

CASSANDRA: DRUG GENE ASSOCIATION PREDICTION VIA
TEXT MINING AND ONTOLOGIES

Dissertation

zur Erlangung des akademischen Grades
Doctor of Philosophy (PhD)

vorgelegt an der
Technischen Universität Dresden
Fakultät Informatik

eingereicht von
Dipl.-Ing. MARIA KISSA

geboren am 04 March 1986 in Athen, Griechenland

Gutachter: Prof. Dr. Michael Schroeder
Technische Universität Dresden
Nigam H. Shah, MBBS, PhD
Stanford University

Tag der Verteidigung: 20. 01. 2015

Dresden, im Januar 2015

Declaration of Authorship

I, Maria KISSA, declare that this thesis titled, 'CASSANDRA: drug gene association prediction via text mining and ontologies' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

List of Publications

1. Maria Kissa, George Tsatsaronis and Michael Schroeder. **Prediction of drug-gene associations via ontological profile similarity with application to drug repositioning.** In *Methods*, 2014.

Contributions: MK and MS conceived and designed the respective methodology. MK implemented the algorithm, performed the experiments, analyzed the data and wrote the major part of this manuscript.

This paper forms the basis of this thesis

2. Maria Kissa, Michael Schroeder and George Tsatsaronis. **Towards an Integrated Compound to Compound Relatedness Measure.** In *Proceedings of the ISMB BioLINK SIG: Linking Literature, Information and Knowledge for Biology.*, 2013.

Contributions: MK and GT conceived and designed the respective methodology. They implemented the algorithm, performed the experiments, analyzed the data and wrote the major part of this manuscript.

3. George Tsatsaronis, Alina Petrova, Maria Kissa, Yue Ma, Felix Distel and Franz Baader and Michael Schroeder. **Learning Formal Definitions for Biomedical Concepts.** In *Proceedings of the 10th OWL: Experiences and Directions Workshop (OWLED 2013)*. 2013.

Contributions: MK analyzed part of the data and contributed to the writing of the manuscript.

“It is the mark of an educated mind to be able to entertain a thought without accepting it. ”

Aristotle

Abstract

The amount of biomedical literature has been increasing rapidly during the last decade. Text mining techniques can harness this large-scale data, shed light onto complex drug mechanisms, and extract relation information that can support computational polypharmacology. In this work, we introduce CASSANDRA, a fully corpus-based and unsupervised algorithm which uses the *MEDLINE* indexed titles and abstracts to infer drug gene associations and assist drug repositioning. CASSANDRA measures the *Pointwise Mutual Information (PMI)* between biomedical terms derived from *Gene Ontology (GO)* and *Medical Subject Headings (MeSH)*. Based on the *PMI* scores, drug and gene profiles are generated and candidate drug gene associations are inferred when computing the relatedness of their profiles. Results show that an *Area Under the Curve (AUC)* of up to 0.88 can be achieved. The algorithm can successfully identify direct drug gene associations with high precision and prioritize them over indirect drug gene associations. Validation shows that the statistically derived profiles from literature perform as good as (and at times better than) the manually curated profiles. In addition, we examine CASSANDRA's potential towards drug repositioning. For all *FDA*-approved drugs repositioned over the last 5 years, we generate profiles from publications before 2009 and show that the new indications rank high in these profiles. In summary, co-occurrence based profiles derived from the biomedical literature can accurately predict drug gene associations and provide insights onto potential repositioning cases.

Acknowledgements

I would like to express my deepest thanks and gratitude to my supervisor Professor Michael Schroeder for giving me the opportunity to work in such an amazing topic and in such a wonderful and international environment. His constant presence, guidance and encouragement supported me greatly throughout this challenging academic journey.

My deepest thanks also go to Dr. George Tsatsaronis, our multi-tasking postdoc, who despite the vast volume of responsibilities, he always found his way into transforming my questions into fruitful and stimulating scientific discussions.

Apart from Michael and George, I would like to specifically thank my colleagues and friends Alina, Daniel and Norhan with whom I shared both my scientific and general concerns. Our discussions and time together had always been a pleasant and entertaining break during these demanding *PhD* years.

Of course, I would like to thank all the past and present members of the Schroeder group. This fantastic working and collaborative environment couldn't be nothing else but an irreplaceable factor towards the successful realization of the following work.

No words are enough to express my love and gratitude to my husband Nikos. Patient, confident, encouraging, motivating, energetic, supporting, always present, always there for me, always believing in me; my personal and ultimate life coach.

Last but not least, come my friends and family. Thank you guys. For being there for me, for accepting me, for laughing and crying with me, for opposing me, for listening to me, for advising me, I thank you from heart.

Maryllia

Contents

Declaration of Authorship	i
Abstract	iv
Acknowledgements	v
Contents	vi
List of Tables	viii
List of Figures	ix
1 Motivation	1
1.1 Problem definition and proposed approach	3
1.2 Thesis Outline	4
2 Background	7
2.1 Drug Repositioning	8
2.1.1 Computational polypharmacology	10
2.2 Literature Based Discovery in the Biomedical Domain	13
2.3 Biomedical Ontologies	18
2.4 Mining biomedical text	21
2.4.1 Annotation of Biomedical Terms	22
2.4.2 Relation Extraction	23
2.5 Drug gene association prediction	25
3 Materials and Methods	31
3.1 Materials and Methods - Implementation	32
3.1.1 Resources - Text and terms	32
3.1.2 Recognition of terms in text	34

3.1.3	Profile generation for drugs and genes	36
3.1.4	Computation of semantic relatedness between the profiles	40
3.2	Materials and Methods - Evaluation	42
3.2.1	Evaluation datasets	42
3.2.1.1	Evaluation datasets for drug gene association prediction	42
3.2.1.2	Evaluation sets for drug repositioning	46
3.2.2	Alternative semantic similarity metrics	47
4	Results	49
4.1	<i>MEDLINE</i> statistics	50
4.2	Evaluation datasets statistics	53
4.3	Drug gene association prediction	57
4.3.1	Performance evaluation - Ontological profiles	59
4.3.1.1	Co-occurrence based profiles	60
4.3.1.2	Manually curated profiles	65
4.3.1.3	Co-occurrence based vs. manually curated profiles	68
4.3.2	Performance evaluation - Semantic similarity	73
4.4	Drug repositioning	82
4.5	Case studies	83
4.5.1	Cathine- <i>GHRL</i> association	83
4.5.2	Fenethylamine- <i>ApoE</i> association	86
4.5.3	Milnacipran- <i>SLC6A4</i> association	89
5	Discussion	93
5.1	Drug gene association prediction	95
5.2	Exploring the literature	98
5.2.1	Focusing on abstracts and titles	99
5.2.2	Using co-occurrence	100
5.3	The role of ontologies	103
5.3.1	Estimating the semantic similarity	105
6	Conclusion	107
6.1	Contribution	108
6.2	Future directions	110
7	Supplementary Material	113
	Bibliography	117

