinformatics (Oxford, England)* 2004, **20**(1):93-99.
10. Hamadeh HK, Jayadev S, Gaillard ET, Huang Q, Stoll R, Blanchard K, Chou J, Tucker CJ, Collins J, Maronpot R *et al*: **Integration of clinical and gene expression endpoints to explore furan-mediated hepatotoxicity**. *Mutation research* 2004, **549**(1-2):169-183.
11. Steiner G, Suter L, Boess F, Gasser R, de Vera MC, Albertini S, Ruepp S: **Discriminating different classes of toxicants by transcript profiling**. *Environmental health perspectives* 2004, **112**(12):1236-1248.
12. **http://www.genelogic.com/knowledge-suites/toxexpress-program**
13. **http://www.entelos.com/browse.php?ID=technologies&TOPIC=dmtx**
14. Waters M, Stasiewicz S, Merrick BA, Tomer K, Bushel P, Paules R, Stegman N, Nehls G, Yost KJ, Johnson CH *et al*: **CEBS--Chemical Effects in Biological Systems: a public data repository integrating study design and toxicity data with microarray and proteomics data**. *Nucleic acids research* 2008, **36**(Database issue):D892-900.
15. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN *et al*: **The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease**. *Science (New York, NY* 2006, **313**(5795):1929-1935.
16. Jaeschke H, McGill MR, Ramachandran A: **Oxidant stress, mitochondria, and cell death mechanisms in drug-induced liver injury: lessons learned from acetaminophen hepatotoxicity**. *Drug metabolism reviews* 2012, **44**(1):88-106.
17. **http://www.geneontology.org/**

18. **http://www.genome.jp/kegg/**
19. Database CT: **http://ctdbase.org/**.
20. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES *et al*: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles**. *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**(43):15545-15550.
21. Goeman JJ, Oosting J, Cleton-Jansen AM, Anninga JK, van Houwelingen HC: **Testing association of a pathway with survival using gene expression data**. *Bioinformatics (Oxford, England)* 2005, **21**(9):1950-1957.
22. **http://www.ingenuity.com/products/ipa-tox.html**
23. **http://www.genego.com/metadrug.php**
24. ToxProfiler: **http://ntc.voeding.tno.nl/tbase/toxprofiler_public/**.
25. Afshari CA, Hamadeh HK, Bushel PR: **The evolution of bioinformatics in toxicology: advancing toxicogenomics**. *Toxicol Sci* 2011, **120 Suppl 1**:S225-237.
26. Fielden MR, Eynon BP, Natsoulis G, Jarnagin K, Banas D, Kolaja KL: **A gene expression signature that predicts the future onset of drug-induced renal tubular toxicity**. *Toxicologic pathology* 2005, **33**(6):675-683.
27. Huang L, Heinloth AN, Zeng ZB, Paules RS, Bushel PR: **Genes related to apoptosis predict necrosis of the liver as a phenotype observed in rats exposed to a compendium of hepatotoxicants**. *BMC genomics* 2008, **9**:288.
28. Baumgartner WA, Jr., Cohen KB, Fox LM, Acquaah-Mensah G, Hunter L: **Manual curation is not sufficient for annotation of genomic databases**. *Bioinformatics (Oxford, England)* 2007, **23**(13):i41-48.
29. Altman RB, Bergman CM, Blake J, Blaschke C, Cohen A, Gannon F, Grivell L, Hahn U, Hersh W, Hirschman L *et al*: **Text mining for biology--the way forward: opinions from leading scientists**. *Genome biology* 2008, **9 Suppl 2**:S7.
30. Cohen KB, Hunter L: **Getting started in text mining**. *PLoS computational biology* 2008, **4**(1):e20.
31. Hersh WR: **Information retrieval: a health and biomedical perspective**: Springer; 2009.
32. Schuemie M, Jelier R, Kors JA: **Peregrine: Lightweight gene name normalization by dictionary lookup.** In: *Biocreative 2 workshop: 2007; Madrid*.
33. **https://trac.nbic.nl/data-mining/**
34. **http://www.acronymfinder.com/**
35. Schuemie MJ, Kors JA, Mons B: **Word sense disambiguation in the biomedical domain: an overview**. *J Comput Biol* 2005, **12**(5):554-565.
36. Alako BT, Veldhoven A, van Baal S, Jelier R, Verhoeven S, Rullmann T, Polman J, Jenster G: **CoPub Mapper: mining MEDLINE based on search term co-publication**. *BMC bioinformatics* 2005, **6**:51.
37. Barbosa-Silva A, Soldatos TG, Magalhaes IL, Pavlopoulos GA, Fontaine JF, Andrade-Navarro MA, Schneider R, Ortega JM: **LAITOR--Literature Assistant for Identification of Terms co-Occurrences and Relationships**. *BMC bioinformatics* 2011, **11**:70.
38. Hoffmann R, Valencia A: **A gene network for navigating the literature**. *Nature genetics* 2004, **36**(7):664.
39. Jenssen TK, Laegreid A, Komorowski J, Hovig E: **A literature network of human genes for high-throughput analysis of gene expression**. *Nature genetics* 2001, **28**(1):21-28.
40. Garten Y, Altman RB: **Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text**. *BMC bioinformatics* 2009, **10 Suppl 2**:S6.
41. Coulet A, Shah NH, Garten Y, Musen M, Altman RB: **Using text to build semantic networks for pharmacogenomics**. *Journal of biomedical informatics* 2010, **43**(6):1009-1019.

42. Bandy J, Milward D, McQuay S: **Mining protein-protein interactions from published literature using Linguamatics I2E**. *Methods in molecular biology (Clifton, NJ* 2009, **563**:3-13.
43. Kemper B, Matsuzaki T, Matsuoka Y, Tsuruoka Y, Kitano H, Ananiadou S, Tsujii J: **PathText: a text mining integrator for biological pathway visualizations**. *Bioinformatics (Oxford, England)* 2010, **26**(12):i374-381.
44. Krallinger M, Rodriguez-Penagos C, Tendulkar A, Valencia A: **PLAN2L: a web tool for integrated text mining and literature-based bioentity relation extraction**. *Nucleic acids research* 2009, **37**(Web Server issue):W160-165.
45. Swanson DR: **Migraine and magnesium: eleven neglected connections**. *Perspectives in biology and medicine* 1988, **31**(4):526-557.
46. Swanson DR: **Fish oil, Raynaud's syndrome, and undiscovered public knowledge**. *Perspectives in biology and medicine* 1986, **30**(1):7-18.
47. Tsuruoka Y, Miwa M, Hamamoto K, Tsujii J, Ananiadou S: **Discovering and visualizing indirect associations between biomedical concepts**. *Bioinformatics (Oxford, England)* 2011, **27**(13):i111-119.
48. Iossifov I, Rodriguez-Esteban R, Mayzus I, Millen KJ, Rzhetsky A: **Looking at cerebellar malformations through text-mined interactomes of mice and humans**. *PLoS computational biology* 2009, **5**(11):e1000559.
49. Jelier R, t Hoen PA, Sterrenburg E, den Dunnen JT, van Ommen GJ, Kors JA, Mons B: **Literature-aided meta-analysis of microarray data: a compendium study on muscle development and disease**. *BMC bioinformatics* 2008, **9**:291.
50. Hettne KM, Weeber M, Laine ML, ten Cate H, Boyer S, Kors JA, Loos BG: **Automatic mining of the literature to generate new hypotheses for the possible link between periodontitis and atherosclerosis: lipopolysaccharide as a case study**. *Journal of clinical periodontology* 2007, **34**(12):1016-1024.
51. Schuemie M, Chichester C, Lisacek F, Coute Y, Roes PJ, Sanchez JC, Kors J, Mons B: **Assignment of protein function and discovery of novel nucleolar proteins based on automatic analysis of MEDLINE**. *Proteomics* 2007, **7**(6):921-931.
52. Hristovski D, Friedman C, Rindflesch TC, Peterlin B: **Exploiting semantic relations for literature-based discovery**. *AMIA Annual Symposium proceedings / AMIA Symposium* 2006:349-353.
53. Yetisgen-Yildiz M, Pratt W: **Using statistical and knowledge-based approaches for literature-based discovery**. *Journal of biomedical informatics* 2006, **39**(6):600-611.
54. Hristovski D, Peterlin B, Mitchell JA, Humphrey SM: **Using literature-based discovery to identify disease candidate genes**. *International journal of medical informatics* 2005, **74**(2-4):289-298.
55. Chen H, Sharp BM: **Content-rich biological network constructed by mining PubMed abstracts**. *BMC bioinformatics* 2004, **5**:147.
56. Rzhetsky A, Iossifov I, Koike T, Krauthammer M, Kra P, Morris M, Yu H, Duboue PA, Weng W, Wilbur WJ *et al*: **GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data**. *Journal of biomedical informatics* 2004, **37**(1):43-53.
57. Weeber M, Vos R, Klein H, Berg LTWDJ-VD, Aronson AR, Molema G: **Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide**. *J Am Med Inform Assoc* 2003, **10**(3):252-259.
58. Wren JD, Bekeredjian R, Stewart JA, Shohet RV, Garner HR: **Knowledge discovery by automated identification and ranking of implicit relationships**. *Bioinformatics (Oxford, England)* 2004, **20**(3):389-398.
59. Hettne KM, de Mos M, de Bruijn AG, Weeber M, Boyer S, van Mulligen EM, Cases M, Mestres J, van der Lei J: **Applied information retrieval and multidisciplinary research: new mechanistic hypotheses in complex**

regional pain syndrome. *Journal of biomedical discovery and collaboration* 2007, **2**:2.
60. de Mos M, Laferriere A, Millecamps M, Pilkington M, Sturkenboom MC, Huygen FJ, Coderre TJ: **Role of NFkappaB in an animal model of complex regional pain syndrome-type I (CRPS-I)**. *J Pain* 2009, **10**(11):1161-1169.
61. van Haagen HH, t Hoen PA, Botelho Bovo A, de Morree A, van Mulligen EM, Chichester C, Kors JA, den Dunnen JT, van Ommen GJ, van der Maarel SM *et al*: **Novel protein-protein interactions inferred from literature context**. *PloS one* 2009, **4**(11):e7894.
62. Frijters R, van Vugt M, Smeets R, van Schaik R, de Vlieg J, Alkema W: **Literature mining for the discovery of hidden connections between drugs, genes and diseases**. *PLoS computational biology* 2010, **6**(9).
63. Jelier R, Schuemie MJ, Roes PJ, van Mulligen EM, Kors JA: **Literature-based concept profiles for gene annotation: the issue of weighting**. *International journal of medical informatics* 2008, **77**(5):354-362.
64. Jelier R, Schuemie MJ, Veldhoven A, Dorssers LC, Jenster G, Kors JA: **Anni 2.0: a multipurpose text-mining tool for the life sciences**. *Genome biology* 2008, **9**(6):R96.
65. Jelier R, Schuemie MJ, Veldhoven A, Dorssers LC, Jenster G, Kors JA: **Anni 2.0: a multipurpose text-mining tool for the life sciences**. *Genome biology* 2008, **9**(6):R96.
66. Soldatos TG, O'Donoghue SI, Satagopam VP, Jensen LJ, Brown NP, Barbosa-Silva A, Schneider R: **Martini: using literature keywords to compare gene sets**. *Nucleic acids research* 2010, **38**(1):26-38.
67. Sartor MA, Mahavisno V, Keshamouni VG, Cavalcoli J, Wright Z, Karnovsky A, Kuick R, Jagadish HV, Mirel B, Weymouth T *et al*: **ConceptGen: a gene set enrichment and gene set relation mapping tool**. *Bioinformatics (Oxford, England)* 2010, **26**(4):456-463.
68. **http://gene2mesh.ncibi.org**
69. Frijters R, Heupers B, van Beek P, Bouwhuis M, van Schaik R, de Vlieg J, Polman J, Alkema W: **CoPub: a literature-based keyword enrichment tool for microarray data analysis**. *Nucleic acids research* 2008, **36**(Web Server issue):W406-410.
70. Leong HS, Kipling D: **Text-based over-representation analysis of microarray gene lists with annotation bias**. *Nucleic acids research* 2009, **37**(11):e79.
71. Kuffner R, Fundel K, Zimmer R: **Expert knowledge without the expert: integrated analysis of gene expression and literature to derive active functional contexts**. *Bioinformatics (Oxford, England)* 2005, **21 Suppl 2**:ii259-267.
72. Minguez P, Al-Shahrour F, Montaner D, Dopazo J: **Functional profiling of microarray experiments using text-mining derived bioentities**. *Bioinformatics (Oxford, England)* 2007, **23**(22):3098-3099.
73. **http://toxnet.nlm.nih.gov/**
74. **http://www.biocreative.org/**
75. Zimmermann M, Fluck J, Thi le TB, Kolarik C, Kumpf K, Hofmann M: **Information extraction in the life sciences: perspectives for medicinal chemistry, pharmacology and toxicology**. *Current topics in medicinal chemistry* 2005, **5**(8):785-796.
76. Kim JD, Ohta T, Tateisi Y, Tsujii J: **GENIA corpus--semantically annotated corpus for bio-textmining**. *Bioinformatics (Oxford, England)* 2003, **19 Suppl 1**:i180-182.
77. Klinger R, Kolarik C, Fluck J, Hofmann-Apitius M, Friedrich CM: **Detection of IUPAC and IUPAC-like chemical names**. *Bioinformatics (Oxford, England)* 2008, **24**(13):i268-i276.
78. Frijters R, Verhoeven S, Alkema W, van Schaik R, Polman J: **Literature-based compound profiling: application to toxicogenomics**. *Pharmacogenomics* 2007, **8**(11):1521-1534.