List of Tables

2.1	Tools for <i>Literature Based Discovery</i>	17
2.2	Co-occurrence vs. Patterns	25
2.3	Drug gene association prediction methodologies	29
3.1	Drug representation with regular expressions - Examples . .	36
4.1	Statistics of term annotations in <i>MEDLINE</i>	50
4.2	Co-occurrences for drugs, genes and ontological terms in <i>MEDLINE</i>	51
4.3	Ontological terms co-occurrences in <i>MEDLINE</i>	51
4.4	Evaluation dataset statistics	54
4.5	<i>Arithmetic mean</i> of statistically significant concepts in profiles	62
4.6	<i>AUC</i> values when including non-profiled drugs and genes . .	64
4.7	<i>AUC</i> values for different type of ontological profiles	71
4.8	Overall performance scores - Datasets for co-occurrence based profiles	78
4.9	Overall performance scores - Datasets for manually curated profiles	79
4.10	<i>nPMI</i> computations - Examples	81
4.11	Examples of drug repositioning potential	84
4.12	<i>Cathine-GHRL</i> textual findings	87
4.13	<i>Fenethylamine-ApoE</i> textual findings	91
4.14	<i>Milnacipran-SLC6A4</i> textual findings	92
7.1	<i>Arithmetic means</i> for different semantic similarity metrics . .	113

List of Figures

1.1	Overview of CASSANDRA	4
2.1	Drug repositioning vs. conventional drug development	9
2.2	The Swanson Hypothesis	13
2.3	The open and closed <i>ABC</i> model	15
3.1	Interspecies gene mention normalisation with <i>GNAT</i>	35
3.2	Annotation of <i>MEDLINE</i> abstracts with <i>GoPubMed</i>	37
3.3	Drug gene association datasets compilation	45
4.1	Graphical representation of co-occurrences for drugs and genes	52
4.2	True drug gene associations <i>Venn</i> Diagram	55
4.3	<i>Venn</i> Diagrams for drugs and genes in the evaluation datasets	56
4.4	<i>ROC</i> curves for co-occurrence based profiles	61
4.5	<i>ROC</i> curves for <i>Human</i> and non- <i>Human</i> genes	63
4.6	<i>ROC</i> curves for manually curated profiles	66
4.7	<i>PR</i> curves for all datasets	67
4.8	<i>ROC</i> curves - Manually curated vs. co-occurrence based profiles	69
4.9	<i>PR</i> curves - Manually curated vs. co-occurrence based profiles	70
4.10	<i>ROC</i> curves for different metrics of semantic similarity	74
4.11	<i>PR</i> curves for different metrics of semantic similarity	75
4.12	Density distributions for different measures of semantic similarity	76
4.13	<i>Cathine</i> - <i>GHRL</i>	85
4.14	<i>Fenethylamine</i> - <i>ApoE</i>	88
4.15	<i>Milnacipran</i> - <i>SLC6A4</i>	90
7.1	All ratios <i>PR</i> curves for all datasets	114
7.2	All ratios <i>PR</i> curves - Manually curated vs. co-occurrence based profiles	115

Chapter 1

Motivation

Drug discovery is an expensive and time-consuming process with a low rate of success. The average cost for launching a new drug into the market is estimated to 1.8 billion dollars (Paul et al., 2010) and the traditional time line until a drug is made available for use ranges from 10-17 years. In spite of that, the drugs that make it to the market are very few. Notably, from 1999 to 2008, only 50 compounds were approved by the *Food and Drug Administration (FDA)* in the U.S., out of which 17 were identified as arising from target-based discovery methods (Hurle et al., 2013). For these reasons, drug repositioning constitutes a popular alternative to conventional drug research and development for the past few years. Drug repositioning, meaning the task of finding new targets for old drugs, accelerates the process of drug development, minimizes the associated costs, and, in parallel, contributes to the prevention of noxious adverse events and toxicological liabilities. Via drug repositioning, abandoned drugs come back to use and successful drugs expand their therapeutic applications.

*Why Drug
Repositioning?*

Knowledge pertaining to drug gene associations is considered valuable and can contribute to drug discovery and repositioning. Unravelling putative associations between drugs and gene products can shed light onto the processes of drug delivery and its effects, such as the changes in the cellular metabolism and the occurrence of unexpected adverse events. Such information is scattered across the biomedical literature, the volume of which has been increasing rapidly during the past years. Computational methodologies and more specifically text mining can harness the data that

*Information
"hidden" in
literature*

is publicly available in biomedical articles. With the help of *Information Extraction (IE)* techniques facts and arguments pertaining to drugs and gene products can be retrieved. Linking this textual evidence can lead to the inference of indirect relationships between drugs and genes and hence support computational drug repositioning.

Generally, the inference of implicit knowledge from seemingly unrelated facts has been called *Literature-Based Discovery (LBD)* (Andronis et al., 2011). *Literature-Based Discovery* was successfully applied for the first time on drug discovery by Don R. Swanson in 1986. Swanson spotted two apparently unrelated facts that were reported separately in literature and then brought them together in the formulation of a hidden hypothesis. The first refers to the beneficial properties of *fish oil* towards the reduction of *blood viscosity*. The second refers to high *blood viscosity* as a symptom of a peripheral circulatory disorder known as *Raynauds' Syndrome*. Swanson generated the hypothesis that *fish oil* may have a beneficial effect towards the alleviation of *Raynauds' Syndrome*. This hypothesis was later experimentally validated by the work of DiGiacomo et al. (1989).

The Swanson hypothesis

Consequently, *Information Extraction* techniques and automated *Literature-Based Discovery* constitute valuable tools towards the establishment of hidden hypotheses between biomedical entities and can be utilized to form putative associations between drugs and genes. To establish such associations, as an analogous process to Swanson's *ABC* model, the use of intermediate biomedical concepts becomes critical; two unrelated concepts *A* and *C* (i.e., *fish oil* and *Raynaud's Syndrome*) are indirectly connected via a concept *B* (i.e., *blood viscosity*). Such concepts are provided by biomedical ontologies. Ontologies are hierarchically structured terminologies that capture and formally represent knowledge as a set of concepts within a domain. In the case of the biomedical domain, ontologies have been extensively used towards three major directions: the management of biomedical knowledge, the integration of data and the decision support and reasoning over the concepts that constitute the ontologies (Bodenreider, 2008). Hence, biomedical ontologies can be applied in tandem with the *ABC* model towards the extraction of implicit knowledge and more specifically towards the retrieval of potential drug gene associations.

The contribution of ontologies

1.1 Problem definition and proposed approach

The open problem that this work aims to address is the automated extraction of putative drug gene associations from biomedical text, as a way to boost computational drug repositioning. The current study proposes the application of standardized text mining techniques and the integration of biomedical ontologies towards the identification of drug gene associations. We mine the vast volume of biomedical literature to construct corpus-based profiles of ontological terms for both genes and drugs. These profiles are, then in turn used, to quantify the degree of relatedness between a drug and a gene and hence to establish putative drug gene associations.

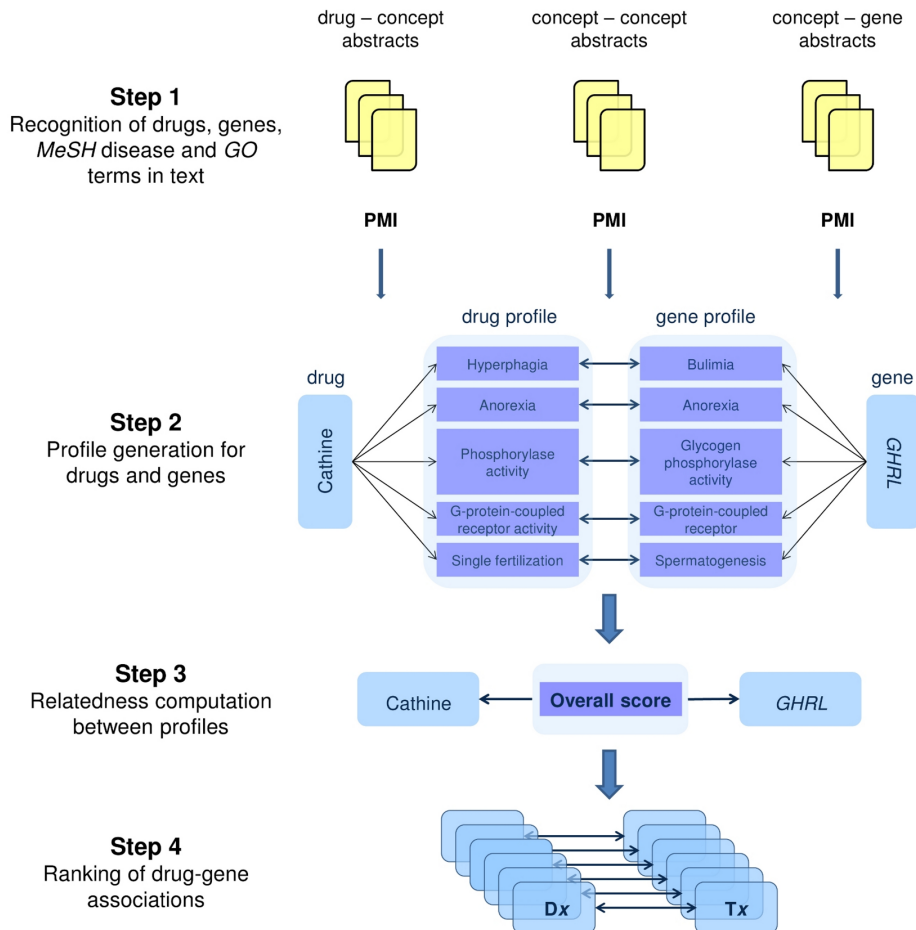
*Mine implicit
drug gene
associations*

More specifically, we introduce CASSANDRA; an unsupervised algorithm that predicts new drug gene associations solely by the systematic co-occurrence analysis of the biomedical terms in all the scientific publications indexed by *MEDLINE*. *MEDLINE* is a freely available bibliographic database which contains journal citations and abstracts for biomedical literature from around the world. The presented method identifies the co-occurrences of ontological concepts with drugs and genes in *MEDLINE* titles and abstracts. The ontological terms are obtained from two popular ontologies of the biomedical domain, i.e., the *Gene Ontology (GO)* (Ashburner et al., 2000) and *Medical Subject Headings*¹ (*MeSH*). CASSANDRA utilizes this co-occurrence information to rank the most related *GO* and *MeSH* concepts to the drug and the gene respectively. These concepts form an individual profile for each drug and gene. Then, by quantifying the statistical semantic relatedness between these profiles, the suggested algorithm assesses and prioritizes the associations between drugs and genes. Notably, the generated profiles can provide an insight into biomedical properties for drugs and genes and contribute to the inference of associations that might not have been included in a database nor reported in the literature.

*Ontological
co-occurrence
based profiles*

¹<http://www.nlm.nih.gov/mesh/>

Figure 1.1: Overview of CASSANDRA



The figure illustrates the major steps of CASSANDRA. The first step involves the recognition of drug and gene names, *MeSH Disease* and *GO* terms in the biomedical text. Then in Step 2, ontological profiles are assigned to drugs and genes, based on the *Pointwise Mutual Information (PMI)* between drugs/genes and ontological terms. The third step involves the computation of the statistical semantic similarity between the ontological profiles. Finally, all pairs of a drug D_x and a target gene T_x are ranked based on the semantic similarity of their profiles.

1.2 Thesis Outline

The following thesis is structured in five main sections. In Chapter 2 (*Background*) the introduction of the thesis follows. In this section, we discuss

the advances and applications of *LBD* along with the use of terminologies towards the establishment of indirect hypotheses and the elucidation of hidden relations between biomedical entities. Additionally, relevant methodologies and algorithms towards the prediction of drug gene associations are presented and compared based on their main characteristics.

Chapter 3 (*Materials and Methods*) describes the materials, text mining tools terminologies and mathematical formulas that CASSANDRA utilizes for the automatic identification of putative drug gene associations from biomedical text. Apart from the information regarding the generation of ontological profiles for drugs and genes and the computation of the statistical semantic relatedness between drugs and genes, the chapter includes the resources and steps that were utilized for the generation of the evaluation datasets.

In Chapter 4 (*Results*) the co-occurrence statistics of drugs, genes, *MeSH Disease* and *GO* terms in *MEDLINE* indexed abstracts and titles are provided. The generation of the datasets used for the algorithm's evaluation and their content is analytically described. This section, provides extensively detailed results that assess the good performance of the algorithm for both literature based and manually curated profiles when tested on all provided evaluation datasets. The proposed measure of semantic relatedness is compared against other traditional measures of semantic similarity. The role of ontologies in the overall performance is also examined. Additionally, Chapter 4 includes case studies of 3 manually evaluated drug gene associations proposed by CASSANDRA. Results regarding the potential of the suggested methodology towards drug repositioning are also provided.

Chapter 5 (*Discussion*) discusses the major characteristics of the algorithm. The role of several factors and decisions taken during the algorithm's implementation are being analyzed. CASSANDRA is compared against other relevant works in the field of automatic *Literature- Based Discovery* and drug target interaction prediction.

The thesis is concluded in Chapter 6 (*Conclusion*) wherein the major contributions of CASSANDRA are summarized. This chapter also reports the limitations and future optimizations of the suggested algorithm towards the prediction of drug gene associations.