79. van Dartel DA, Pennings JL, Hendriksen PJ, van Schooten FJ, Piersma AH: **Early gene expression changes during embryonic stem cell differentiation into cardiomyocytes and their modulation by monobutyl phthalate**. *Reproductive toxicology (Elmsford, NY* 2009, **27**(2):93-102.
80. Patel CJ, Butte AJ: **Predicting environmental chemical factors associated with disease-related gene expression data**. *BMC medical genomics* 2010, **3**:17.
81. Chaussabel D, Sher A: **Mining microarray expression data by literature profiling**. *Genome biology* 2002, **3**(10):RESEARCH0055.
82. Glenisson P, Coessens B, Van Vooren S, Mathys J, Moreau Y, De Moor B: **TXTGate: profiling gene groups with text-based information**. *Genome biology* 2004, **5**(6):R43.
83. Jelier R, Jenster G, Dorssers LC, Wouters BJ, Hendriksen PJ, Mons B, Delwel R, Kors JA: **Text-derived concept profiles support assessment of DNA microarray data for acute myeloid leukemia and for androgen receptor stimulation**. *BMC bioinformatics* 2007, **8**:14.
84. Smalheiser NR, Swanson DR: **Linking estrogen to Alzheimer's disease: an informatics approach**. *Neurology* 1996, **47**(3):809-810.
85. Weeber M, Vos R, Klein H, De Jong-Van Den Berg LT, Aronson AR, Molema G: **Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide**. *J Am Med Inform Assoc* 2003, **10**(3):252-259.
86. Sanfilippo A, Posse C, Gopalan B, Riensche R, Beagley N, Baddeley B, Tratz S, Gregory M: **Combining hierarchical and associative gene ontology relations with textual evidence in estimating gene and gene product similarity**. *IEEE Trans Nanobioscience* 2007, **6**(1):51-59.
87. Schuemie M, Chichester C, Lisacek F, Coute Y, Roes P-J, Sanchez JC, Kors J, Mons B: **Assignment of protein function and discovery of novel nucleolar proteins based on automatic analysis of MEDLINE**. *Proteomics* 2007, **7**(6):921-931.
88. Ananiadou S, Nenadic G: **Automatic Terminology Management in Biomedicine**. In: *Text Mining for Biology and Biomedicine.* Edited by Ananiadou S, McNaught J. Boston: Artech House; 2006: 67–92.
89. Krauthammer M, Nenadic G: **Term identification in the biomedical literature**. *Journal of biomedical informatics* 2004, **37**(6):512-526.
90. Cohen AM, Hersh WR: **A survey of current work in biomedical text mining**. *Brief Bioinform* 2005, **6**(1):57-71.
91. Erhardt RA, Schneider R, Blaschke C: **Status of text-mining techniques applied to biomedical text**. *Drug Discov Today* 2006, **11**(7-8):315-325.
92. Zweigenbaum P, Demner-Fushman D, Yu H, Cohen KB: **Frontiers of biomedical text mining: current progress**. *Brief Bioinform* 2007, **8**(5):358-375.
93. Bodenreider O: **The Unified Medical Language System (UMLS): integrating biomedical terminology**. *Nucleic acids research* 2004, **32**(Database issue):D267-270.
94. **UMLS glossary** [http://www.nlm.nih.gov/research/umls/new_users/glossary.html]
95. Cabré Castellví TM: **Theories of terminology: their description, prescription and explanation**. *Terminology* 2003, **9**(2):163–199.
96. Sager J: **A Practical Course in Terminology Processing**. Amsterdam: John Benjamins Publishing Company; 1990.
97. Wüster E: **General terminology theory – fine line between linguistics, logic, ontology, information science and business sciences**. *Linguistics* 1974, **119**(1):61–106.
98. Srinivasan S, Rindflesch TC, Hole WT, Aronson AR, Mork JG: **Finding UMLS Metathesaurus concepts in MEDLINE**. *Proc AMIA Symp* 2002:727-731.

99. McCray AT, Burgun A, Bodenreider O: **Aggregating UMLS semantic types for reducing conceptual complexity**. *Stud Health Technol Inform* 2001, **84**(Pt 1):216-220.
100. McCray AT, Browne AC, Bodenreider O: **The lexical properties of the gene ontology**. *Proc AMIA Symp* 2002:504-508.
101. **Filtering the UMLS Metathesaurus for MetaMap** [http://skr.nlm.nih.gov/papers/references/filtering07.pdf]
102. Aronson AR: **Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program**. *Proc AMIA Symp* 2001:17-21.
103. **UMLS knowledge server**. In.; 2007.
104. Schwartz AS, Hearst MA: **A simple algorithm for identifying abbreviation definitions in biomedical text**. *Pacific Symposium on Biocomputing* 2003:451-462.
105. Torii M, Hu ZZ, Song M, Wu CH, Liu H: **A comparison study on algorithms of detecting long forms for short forms in biomedical text**. *BMC bioinformatics* 2007, **8 Suppl 9**:S5.
106. Xu Y, Wang Z, Lei Y, Zhao Y, Xue Y: **MBA: a literature mining system for extracting biomedical abbreviations**. *BMC bioinformatics* 2009, **10**:14.
107. **PubMed stop words** [http://www.ncbi.nlm.nih.gov/books/bv.fcgi?highlight=stopwords&rid=helppubmed.table.pubmedhelp.T43]
108. Schuemie MJ, Jelier R, Kors JA: **Peregrine: Lightweight gene name normalization by dictionary lookup**. In: *Proceedings of the Biocreative 2 workshop: 2007*.
109. Muller HM, Kenny EE, Sternberg PW: **Textpresso: an ontology-based information retrieval and extraction system for biological literature**. *PLoS Biol* 2004, **2**(11):e309.
110. Hettne KM, Stierum RH, Schuemie MJ, Hendriksen PJ, Schijvenaars BJ, Mulligen EM, Kleinjans J, Kors JA: **A dictionary to identify small molecules and drugs in free text**. *Bioinformatics (Oxford, England)* 2009, **25**(22):2983-2991.
111. Kolarik C, Klinger R, Friedrich CM, Hofmann-Apitius M, Fluck J: **Chemical names: terminological resources and corpora annotation.** In: *Workshop on Building and evaluating resources for biomedical text mining (6th edition of the Language Resources and Evaluation Conference): 2008*.
112. Kemp N, Michael L: **Extraction of Information from the Text of Chemical Patents. 1. Identification of Specific Chemical Names.** *J Chem Inf Comput Sci* 1998, **38**:544-551.
113. Murray-Rust P, Mitchell JBO, Rzepa HS: **Chemistry in bioinformatics**. *BMC bioinformatics* 2005, **6**:141.
114. Murray-Rust P: **Chemistry for everyone**. *Nature* 2008, **451**(7179):648-651.
115. Williams AJ: **A perspective of publicly accessible/open-access chemistry databases**. *Drug Discov Today* 2008, **13**(11-12):495-501.
116. Williams AJ: **Internet-based tools for communication and collaboration in chemistry**. *Drug Discov Today* 2008, **13**(11-12):502-506.
117. Banville DL: **Mining chemical structural information from the drug literature**. *Drug Discov Today* 2006, **11**(1-2):35-42.
118. Corbett P, Copestake A: **Cascaded classifiers for confidence-based chemical named entity recognition**. *BMC bioinformatics* 2008, **9 Suppl 11**:S4.
119. Corbett P, Murray-Rust P: **High-Throughput Identification of Chemistry in Life Science Texts**. In: *CompLife 2006: 2006; Cambridge, UK*. Springer Berlin / Heidelberg: 107-118.
120. Degtyarenko K, de Matos P, Ennis M, Hastings J, Zbinden M, McNaught A, AlcÃintara R, Darsow M, Guedj Ml, Ashburner M: **ChEBI: a database and ontology for chemical entities of biological interest**. *Nucleic acids research* 2008, **36**(Database issue):D344-D350.