Chapter 2

Background

In this work we introduce CASSANDRA; an algorithm for the automated extraction of candidate drug gene associations from biomedical text on the large scale. CASSANDRA focuses on *Literature Based Discovery* and utilizes standardized text mining techniques and ontologies to infer drug gene associations and contribute to computational drug repurposing. This chapter discusses the studies and scientific background that motivated the implementation of CASSANDRA. More specifically, the following questions are addressed

- What is drug repositioning? Why is computational polypharmacology important?
- What is *Literature Based Discovery LBD*? Which are the major principles and studies in the *LBD* domain?
 - Which is the role and contribution of biomedical ontologies?
 - Which are the main methodologies towards information extraction from biomedical text?
- What is the state of the art in the field of computational drug gene association prediction?

2.1 Drug Repositioning

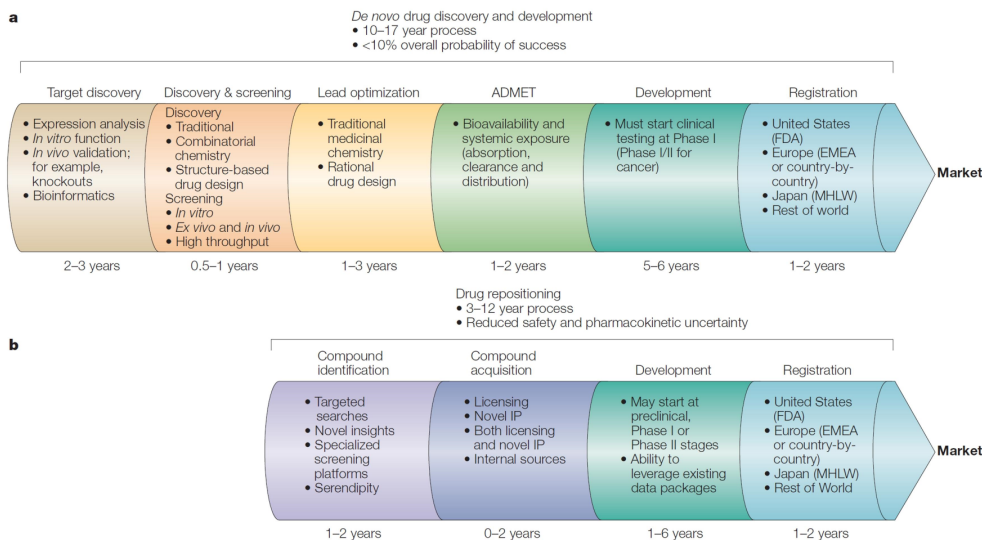
The Research and Development (*R&D*) of new drugs is a particularly time consuming and costly process. Despite the imminent effort of this task, few are the drugs that finally make it to the market. *Polypharmacology* (or *Drug Repositioning*) has lately entered the picture of drug research and development and has been considered the means to overcome the complexity, delays or possible deadlocks of this tedious process. Polypharmacology focuses on drug molecules that interact with multiple targets (Reddy and Zhang, 2013). The old dogma of the *Magic Bullet* (*one drug one target*) introduced by Paul Erlich has been replaced by the *one drug multiple targets* philosophy of polypharmacology. In particular, drug repositioning focuses on the identification of unknown targets for already existing drugs, thus presenting an alternative means in drug research and development. Figure 1.1 demonstrates a comparison between conventional drug research and development and drug repurposing (Ashburn and Thor, 2004).

*What is Drug
Repositioning*

Drug repositioning emerged as a new paradigm after the discovery of drugs with multi-targeting activities that could have either a therapeutic or noxious effect. Such cases are quite a few.

Notably, the story of drug repositioning dates back almost 60 years. A very old case is that of the drug *Plaquenil*. *Plaquenil* (*Hydroxychloroquine*) has been used in the beginning of the 20th century as an antimalarial agent. In 1955, *Plaquenil* was approved by the *U.S. Food and Drug Administration* (*FDA*¹) for the treatment of *Lupus Erythematosus*, a multicomplex autoimmune disease of unknown etiology. Due to its immunosuppressive, anti-inflammatory and antithrombotic properties, *Plaquenil* has become since then the most commonly prescribed antimalarial medication for *Lupus* in the U.S. (Fessler et al., 2005). Perhaps the most famous repositioning case is that of the drug *Sildenafil* (*Viagra*). *Sildenafil* was initially synthesised as an therapeutic agent against *Erectile Dysfunction* and was *FDA*-approved in 1998. Nevertheless, the efficacy of *Sildenafil* in the treatment of *Pulmonary Arterial Hypertension* (*PAH*) led to the extension of its approved application in 2005 (Richalet et al., 2005).

*Expand uses of
successful
agents*

Figure 2.1: Drug repositioning vs. conventional drug development

De novo drug discovery and development is a 10 to 17 year process with a probability of success lower than 10%. Drug repositioning offers the possibility of reduced time and risk as several phases common to de novo drug discovery and development can be bypassed because repositioning candidates have frequently been through several phases of development for their original indication (Ashburn and Thor, 2004).

Another interesting repositioning case is that of the drug *Thalidomide*. *Thalidomide* entered the market on 1957 as a sedative drug and it was used to alleviate morning sickness and nausea in pregnant women. However, the latter led to unprecedented ramifications. Worldwide, there were around 10,000 cases of children born with limb malformation and other developmental defects which were attributed to *Thalidomide*'s use. There has been a long study regarding the etiology behind the teratogenic effects of *Thalidomide*. Finally, in 2010 it became known that *Thalidomide* targets and inactivates the protein *Cereblon* (*CRBN*) which is important for the limb outgrowth and expression of the fibroblast growth factor *Fgf8*. (Ito et al., 2010). Despite the unwanted effects, the antiinflammatory properties of *Thalidomide* made the drug re-emerge as a therapeutic agent against *Erythema Nodosum Leprosum* (*ENL*). FDA approved the respective use of the drug in 1998 and 8 years later, in 2006, granted *Thalidomide* as the first-line medication in the treatment of a specific type of bone marrow

*Bring old
drugs back to
life*

cancer, i.e., *Multiple Myeloma (MM)*. *Thalidomide* was found to inhibit adhesion of *Multiple Myeloma* to bone marrow stromal cells and thereby decreases tumor cell growth, survival and drug resistance conferred by the bone marrow milieu. (Breitkreutz and Anderson, 2008).

However, there are cases wherein the *one drug-multiple targets* philosophy has fired back, leading to serious toxicological liabilities and the subsequent withdrawal of a drug from the market. *Alatrofloxacin* constitutes a representative example; manufactured as a potent antibiotic, the drug was found to cause severe liver toxicity, even of lethal outcome (Qureshi et al., 2011). Fatal *Rhabdomyolysis* was the unwanted adverse event after the administration of *Cerivastatin*. The drug was initially developed to prevent *Cardiovascular Disease* by reducing the levels of cholesterol. However, its use was followed up by the decomposition of damaged muscle tissue, a condition known as *Rhabdomyolysis* (Psaty, Bruce M. et al., 2004).

*Unravel
unwanted
adverse events*

The examples described above demonstrate the broad range of prospects related to drug repurposing. Identifying unknown targets for existing drugs can expand the therapeutic applications of the successful agents (e.g., the cases of *Plaquenil* and *Sildenafilfil*), bring abandoned compounds back to life (as in the case of *Thalidomide*) or unravel any unwanted adverse events (such in the case of *Alatrofloxacin* and *Cerivastatin*). In parallel, the benefits in the financial and timescale related demands of drug research and development are evident, thus making drug repurposing a desirable alternative.

However, a plausible question arises; how feasible is it to cover all possible targets on an experimental full-scale level, and thus enable drug repurposing? This fact poses an significant limitation to polypharmacology and that is exactly wherein the computational methods come into play. In the following section, we discuss the progress and efforts in the domain of computational polypharmacology.

2.1.1 Computational polypharmacology

As it has been mentioned above, drug repositioning poses an industry-wide challenge towards the experimental identification of unknown drug targets

and hidden drug functionalities. Current experimental approaches include large scale *-omic* (genomic and proteomic) analyses, and *siRNA* screens. The former examines the modifications in gene expression levels and the post-translational products of genes to determine disease mechanisms and drug responses (e.g., Kim et al. (2011)). The latter elucidates the role of individual proteins (potential drug targets or off-targets) in the cell by examining the impact that the silencing of their coding gene has on the signal transduction (Jackson and Linsley, 2010). Although successfully applied, the aforementioned approaches result in vast volumes of biomedical information. Utilizing this information to enable drug discovery has, in turn, posed a subsequent challenge in terms of time and complexity (Betz et al., 2005). For that reason, computational methods that expedite drug discovery have been in the spotlight lately.

Although a new field of scientific interest, computational drug repurposing has been rapidly advancing and already counts several success stories. Following two basic directions towards drug repurposing, the computational methods either harness disease/phenotypic similarities and result in novel drug disease relations or take a leap further to identify the exact unknown targets of a drug. A representative example of the former is the *Connectivity Map* (Lamb et al., 2006). The *Connectivity Map* constitutes a reference collection of gene expression profiles for 164 bioactive small molecules (perturbagens) on four human cancer cells lines. Given a gene signature query, the system uses pattern-matching algorithms to return a ranked list of strongly correlated to weakly correlated gene profiles and hence, perturbagens. Drug molecules correspond to a gene expression state and whether the gene signature query constitutes a drug or a disease related phenotype, drug-drug or drug-disease relations can be suggested. Sirota et al. (2011) expanded the application of the *Connectivity Map* and included in their study gene expression profiles for 100 diseases. Based on the hypothesis that a drug with a gene expression signature opposite to that of a disease can be a therapeutic alternative towards the respective disease, they systematically computed the *negative* similarity between drug and disease gene expression profiles. Altogether, they resulted in individual therapeutic predictions for 53 diseases. Most importantly, the aforementioned study resulted in the successful computational repurposing of the anticonvulsant drug *Topimaratate* to *Inflammatory Bowel Disease* (Dudley

*Drug-Disease
methods*

et al., 2011a). Experimental validation in rodents showed that *Topiramate* significantly reduced gross pathological signs and microscopic damage in primary affected colon tissue which constitute manifestations of the *Inflammatory Bowel Disease*. In a recent study, Jin et al. (2014) modified the scoring scheme of *Connectivity Map* and experimentally evaluated their predictions synergistic activity of the drugs *Trolox C* and *Cytisine* for the treatment of *Diabetes, Type 2*.

Emphasizing on target identification, Keiser et al. (2007) classified target proteins based on the set-wise chemical similarity among their ligands. Given a drug query, the *Similarity Ensemble Approach* that they introduced suggests these target proteins whose known ligands share common chemical features with the respective drug. This approach, led to the finding that the drug *Methadone*, apart from an μ -opioid receptor modulator, is also a potent antagonist of the *M3 muscarinic receptor*. This finding was also experimentally validated. Campillos et al. (2008) computationally applied the hypothesis that drugs causing the same adverse events may share the same off-targets. The authors built a network of 1,018 side effect-based drug-drug connections and experimentally confirmed 13 novel drug-target interactions. In a following study, Lounkine and colleagues introduced a conversed approach and used target protein predictions to associate drugs with unintended adverse events (Lounkine et al., 2012). They focused on a set 656 marketed drugs and 73 targets with experimentally established associations to certain adverse events. Via the *Similarity Ensemble Approach*, they calculated the drug-target similarity and then they assigned novel side effects to drugs. Their work resulted in the experimental validation of 125 novel drug-target interactions.

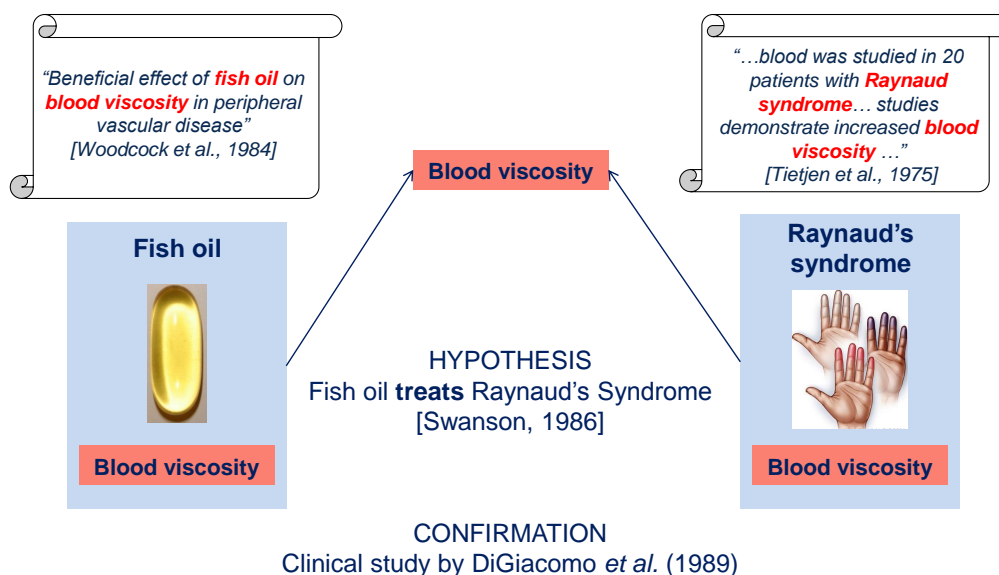
The studies described above suggest the efficient contribution of computational methods to polypharmacology and particularly drug repositioning. Their successful application to drug discovery has triggered a plethora of following up studies that span across different data types (structural, textual or transcriptional/genomic data) (Hurle et al., 2013) and combine different strategies, e.g., networks, text mining, machine learning (Dudley et al., 2011b). As the biomedical data grows, the necessity for such methodologies grows, as well.