121. Rebholz-Schuhmann D, Arregui M, Gaudan S, Kirsch H, Jimeno A: **Text processing through Web services: calling Whatizit**. *Bioinformatics (Oxford, England)* 2008, **24**(2):296-298.
122. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M: **DrugBank: a knowledgebase for drugs, drug actions and drug targets**. *Nucleic acids research* 2008, **36**(Database issue):D901-906.
123. Kolarik C, Hofmann-Apitius M, Zimmermann M, Fluck J: **Identification of new drug classification terms in textual resources**. *Bioinformatics (Oxford, England)* 2007, **23**(13):i264-272.
124. Segura-Bedmar I, Martinez P, Segura-Bedmar M: **Drug name recognition and classification in biomedical texts. A case study outlining approaches underpinning automated systems**. *Drug Discov Today* 2008, **13**(17-18):816-823.
125. Agarwal P, Searls DB: **Literature mining in support of drug discovery**. *Brief Bioinform* 2008, **9**(6):479-492.
126. Singh SB, Hull RD, Fluder EM: **Text Influenced Molecular Indexing (TIMI): a literature database mining approach that handles text and chemistry**. *J Chem Inf Comput Sci* 2003, **43**(3):743-752.
127. Walker MJ, Hull RD, Singh SB: **CKB - the compound knowledge base: a text based chemical search system**. *J Chem Inf Comput Sci* 2002, **42**(6):1293-1295.
128. Zhu S, Okuno Y, Tsujimoto G, Mamitsuka H: **A probabilistic model for mining implicit 'chemical compound-gene' relations from literature**. *Bioinformatics (Oxford, England)* 2005, **21 Suppl 2**:ii245-ii251.
129. Weisgerber DW: **Chemical Abstracts Service Chemical Registry System: history, scope, and impacts.** *Journal of the American Society for Information Science* 1997, **48**:349-360.
130. Lipscomb CE: **Medical Subject Headings (MeSH)**. *Bull Med Libr Assoc* 2000, **88**(3):265-266.
131. Hanisch D, Fundel K, Mevissen H-T, Zimmer R, Fluck J: **ProMiner: rule-based protein and gene entity recognition**. *BMC bioinformatics* 2005, **6 Suppl 1**:S14.
132. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Edgar R, Federhen S *et al*: **Database resources of the National Center for Biotechnology Information**. *Nucleic acids research* 2008, **36**(Database issue):D13-21.
133. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T *et al*: **KEGG for linking genomes to life and the environment**. *Nucleic acids research* 2008, **36**(Database issue):D480-484.
134. Goto S, Okuno Y, Hattori M, Nishioka T, Kanehisa M: **LIGAND: database of chemical compounds and reactions in biological pathways**. *Nucleic acids research* 2002, **30**(1):402-404.
135. Wishart DS, Knox C, Guo AC, Eisner R, Young N, Gautam B, Hau DD, Psychogios N, Dong E, Bouatra S *et al*: **HMDB: a knowledgebase for the human metabolome**. *Nucleic acids research* 2009, **37**(Database issue):D603-610.
136. McCray AT, Bodenreider O, Malley JD, Browne AC: **Evaluating UMLS strings for natural language processing**. *Proc AMIA Symp* 2001:448-452.
137. Wilbur WJ, Hazard GF, Divita G, Mork JG, Aronson AR, Browne AC: **Analysis of biomedical text for chemical names: a comparison of three methods**. In: *Proc AMIA Symp: 1999*. 176-180.
138. Hettne KM, van Mulligen E, Schuemie M, Schijvenaars B, Kors JA: **Rewriting and suppressing UMLS terms for improved biomedical term identification**. *Journal of Biomedical Semantics* 2009, **Submitted**.
139. Schuemie M, Jelier R, Kors J: **Peregrine: Lightweight gene name normalization by dictionary lookup**. In: *Proceedings of the Biocreative 2 workshop: 2007; Madrid*.

140. Morgan AA, Lu Z, Wang X, Cohen AM, Fluck J, Ruch P, Divoli A, Fundel K, Leaman R, Hakenberg J *et al*: **Overview of BioCreative II gene normalization**. *Genome biology* 2008, **9 Suppl 2**:S3.
141. Schuemie MJ, Mons B, Weeber M, Kors JA: **Evaluation of techniques for increasing recall in a dictionary approach to gene and protein name identification**. *Journal of biomedical informatics* 2007, **40**(3):316-324.
142. Richard AM, Gold LS, Nicklaus MC: **Chemical structure indexing of toxicity data on the internet: moving toward a flat world**. *Curr Opin Drug Discov Devel* 2006, **9**(3):314-325.
143. McCray AT, Srinivasan S, Browne AC: **Lexical methods for managing variation in biomedical terminologies**. *Proc Annu Symp Comput Appl Med Care* 1994:235-239.
144. Corbett P, Batchelor C, Teufel S: **Annotation of Chemical Named Entities**. In: *BioNLP 2007: Biological, translational, and clinical language processing: 2007; Prague*. 57–64.
145. Bingjun S, Prasenjit M, Giles CL: **Mining, indexing, and searching for textual chemical molecule information on the web**. In: *Proceeding of the 17th international conference on World Wide Web; Beijing, China*. ACM 2008.
146. Bingjun S, Qingzhao T, Prasenjit M, Giles CL: **Extraction and search of chemical formulae in text documents on the web**. In: *Proceedings of the 16th international conference on World Wide Web; Banff, Alberta, Canada*. ACM 2007.
147. Chen JH, Linstead E, Swamidass SJ, Wang D, Baldi P: **ChemDB update--full-text search and virtual chemical space**. *Bioinformatics (Oxford, England)* 2007, **23**(17):2348-2351.
148. Schulz M, Uhlendorf J, Klipp E, Liebermeister W: **SBMLmerge, a system for combining biochemical network models**. *Genome Inform* 2006, **17**(1):62-71.
149. Smith L, Tanabe LK, Ando RJ, Kuo CJ, Chung IF, Hsu CN, Lin YS, Klinger R, Friedrich CM, Ganchev K *et al*: **Overview of BioCreative II gene mention recognition**. *Genome biology* 2008, **9 Suppl 2**:S2.
150. Wren J: **A scalable machine-learning approach to recognize chemical names within large text databases**. *BMC bioinformatics* 2006, **7 Suppl 2**:S3.
151. Yu H, Hripcsak G, Friedman C: **Mapping abbreviations to full forms in biomedical articles**. *J Am Med Inform Assoc* 2002, **9**(3):262-272.
152. Yu H, Kim W, Hatzivassiloglou V, Wilbur WJ: **Using MEDLINE as a knowledge source for disambiguating abbreviations and acronyms in full-text biomedical journal articles**. *Journal of biomedical informatics* 2007, **40**(2):150-159.
153. **ChemSpider** [http://www.chemspider.com/]
154. **ChemMantis** [http://www.chemspider.com/blog/welcome-chemmantis-to-chemzoo-and-a-call-for-contributions-to-the-community.html]
155. Bretcher J: **Name=Struct: A Practical Approach to the Sorry State of Real-Life Chemical Nomenclature**. *J Chem Inf Comput Sci* 1999, **39**:943-950.
156. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M: **From genomics to chemical genomics: new developments in KEGG**. *Nucleic acids research* 2006, **34**(Database issue):D354-D357.
157. **ChemIDplus Fact Sheet** [http://www.nlm.nih.gov/pubs/factsheets/chemidplusfs.html]
158. **CAS REGISTRY and CAS Registry Numbers** [http://www.cas.org/expertise/cascontent/registry/regsys.html]
159. **Beilstein database** [http://en.wikipedia.org/wiki/Beilstein_database]
160. **EINECS numbers** [http://en.wikipedia.org/wiki/EINECS]
161. Kanehisa M, Goto S: **KEGG: Kyoto Encyclopedia of Genes and Genomes**. *Nucleic acids research* 2000, **28**(1):27-30.
162. Kandasamy K, Mohan SS, Raju R, Keerthikumar S, Kumar GSS, Venugopal AK, Telikicherla D, Navarro JD, Mathivanan S, Pecquet C *et al*: **NetPath: a public**