*Drug-Target
methods*

2.2 Literature Based Discovery in the Biomedical Domain

Literature Based Discovery has been defined as the process of (semi)-automatic inference of implicit knowledge out of literature (Weeber et al., 2005). The first and most indicative paradigm of literature-based hypothesis was introduced by Don R. Swanson in 1986 (Swanson, 1986) (Figure 2.2).

Figure 2.2: The Swanson Hypothesis



The figure illustrates the first successful application of the *ABC* model. Two seemingly unrelated facts form the hypothesis that fish oil is a treatment alternative to *Raynaud's Syndrome*. The hypothesis was initially established by Swanson (1986) and DiGiacomo et al. provided the experimental validation 3 years later.

Swanson spotted two independent reported facts in literature that, nevertheless, shared one common factor; *blood viscosity*. The former refers to

The Swanson Hypothesis

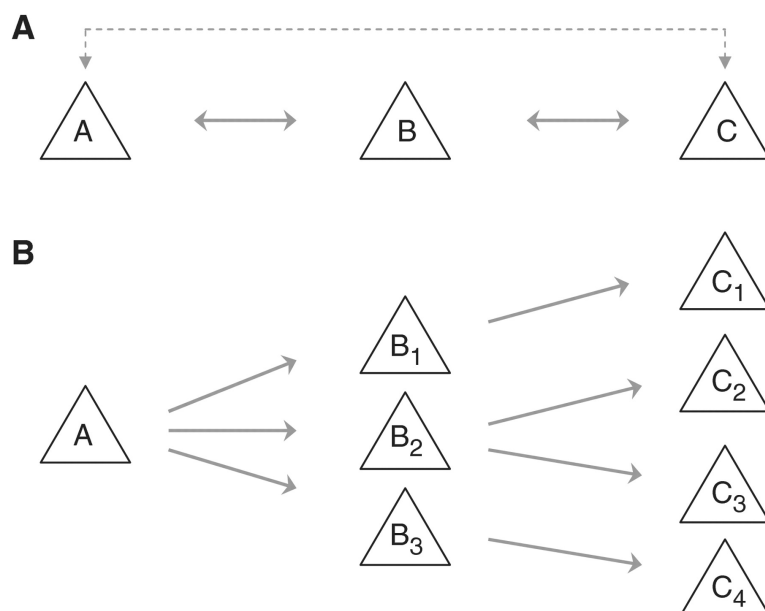
Raynaud's Syndrome, a circulatory disorder which exhibits excessively reduced blood flow to fingers and toes, and hence results in increased *blood viscosity* (Tietjen et al., 1975). The latter reports that *Dietary Fish Oil* appears to lower *blood viscosity* (Woodcock et al., 1984). Observing that *Raynaud's Syndrome* and *Dietary Fish Oil* had never been reported together in literature before, Swanson connected these two assertions and formed the hypothesis that *Dietary Fish Oil* may have a beneficial effect on *Raynaud's Syndrome*. Indeed, this hypothesis was afterwards experimentally validated by DiGiacomo et al. (1989). On the same fashion, Swanson along with his colleague Smalheiser furtherly exploited the *mutually isolated literatures* and postulated several other hypotheses (Swanson, 1990; Swanson and Smalheiser, 1996, 1998). In one of them, they suggested the therapeutic effects of *Magnesium* in *Migraine* (Swanson, 1988). Although, this hypothesis has not been experimentally confirmed, empirical treatment of *Migraine* patients with *Magnesium* has shown promising results (Mauskop and Varughese, 2012).

Establishing indirect associations between two concepts *A* and *C* via an intermediate concept *B* has been referred to either as the *Swanson Hypothesis* or the *ABC* model. The *ABC* model has been repeatedly used to discover hidden associations in the biomedical domain. The developed methodologies are either used to recover the original hypotheses proposed by Swanson (Cameron et al., 2013; Cohen et al., 2010a; Srinivasan, 2004; Weeber et al., 2001), or they take a step further and establish new hypotheses (Gramatica et al., 2014; Dong et al., 2014; Baker and Hemminger, 2010; Ahlers et al., 2007; Srinivasan and Libbus, 2004; Wren et al., 2004; Weeber et al., 2003). Weeber et al. (2005) elaborated further on the *ABC* model and discriminated the *Literature Based Discovery* methods into *closed* and *open* discovery methods. As shown in Figure 2.3 the *open* model, the concepts *A* and *C* are given and the hypothesis of *A*'s connection to *C* has to be established by the identification of intermediate concepts *B*. On the other hand, the *closed* model focuses on the concept *A* and using this as a starting point browses the intermediate concepts *B* that indirectly connect it to various final concepts *C*.

Several algorithms and automated tools have adopted the *ABC* model to assist *Literature Based Discovery* in the biomedical domain (see Table 2.1). Smalheiser and Swanson were the first to provide *ARROWSMITH*, a

*ABC model -
definition*

*ABC model -
tools and
algorithms*

Figure 2.3: The open and closed *ABC* model

The two representations of *ABC* model. (A) *Closed* discovery model: the concepts *A* and *B* share an implicit connection if they are explicitly associated to a common concept *B*. (B) *Open* discovery model: Concept *A* is indirectly connected to several concepts *C* via a set of intermediate concepts *B*. Figure as in Andronis et al. (2011).

closed discovery automated system (Smalheiser and Swanson, 1998). Using only the titles from the *MEDLINE* indexed articles and for two given concepts *A* and *C*, *ARROWSMITH* retrieves the article sets pertaining to *A* and *B* respectively and generates a set of intermediate terms *B* (i.e., words/phrases) that are found to overlap in these article sets. *FACTA+* also utilizes *MEDLINE* and uses *open* discovery to retrieve indirect associations between drugs, diseases, chemical compounds and proteins/genes (Tsuruoka et al., 2011); the system accepts a set of keywords as a query input and returns all possible directly associated concepts. It then uses these so-called *pivot* concepts as intermediates and retrieves the directly associated to them *target* concepts. The query can be either a single term or a biomolecular event, whilst the *query-pivot* and *pivot-target* concept relations are suggested based on co-occurrence statistics. Quite similar to *FACTA+* is the *CoPub Discovery* system (Frijters et al., 2010) which is also based on co-occurrence statistics. *CoPub Discovery* supports both

closed and *open* discovery between drugs, genes and diseases with the use of biological processes, pathways and genes as the intermediate concepts.

Another automated tool supporting both *open* and *closed* discovery is *BITOLA* (Hristovski et al., 2005). *BITOLA* uses a two step establishment of indirect relationships between disease terms and genes. First, it establishes associations between diseases or genes and *Medical Subject Headings* (*MeSH*)³ descriptors based on co-occurrence statistics. The higher the number of *MeSH* descriptors connecting the disease term and the gene, the stronger the implicit association between them. *MeSH* is a controlled vocabulary created for indexing articles in the life sciences, whilst the *MeSH* descriptors are the *MeSH* terms that are assigned to *MEDLINE* indexed articles through manual curations. Such terms are used in the methodology of Baker and Hemminger (2010) wherein indirect associations are established between fixed chemicals and disease terms via proteins.

Although, the majority of *ABC* model methodologies uses co-occurrence statistics, there are a few methods that use *Natural Language Processing* techniques to establish relationships between a concept *A* or *C* with an intermediate concept *B* (Cairelli et al., 2013; Cohen et al., 2012, 2010a; Hristovski et al., 2008). Table 2.1 provides an overview of the most distinctive tools implemented for *Literature Based Discovery*. As shown, most of the tools utilize *MEDLINE*, but few of them deviate from the traditional *ABC* model. The majority remains to the hypothesis retrieval without applying any further refinement. Additionally, although ontologies constitute useful resources of terms for the hypothesis establishment, these are not extensively used. The tools consider rather their own specified terminologies, as in the case of *CoPub Discovery* for example.

The seminal *ABC* model goes beyond literature applications. Observing things from a more general perspective, several works in the domain of life sciences have generated hypotheses based on the implicit connections between two entities. For example, Campillos et al. (2008) use side effects as the intermediate concepts to relate drugs to targets. The *Connectivity Map* can be also considered an expansion of the *ABC* model (Lamb et al., 2006); the gene expression profiles are connecting two drugs or a drug with a disease. *WENDI* is another tool that concentrates data from various

³<http://www.nlm.nih.gov/mesh/>

Table 2.1: Tools for Literature Based Discovery

Tools	MEDLINE abstracts	Relation		Discovery		LBD extended			Concepts			Associations refinement	
		co-occurrence	pattern	open	closed		GO	MeSH	UMLS				
ARROWSMITH		✓			✓								
BITOLA	✓	✓	✓	✓	✓				✓				✓
LitLinker	✓	✓		✓					✓				✓
EpiphaNet	✓		✓	✓	✓			✓				✓	
CoPub Discovery	✓	✓		✓	✓								✓
FACTA+	✓	✓		✓	✓								
SemanticMEDLINE	✓		✓	✓	✓			✓					✓

The table demonstrates the key features of several tools for the automatic *Literature Based Discovery*. The tools utilize *MEDLINE* abstracts (except *ARROWSMITH* which utilizes titles) and apply either co-occurrence or patterns/rules to extract pairwise relations between the concepts. They support either *open* or *closed* discovery. Few of them deviate from the traditional *ABC* model and extend *Literature Based Discovery*. Most tools remain to hypothesizes generation without applying further refinement. Additionally, although constituting useful resources of terms, ontologies are not extensively utilized.

biomedical repositories and uses this data to indirectly associate drugs with genes and diseases (Zhu et al., 2010).

Apparently, two are the prerequisites when applying the *ABC* model on the biomedical literature. The former is a set of terms/entities found in text and the latter is to establish pairwise relations between them. Hence, the arising questions are; which are the terms that are important for *Literature Based Discovery* in the biomedical domain? How can we extract these terms and any existing pairwise relations between them? On the one hand, there is the necessity for established dictionaries and structured terminologies (*ontologies*) indexing the biomedical literature. On the other hand, there is the necessity for *text mining tools* that are able to successfully spot the references of such terms in text and retrieve the relations between them.

*Text mining
and ontologies
for ABC
model*

2.3 Biomedical Ontologies

In computer science, an *ontology* is defined as the technique used to represent and disseminate knowledge about a specific domain by modeling the elements in that domain and the relationships between them (Gruber, 1991; Bodenreider and Stevens, 2006). According to Maojo et al. (2011), ontologies involve: (a) modelling primitives that include objects, classes or categories (e.g., cells, organs, persons), (b) semantic relationships between these primitives (e.g., kidney *is_part_of* human body), (c) properties pertaining to each class (descriptive or functional).

Definition

Bodenreider (2008) classified the role of ontologies into three major categories. The first is the knowledge management, such as indexing and information retrieval. For example, *Medical Subject Headings (MeSH)* are used to index the *MEDLINE* articles. Shah et al. (2009a) implement a prototype system for the automated annotation and indexing of gene-expression data sets, image descriptions, clinical trial reports and *MEDLINE* indexed abstracts with concepts from the appropriate ontologies. *SemRep* extracts semantic predications (subject-predicate-object triples) from text based on *UMLS (Unified Medical Language System)* concepts, (Rindfleisch and Fisman, 2003).

*The role of
biomedical
ontologies -
Indexing*

The second role is to integrate (heterogeneous) data and disseminate it. A prerequisite for that is, of course, semantic interoperability. *Gene Ontology (GO)* is such a resource (Ashburner et al., 2000). *GO* is a popular and publicly available controlled vocabulary that concentrates information regarding genes, gene products and their attributes. Its major goal is to unify the genes' representation across databases and provides tools that allow easy access to the *GO* data and annotations. Apart from gene annotation, *GO* has also been used to structure *MEDLINE* indexed articles and abstracts by the semantic knowledge-based search engine *GoPubMed* (Doms and Schroeder, 2005).

*Data
Integration*

The third role of ontologies is to assist decision support and automated reasoning. Blonde et al. (2011) performed a semi-automated approach to reason over different ontologies and managed to infer 158 million previously hidden knowledge statements. The *Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT)*⁴ is a formalized comprehensive terminology for the *Electronic Health Record (EHR)*. Its formal model has allowed reasoning services to derive implicit relations (subsumptions) from the ones explicitly represented by automatically computing the axioms responsible for these relations (Baader and Suntisrivaraporn, 2008). In the same context, Magka et al. (2014) implement a reasoning algorithm that significantly speeds up the automated classification of the chemical compounds included in the *ChEBI (Chemical Entities of Biological Interest)* ontology and contribute to the ontology's curation by identifying missing and contradictory subsumptions.

Reasoning

Ontologies have been also used as means to establish or unravel relations between biomedical terms and hence to assist information extraction. Srinivasan and Libbus (2004) were among the first to use the *MeSH* terms that index *MEDLINE* articles to establish topic profiles and generate hypotheses similar to the *Swanson* example. Several methodologies have since then exploited profiles of *MeSH* terms to interrelate biomedical entities (Baker and Hemminger, 2010; Cheung et al., 2012, 2013; Dong et al., 2014). *BioMine* collects data from several biomedical knowledge bases and, among others, utilizes *GO* terms towards link prediction in a large network of biomedical entities (Eronen and Toivonen, 2012). Schlicker et al. (2010) made a step

*Connect
biomedical
entities via
ontologies*

⁴<http://www.ihtsdo.org/snomed-ct/snomed-ct0/>

further and prioritized disease-gene pairs based on the similarity of their *GO* profiles. Plake (2010) apply an elementary version of the same idea to relate drugs and genes. Several studies have used phenotypic terminologies on that respect, as well (Oellrich et al., 2014; Smedley et al., 2013; Washington et al., 2009). To extract drug target information, Hoehndorf et al. (2013) compute the similarity between mouse model and drug-induced phenotypes with the use of *Mammalian Phenotype Ontology (MP)* (Smith et al., 2004) and the *Human Phenotype Ontology (HPO)* (Robinson et al., 2008).