resource of curated signal transduction pathways. *Genome biology*, **11**(1):R3.
163. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nature genetics* 2000, **25**(1):25-29.
164. Huang da W, Sherman BT, Lempicki RA: **Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists**. *Nucleic acids research* 2009, **37**(1):1-13.
165. Khatri P, Draghici S: **Ontological analysis of gene expression data: current tools, limitations, and open problems**. *Bioinformatics (Oxford, England)* 2005, **21**(18):3587-3595.
166. Barry WT, Nobel AB, Wright FA: **Significance analysis of functional categories in gene expression studies: a structured permutation approach**. *Bioinformatics (Oxford, England)* 2005, **21**(9):1943-1949.
167. Pavlidis P, Qin J, Arango V, Mann JJ, Sibille E: **Using the gene ontology for microarray data mining: a comparison of methods and application to age effects in human prefrontal cortex**. *Neurochem Res* 2004, **29**(6):1213-1222.
168. Mootha VK, Lindgren CM, Eriksson K-F, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, RidderstrÃ¥le M, Laurila E *et al*: **PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes**. *Nature genetics* 2003, **34**(3):267-273.
169. Goeman JJ, van de Geer SA, van Houwelingen JC: **Testing against a high-dimensional alternative**. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 2006, **68**:477-493.
170. Hummel M, Meister R, Mansmann U: **GlobalANCOVA: exploration and assessment of gene group effects**. *Bioinformatics (Oxford, England)* 2008, **24**(1):78-85.
171. Goeman JJ, Bühlmann P: **Analyzing gene expression data in terms of gene sets: methodological issues**. *Bioinformatics (Oxford, England)* 2007, **23**(8):980-987.
172. Barrell D, Dimmer E, Huntley RP, Binns D, O'Donovan C, Apweiler R: **The GOA database in 2009--an integrated Gene Ontology Annotation resource**. *Nucleic acids research* 2009, **37**(Database issue):D396-D403.
173. Khatri P, Done B, Rao A, Done A, Draghici S: **A semantic analysis of the annotations of the human genome**. *Bioinformatics (Oxford, England)* 2005, **21**(16):3416-3421.
174. He X, Sarma MS, Ling X, Chee B, Zhai C, Schatz B: **Identifying overrepresented concepts in gene lists from literature: a statistical approach based on Poisson mixture model**. *BMC bioinformatics* 2010, **11**:272.
175. Blaschke C, Oliveros JC, Valencia A: **Mining functional information associated with expression arrays**. *Functional & integrative genomics* 2001, **1**(4):256-268.
176. Homayouni R, Heinrich K, Wei L, Berry MW: **Gene clustering by latent semantic indexing of MEDLINE abstracts**. *Bioinformatics (Oxford, England)* 2005, **21**(1):104-115.
177. Raychaudhuri S, Altman RB: **A literature-based method for assessing the functional coherence of a gene group**. *Bioinformatics (Oxford, England)* 2003, **19**(3):396-401.
178. Shatkay H, Feldman R: **Mining the biomedical literature in the genomic era: an overview**. *J Comput Biol* 2003, **10**(6):821-855.
179. Febbo PG, Mulligan MG, Slonina DA, Stegmaier K, Di Vizio D, Martinez PR, Loda M, Taylor SC: **Literature Lab: a method of automated literature interrogation to infer biology from microarray analysis**. *BMC genomics* 2007, **8**:461.
180. Huang ZX, Tian HY, Hu ZF, Zhou YB, Zhao J, Yao KT: **GenCLiP: a software program for clustering gene lists by literature profiling and constructing**

gene co-occurrence networks related to custom keywords. *BMC bioinformatics* 2008, **9**:308.
181. Rubinstein R, Simon I: **MILANO--custom annotation of microarray results using automatic literature searches**. *BMC bioinformatics* 2005, **6**:12.
182. Tjioe E, Berry MW, Homayouni R: **Discovering gene functional relationships using FAUN (Feature Annotation Using Nonnegative matrix factorization)**. *BMC bioinformatics* 2010, **11 Suppl 6**:S14.
183. Burkart MF, Wren JD, Herschkowitz JI, Perou CM, Garner HR: **Clustering microarray-derived gene lists through implicit literature relationships**. *Bioinformatics (Oxford, England)* 2007, **23**(15):1995-2003.
184. Leach SM, Tipney H, Feng W, Baumgartner WA, Kasliwal P, Schuyler RP, Williams T, Spritz RA, Hunter L: **Biomedical discovery acceleration, with applications to craniofacial development**. *PLoS computational biology* 2009, **5**(3):e1000215.
185. Kueffner R, Fundel K, Zimmer R: **Expert knowledge without the expert: integrated analysis of gene expression and literature to derive active functional contexts**. *Bioinformatics (Oxford, England)* 2005, **21 Suppl 2**:ii259-ii267.
186. Jelier R, Schuemie MJ, Veldhoven A, Dorssers LCJ, Jenster G, Kors JA: **Anni 2.0: a multipurpose text-mining tool for the life sciences**. *Genome biology* 2008, **9**(6):R96.
187. Srinivasan P: **Text mining: generating hypotheses from MEDLINE**. *JASIST* 2004, **55**:396-413.
188. Bakker ML, Boukens BJ, Mommersteeg MTM, Brons JF, Wakker V, Moorman AFM, Christoffels VM: **Transcription factor Tbx3 is required for the specification of the atrioventricular conduction system**. *Circ Res* 2008, **102**(11):1340-1349.
189. Hoogaars WMH, Engel A, Brons JF, Verkerk AO, de Lange FJ, Wong LYE, Bakker ML, Clout DE, Wakker V, Barnett P *et al*: **Tbx3 controls the sinoatrial node gene program and imposes pacemaker function on the atria**. *Genes Dev* 2007, **21**(9):1098-1112.
190. Horsthuis T, Buermans HPJ, Brons JF, Verkerk AO, Bakker ML, Wakker V, Clout DEW, Moorman AFM, t Hoen PAC, Christoffels VM: **Gene expression profiling of the forming atrioventricular node using a novel tbx3-based node-specific transgenic reporter**. *Circ Res* 2009, **105**(1):61-69.
191. Boukens BJD, Christoffels VM, Coronel R, Moorman AFM: **Developmental basis for electrophysiological heterogeneity in the ventricular and outflow tract myocardium as a substrate for life-threatening ventricular arrhythmias**. *Circ Res* 2009, **104**(1):19-31.
192. Lupoglazoff JM, Cheav T, Baroudi G, Berthet M, Denjoy I, Cauchemez B, Extramiana F, Chahine M, Guicheney P: **Homozygous SCN5A mutation in long-QT syndrome with functional two-to-one atrioventricular block**. *Circ Res* 2001, **89**(2):E16-E21.
193. Simon AM, Goodenough DA, Paul DL: **Mice lacking connexin40 have cardiac conduction abnormalities characteristic of atrioventricular block and bundle branch block**. *Curr Biol* 1998, **8**(5):295-298.
194. Kroon JTM, Wei W, Simon WJ, Slabas AR: **Identification and functional expression of a type 2 acyl-CoA:diacylglycerol acyltransferase (DGAT2) in developing castor bean seeds which has high homology to the major triglyceride biosynthetic enzyme of fungi and animals**. *Phytochemistry* 2006, **67**(23):2541-2549.
195. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AAM, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ *et al*: **A gene-expression signature as a predictor of survival in breast cancer**. *N Engl J Med* 2002, **347**(25):1999-2009.
196. Goeman JJ, Mansmann U: **Multiple testing on the directed acyclic graph of gene ontology**. *Bioinformatics (Oxford, England)* 2008, **24**(4):537-544.

197. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT *et al*: **Gene expression profiling predicts clinical outcome of breast cancer**. *Nature* 2002, **415**(6871):530-536.
198. Dong X, Liu F, Sun L, Liu M, Li D, Su D, Zhu Z, Dong J-T, Fu L, Zhou J: **Oncogenic function of microtubule end-binding protein 1 in breast cancer**. *J Pathol* 2010.
199. Morris PG, Fornier MN: **Ixabepilone and other epothilones: microtubule-targeting agents for metastatic breast cancer**. *Clin Adv Hematol Oncol* 2009, **7**(2):115-122.
200. Chen C-C, Chang T-W, Chen F-M, Hou M-F, Hung S-Y, Chong I-W, Lee S-C, Zhou T-H, Lin S-R: **Combination of multiple mRNA markers (PTTG1, Survivin, UbcH10 and TK1) in the diagnosis of Taiwanese patients with breast cancer by membrane array**. *Oncology* 2006, **70**(6):438-446.
201. Varambally S, Yu J, Laxman B, Rhodes DR, Mehra R, Tomlins SA, Shah RB, Chandran U, Monzon FA, Becich MJ *et al*: **Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression**. *Cancer Cell* 2005, **8**(5):393-406.
202. Jung Y, Park J, Bang Y-J, Kim T-Y: **Gene silencing of TSPYL5 mediated by aberrant promoter methylation in gastric cancers**. *Lab Invest* 2008, **88**(2):153-160.
203. Vachani A, Nebozhyn M, Singhal S, Alila L, Wakeam E, Muschel R, Powell CA, Gaffney P, Singh B, Brose MS *et al*: **A 10-gene classifier for distinguishing head and neck squamous cell carcinoma and lung squamous cell carcinoma**. *Clin Cancer Res* 2007, **13**(10):2905-2915.
204. de Vogel-van den Bosch HM, Bünger M, de Groot PJ, Bosch-Vermeulen H, Hooiveld GJEJ, Müller M: **PPARalpha-mediated effects of dietary lipids on intestinal barrier gene expression**. *BMC genomics* 2008, **9**:231.
205. Capdevila JH, Falck JR, Imig JD: **Roles of the cytochrome P450 arachidonic acid monooxygenases in the control of systemic blood pressure and experimental hypertension**. *Kidney Int* 2007, **72**(6):683-689.
206. Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S *et al*: **Database resources of the National Center for Biotechnology Information**. *Nucleic acids research* 2009, **37**(Database issue):D5-15.
207. UniProt Consortium: **The Universal Protein Resource (UniProt) in 2010**. *Nucleic acids research* 2010, **38**(Database issue):D142-D148.
208. Bruford EA, Lush MJ, Wright MW, Sneddon TP, Povey S, Birney E: **The HGNC Database in 2008: a resource for the human genome**. *Nucleic acids research* 2008, **36**(Database issue):D445-D448.
209. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S *et al*: **Database resources of the National Center for Biotechnology Information**. *Nucleic acids research* 2007, **35**(Database issue):D5-12.
210. Hettne KM, van Mulligen EM, Schuemie MJ, Schijvenaars BJA, Kors JA: **Rewriting and suppressing UMLS terms for improved biomedical term identification**. *Journal of Biomedical Semantics* 2010, **1:5**.
211. Goodman LA, Kruskal WH: **Measures of association for cross classifications**: Springer-Verlag, New York; 1979.
212. Holm S: **A simple sequentially rejective multiple test procedure**. *Scandinavian J Stat* 1979, **6**:65-70.
213. Huber W, von Heydebreck A, Sültmann H, Poustka A, Vingron M: **Variance stabilization applied to microarray data calibration and to the quantification of differential expression**. *Bioinformatics (Oxford, England)* 2002, **18 Suppl 1**:S96-104.
214. van Houwelingen HC, Bruinsma T, Hart AAM, Veer LJVt, Wessels LFA: **Cross-validated Cox regression on microarray gene expression data**. *Stat Med* 2006, **25**(18):3201-3216.

215. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data**. *Biostatistics* 2003, **4**(2):249-264.
216. Chen L, Liu H, Friedman C: **Gene name ambiguity of eukaryotic nomenclatures**. *Bioinformatics (Oxford, England)* 2005, **21**(2):248-256.
217. Smalley JL, Gant TW, Zhang SD: **Application of connectivity mapping in predictive toxicology based on gene-expression similarity**. *Toxicology* 2010, **268**(3):143-146.
218. Bussemaker HJ, Ward LD, Boorsma A: **Dissecting complex transcriptional responses using pathway-level scores based on prior information**. *BMC bioinformatics* 2007, **8**:S6.
219. Davis AP, King BL, Mockus S, Murphy CG, Saraceni-Richards C, Rosenstein M, Wiegers T, Mattingly CJ: **The Comparative Toxicogenomics Database: update 2011**. *Nucleic acids research* 2011, **39**(Database issue):D1067-72.
220. Jelier R, Goeman JJ, Hettne KM, Schuemie MJ, den Dunnen JT, 't Hoen PAC: **Literature-aided interpretation of gene expression data**. *Brief Bioinform* 2011, **12**(5):518-29.
221. Knudsen TB, Martin MT, Kavlock RJ, Judson RS, Dix DJ, Singh AV: **Profiling the activity of environmental chemicals in prenatal developmental toxicity studies using the U.S. EPA's ToxRefDB**. *Reproductive toxicology (Elmsford, NY* 2009, **28**(2):209-219.
222. Boorsma A, Foat BC, Vis D, Klis F, Bussemaker HJ: **T-profiler: scoring the activity of predefined groups of genes using gene expression data**. *Nucleic acids research* 2005, **33**(Web Server issue):W592-595.
223. Nogales-Cadenas R, Carmona-Saez P, Vazquez M, Vicente C, Yang X, Tirado F, Carazo JM, Pascual-Montano A: **GeneCodis: interpreting gene lists through enrichment analysis and integration of diverse biological information**. *Nucleic acids research* 2009, **37**(Web Server issue):W317-322.
224. Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing**. *Journal of the Royal Statistical Society Series B (Methodological)* 1995, **57**(1):289-300.
225. Naciff JM, Khambatta ZS, Reichling TD, Carr GJ, Tiesman JP, Singleton DW, Khan SA, Daston GP: **The genomic response of Ishikawa cells to bisphenol A exposure is dose- and time-dependent**. *Toxicology* 2010, **270**(2-3):137-149.
226. Lin CY, Vega VB, Thomsen JS, Zhang T, Kong SL, Xie M, Chiu KP, Lipovich L, Barnett DH, Stossi F *et al*: **Whole-genome cartography of estrogen receptor alpha binding sites**. *PLoS genetics* 2007, **3**(6):e87.
227. Selvaraj V, Bunick D, Finnigan-Bunick C, Johnson RW, Wang H, Liu L, Cooke PS: **Gene expression profiling of 17beta-estradiol and genistein effects on mouse thymus**. *Toxicol Sci* 2005, **87**(1):97-112.
228. Tijet N, Boutros PC, Moffat ID, Okey AB, Tuomisto J, Pohjanvirta R: **Aryl hydrocarbon receptor regulates distinct dioxin-dependent and dioxin-independent gene batteries**. *Molecular pharmacology* 2006, **69**(1):140-153.
229. Bosse Y, Maghni K, Hudson TJ: **1alpha,25-dihydroxy-vitamin D3 stimulation of bronchial smooth muscle cells induces autocrine, contractility, and remodeling processes**. *Physiol Genomics* 2007, **29**(2):161-168.
230. Li Z, Stonehuerner J, Devlin RB, Huang YC: **Discrimination of vanadium from zinc using gene profiling in human bronchial epithelial cells**. *Environmental health perspectives* 2005, **113**(12):1747-1754.
231. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP: **GenePattern 2.0**. *Nature genetics* 2006, **38**(5):500-501.
232. van Dartel DA, Pennings JL, de la Fonteyne LJ, van Herwijnen MH, van Delft JH, van Schooten FJ, Piersma AH: **Monitoring developmental toxicity in the embryonic stem cell test using differential gene expression of differentiation-related genes**. *Toxicol Sci* 2010, **116**(1):130-139.
233. de Jong E, Barenys M, Hermsen SA, Verhoef A, Ossendorp BC, Bessems JG, Piersma AH: **Comparison of the mouse Embryonic Stem cell Test, the rat**

Whole Embryo Culture and the Zebrafish Embryotoxicity Test as alternative methods for developmental toxicity testing of six 1,2,4-triazoles. *Toxicol Appl Pharmacol* 2011, **253**(2):103-111.
234. Vanden Bossche H, Marichal P, Gorrens J, Coene MC: **Biochemical basis for the activity and selectivity of oral antifungal drugs**. *British journal of clinical practice* 1990, **71**:41-46.
235. Pennings JL, van Dartel DA, Robinson JF, Pronk TE, Piersma AH: **Gene set assembly for quantitative prediction of developmental toxicity in the embryonic stem cell test**. *Toxicology* 2011, **284**(1-3):63-71.
236. Kors JA, Schuemie MJ, Schijvenaars BJA, Weeber M, Mons B: **Combination of genetic databases for improving identification of genes and proteins in text**. In: *BioLINK: 2005; Detroit*.
237. **http://www.ihop-net.org/UniPub/iHOP/**
238. **http://www.chemspider.com/blog/wp-content/uploads/2010/08/CSDocs_GuideToDatabaseCurationAndAnnotation.pdf**
239. Roque FS, Jensen PB, Schmock H, Dalgaard M, Andreatta M, Hansen T, Soeby K, Bredkjaer S, Juul A, Werge T *et al*: **Using electronic patient records to discover disease correlations and stratify patient cohorts**. *PLoS computational biology* 2011, **7**(8):e1002141.
240. Hersh WR, Campbell EH, Evans DA, Brownlow ND: **Empirical, automated vocabulary discovery using large text corpora and advanced natural language processing tools**. *Proc AMIA Annu Fall Symp* 1996:159-163.
241. Williams AJ, Ekins S: **A quality alert and call for improved curation of public chemistry databases**. *Drug Discov Today* 2011, **16**(17-18):747-750.
242. Williams AJ, Ekins S, Tkachenko V: **Towards a gold standard: regarding quality in public domain chemistry databases and approaches to improving the situation**. *Drug Discov Today* 2012.
243. Huang R, Southall N, Wang Y, Yasgar A, Shinn P, Jadhav A, Nguyen DT, Austin CP: **The NCGC pharmaceutical collection: a comprehensive resource of clinically approved drugs enabling repurposing and chemical genomics**. *Science translational medicine* 2011, **3**(80):80ps16.
244. Stobbe MD, Houten SM, Jansen GA, van Kampen AH, Moerland PD: **Critical assessment of human metabolic pathway databases: a stepping stone for future integration**. *BMC systems biology* 2011, **5**:165.
245. Kumar A, Suthers PF, Maranas CD: **MetRxn: a knowledgebase of metabolites and reactions spanning metabolic models and databases**. *BMC bioinformatics* 2012, **13**(1):6.
246. **http://www.openphacts.org/**
247. Gruning BA, Senger C, Erxleben A, Flemming S, Gunther S: **Compounds In Literature (CIL): screening for compounds and relatives in PubMed**. *Bioinformatics (Oxford, England)* 2011, **27**(9):1341-1342.
248. Nobata C, Dobson PD, Iqbal SA, Mendes P, Tsujii J, Kell DB, Ananiadou S: **Mining metabolites: extracting the yeast metabolome from the literature**. *Metabolomics* 2011, **7**(1):94-101.
249. Jessop DM, Adams SE, Willighagen EL, Hawizy L, Murray-Rust P: **OSCAR4: a flexible architecture for chemical text-mining**. *Journal of cheminformatics* 2011, **3**(1):41.
250. **http://www.biosemantics.org/anni**
251. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ *et al*: **The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration**. *Nat Biotechnol* 2007, **25**(11):1251-1255.
252. **http://bioportal.bioontology.org/ontologies**
253. **http://www.wikipedia.org/**
254. **http://www.conceptwiki.org/**

255. Kelder T, van Iersel MP, Hanspers K, Kutmon M, Conklin BR, Evelo CT, Pico AR: **WikiPathways: building research communities on biological pathways**. *Nucleic Acids Res* 2012, **40**(Database issue):D1301-7.
256. Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P: **Drug target identification using side-effect similarity**. *Science (New York, NY* 2008, **321**(5886):263-266.
257. **http://www.pharmgkb.org/**
258. **http://www.genecards.org/**
259. Rhodes J, Boyer S, Kreulen J, Chen Y, Ordonez P: **Mining patents using molecular similarity search**. *Pacific Symposium on Biocomputing* 2007:304-315.
260. Sayle R, Xie PH, Muresan S: **Improved chemical text mining of patents with infinite dictionaries and automatic spelling correction**. *J Chem Inf Model* 2012, **52**(1):51-62.
261. **http://www.ebi.ac.uk/Rebholz-srv/CALBC/corpora/corpora.html**
262. Cano C, Monaghan T, Blanco A, Wall DP, Peshkin L: **Collaborative text-annotation resource for disease-centered relation extraction from biomedical text**. *Journal of biomedical informatics* 2009, **42**(5):967-977.
263. Mons B, Ashburner M, Chichester C, van Mulligen E, Weeber M, den Dunnen J, van Ommen GJ, Musen M, Cockerill M, Hermjakob H *et al*: **Calling on a million minds for community annotation in WikiProteins**. *Genome biology* 2008, **9**(5):R89.
264. **http://genome2.ugr.es/bionotate2/**
265. **http://www.calbc.eu/**
266. Rebholz-Schuhmann D, Yepes AJ, Li C, Kafkas S, Lewin I, Kang N, Corbett P, Milward D, Buyko E, Beisswanger E *et al*: **Assessment of NER solutions against the first and second CALBC Silver Standard Corpus**. *J Biomed Semantics* 2011, **2 Suppl 5**:S11.
267. Kang N, van Mulligen EM, Kors JA: **Training text chunkers on a silver standard corpus: can silver replace gold?** *BMC bioinformatics* 2012, **13**:17.
268. Jimeno-Yepes AJ, Aronson AR: **Knowledge-based biomedical word sense disambiguation: comparison of approaches**. *BMC bioinformatics* 2010, **11**:569.
269. Jimeno-Yepes A, McInnes BT, Aronson AR: **Collocation analysis for UMLS knowledge-based word sense disambiguation**. *BMC bioinformatics* 2011, **12 Suppl 3**:S4.
270. Akkaya C, Conrad A, Wiebe J, Mihalcea R: **Amazon Mechanical Turk for Subjectivity Word Sense Disambiguation**. In: *NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk: 2012; Los Angeles, USA*. 195-203.
271. **http://semanticweb.org/**
272. Hull D, Wolstencroft K, Stevens R, Goble C, Pocock MR, Li P, Oinn T: **Taverna: a tool for building and running workflows of services**. *Nucleic Acids Res* 2006, **34**(Web Server issue):W729-32.
273. Bhagat J, Tanoh F, Nzuobontane E, Laurent T, Orlowski J, Roos M, Wolstencroft K, Aleksejevs S, Stevens R, Pettifer S, Lopez R, Goble CA: **BioCatalogue: a universal catalogue of web services for the life sciences**. *Nucleic Acids Res* 2010, **38**(Web Server issue):W689-94.
274. Goble CA, Bhagat J, Aleksejevs S, Cruickshank D, Michaelides D, Newman D, Borkum M, Bechhofer S, Roos M, Li P, De Roure D: **myExperiment: a repository and social network for the sharing of bioinformatics workflows**. *Nucleic Acids Res* 2010, **38**(Web Server issue):W677-82.
275. Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette J: **Bio2RDF: towards a mashup to build bioinformatics knowledge systems**. *Journal of biomedical informatics* 2008, **41**(5):706-716.
276. Samwald M, Jentzsch A, Bouton C, Kallesoe CS, Willighagen E, Hajagos J, Marshall MS, Prud'hommeaux E, Hassenzadeh O, Pichler E *et al*: **Linked open drug data**

for pharmaceutical research and development. *Journal of cheminformatics* 2011, **3**(1):19.

277. Chen B, Dong X, Jiao D, Wang H, Zhu Q, Ding Y, Wild DJ: **Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data**. *BMC bioinformatics* 2010, **11**:255.

Curriculum Vitae

Kristina Hettne was born in Råda, Sweden, on March 2, 1978. She followed The Health Care Programme at the Swedish Upper Secondary School and graduated in 1997, after which she embarked on an around-the-world trip. Coming back to Sweden in 1998, she followed courses in natural science subjects provided by the Municipal Adult Education in Sweden, preparing her for a university study combining biology and computer science. In 1999 she started her study in Bioinformatics at the University of Skövde, Sweden. In 2003 she obtained her M.Sc. degree in Computer Science, focusing on Bioinformatics with the thesis "Using nuclear receptor interactions as biomarkers for metabolic syndrome". The work was carried out at the Bioinformatics department at AstraZeneca R&D, Mölndal, Sweden. In 2004 she started working as a Scientist at the Safety Assessment department at AstraZeneca R&D Mölndal, Sweden. During her time at AstraZeneca Kristina was involved in research projects aimed at unraveling parts of the biology behind two diseases, the Complex Regional Pain Syndrome I (CRPS-I) and periodontitis, and in research surrounding nuclear receptor pathways and their relevance for the safety of drugs. In August 2006 she started as a PhD student in the Biosemantics group at the Erasmus MC, Rotterdam, on a project to develop class prediction tools for toxicological classification, and to identify molecular and genetic pathways linking expression profiles and specific toxic phenotypes, of which the results are reported in this thesis. The project was a collaboration between the department of Toxicogenomics at Maastricht University and the department of Medical Informatics at Erasmus MC. In June 2011 she joined the Biosemantics group at the Human Genetics department at the Leiden University Medical Center as a postdoctoral researcher, combining her interest in health care, biology, and computer science.

Publication list

Cases M, Garcia-Serna R, **Hettne K**, Weeber M, van der Lei J, Boyer S, Mestres J: **Chemical and biological profiling of an annotated compound library directed to the nuclear receptor family**. *Current topics in medicinal chemistry* 2005, **5**(8):763-772.

Hettne K, Cases M, Boyer S, Mestres J: **Connecting small molecules to nuclear receptor pathways**. *Current topics in medicinal chemistry* 2007, **7**(15):1530-1536.

Hettne KM, de Mos M, de Bruijn AG, Weeber M, Boyer S, van Mulligen EM, Cases M, Mestres J, van der Lei J: **Applied information retrieval and multidisciplinary research: new mechanistic hypotheses in complex regional pain syndrome**. *Journal of biomedical discovery and collaboration* 2007, **2**:2.

van Mulligen EM, Cases M, **Hettne K**, Molero E, Weeber M, Robertson KA, Oliva B, de la Calle G, Maojo V: **Training multidisciplinary biomedical informatics students: three years of experience**. *Journal of the American Medical Informatics Association* 2008, **15**(2):246-254.

Hettne KM, Stierum RH, Schuemie MJ, Hendriksen PJ, Schijvenaars BJ, Mulligen EM, Kleinjans J, Kors JA: **A dictionary to identify small molecules and drugs in free text**. *Bioinformatics (Oxford, England)* 2009, **25**(22):2983-2991.

Hettne KM, van Mulligen EM, Schuemie MJ, Schijvenaars BJ, Kors JA: **Rewriting and suppressing UMLS terms for improved biomedical term identification**. *Journal of biomedical semantics* 2010, **1**(1):5.

Hettne KM, Weeber M, Laine ML, ten Cate H, Boyer S, Kors JA, Loos BG: **Automatic mining of the literature to generate new hypotheses for the possible link between periodontitis and atherosclerosis: lipopolysaccharide as a case study**. *Journal of clinical periodontology* 2007, **34**(12):1016-1024.

Hettne KM, Williams AJ, van Mulligen EM, Kleinjans J, Tkachenko V, Kors JA: **Automatic vs. manual curation of a multi-source chemical dictionary: the impact on text mining**. *Journal of cheminformatics* 2010, **2**(1):3.

Jelier R, Goeman JJ, **Hettne KM**, Schuemie MJ, den Dunnen JT, t Hoen PA: **Literature-aided interpretation of gene expression data with the weighted global test**. *Briefings in bioinformatics* 2011.

Hettne KM, Boorsma A, van Dartel DAM, Goeman JJ de Jong E, Piersma AH, Stierum RH, Kleinjans JC, Kors JA: **Next-generation text-mining mediated chemical-response specific gene sets for interpretation of gene expression data**. *Submitted*.