As shown above, the relationships between biomedical entities (e.g., proteins or diseases and genes) can be established via the same or similar ontological concepts. Hence, a variety of metrics has been implemented to assess the semantic similarity (relatedness) between two ontological concepts. Pesquita et al. (2009) classify these metrics in two basic categories; the edge-based and the node-based approaches. Edge-based approaches mainly consider the similarity of two ontological terms as a function of the distance between them in the ontology (Wu and Palmer, 1994; Leacock et al., 1998); the shortest the path that connects the two terms via their *Least Common Ancestor (LCA)*, the higher their similarity. Node-based approaches apart from hierarchical information also compare certain properties of the ontological terms, such as their *Information Content (IC)* on an specific corpus (Resnik, 1995; Jiang and Conrath, 1997; Lin, 1998; Couto et al., 2005; Pesquita et al., 2008).

*Semantic
similarity
metrics*

Several approaches have explicitly focused on *GO*. Wang et al. (2007) propose an edge-based similarity metric that considers the semantic contribution of each edge type (*is-part-of* or *is-a*). Jain and Bader (2010) use the number of a node's descendants to formulate the topological *Information Content* and cluster relative *GO* terms; terms that belong to the same subset are assigned a higher similarity score. Yang et al. (2012) explore the hierarchy beneath the *GO* terms and model the uncertainty of the nodes based on gene annotation information to improve existing measures of semantic similarity.

The ontologies mentioned above constitute only some popular examples. Currently, there is a variety of biomedical ontologies covering certain sectors of knowledge at different levels of specificity. For example, there are

ontologies dedicated to a specific disease (e.g., the *Cardiovascular Disease Ontology* or *CVDO* (Barton et al., 2014)), a specific type of cells (e.g., *Beta Cells Genomic Ontology* or *BCGO* (Zheng et al., 2013)) or a certain organism (e.g., *Xenopus Anatomy and Development* or *XAO* (Segerdell et al., 2013)). Several initiatives towards the formulation of a common set of principles for different ontologies also exist, such as the *Open Biomedical Ontologies (OBO)* consortium (Smith et al., 2007). As the amount of data produced in biology exponentially increases due to the advent of the *genomic* era and the high-throughput techniques developed in sequencing, drug and phenotypic screening, the role of ontologies becomes more and more significant (Hoehndorf et al., 2012).

2.4 Mining biomedical text

Deciding which biomedical terms/ontological concepts to annotate in text constitutes only the first step towards the extraction of information in the biomedical domain. The successful identification of these very terms and the extraction of any of their relationships reported in text are the following and particularly demanding steps. Given the exponential growth of biomedical literature (Hunter and Cohen, 2006), the automation of this process is itself a significant challenge. Accordingly, traditional biomedical text mining systems usually consist of two modules; the former recognizes biological entities or concepts in text and the latter focuses on the extraction of any relations existing between these entities (Zweigenbaum et al., 2007).

With regards to the biomedical text that constitutes the input of text mining systems, scientific abstracts and titles are widely used mainly due to their public accessibility through *PubMed*⁵ (i.e., an interface to browse the *MEDLINE* database of indexed articles in life sciences) (Vincze et al., 2008). It has been also demonstrated that text mining tools perform better in abstracts than in full-text articles (Cohen et al., 2010b). Gijón-Correas et al. (2014) predict the relatedness of a list of chemicals retrieved from *MEDLINE* indexed abstracts and titles to a query topic. *GoPubMed*

*Abstracts and
full-text*

⁵<http://www.ncbi.nlm.nih.gov/pubmed>

(Doms and Schroeder, 2005) and *PubTator* (Wei et al., 2013) are web-tools that allow users to retrieve articles associated with specific bioconcepts based on their *MEDLINE* abstracts. *LitInspector* performs signal transduction pathway mining again from *MEDLINE* abstracts. *PolySearch* combines abstracts and database factoids to unravel biomedical relations (Cheng et al., 2008) Nevertheless, there exist methodologies that investigate the extraction of biomedical information from full-text articles. In a recent work, Név  ol et al. (2012) propose the use of text mining on full-text articles to tackle the automatic curation of links between biological databases and the literature. *Pharmspresso* automatically extracts pharmacogenomic facts from full text articles (Garten and Altman, 2009). Hakenberg et al. (2010) focus on protein-protein interaction extraction from full-text articles. *GeneView* is a semantic search engine built upon a comprehensively annotated version of *MEDLINE* abstracts and openly available full-text articles (Thomas et al., 2012).

2.4.1 Annotation of Biomedical Terms

The identification of terms in biomedical text is an active field of study. There exist a variety of *Name Entity Recognition* tools often addressing specific terms categories. The most popular term category is that of genes/protein names (Campos et al., 2013; Fontaine et al., 2011; Torii et al., 2009; Hakenberg et al., 2008a; Settles, 2005). There are also species/organism detection methodologies like *LINNAEUS* (Gerner et al., 2010) which are commonly used together with gene/protein name annotators to alleviate the high inter-species ambiguity that characterizes gene names. Other *Named Entity Recognition* tools identify protein mutations (Burger et al., 2014; Winnenburger et al., 2009; Caporaso et al., 2007) or disease names (Leaman et al., 2013). Lately, there has been an effort to tackle annotation of text with chemical names (Rockt  schel et al., 2012; Jessop et al., 2011). Tools and strategies have been also applied towards the annotation of ontological terms in text (Aronson and Lang, 2010; Shah et al., 2009b; Doms and Schroeder, 2005). *Name Entity Recognition* tools usually follow dictionary based matching approaches combined with machine learning components (Huang et al., 2011; Wermter et al., 2009). In some cases only machine

*Term
categories and
approaches*

learning (Leaman et al., 2013) or rule-based approaches are used (Kang et al., 2012).

Each term category entails its own challenges in the process of *Name Entity Recognition*. For example, gene/protein names are characterized by ambiguity not only across species but also within species, with common English words and with medical sublanguage terms (Wermter et al., 2009). Apart from overlapping with gene names, disease names are frequently abbreviated when mentioned in text (Leaman et al., 2013). The case of chemical name recognition poses an even greater challenge due to the highly heterogeneous and various ways of naming them (e.g., chemicals can be referred to by their *IUPAC* (*International Union of Pure and Applied Chemistry*)⁶ or brand name) (Rocktäschel et al., 2012). The recognition of concepts (ontological terms) in text involves an additional complication, since there is often a disconnect between what is captured in an ontology and what is found to be explicitly stated in text (Funk et al., 2014).

Challenges

2.4.2 Relation Extraction

After the identification of biomedical entities in text, the establishment of the relations between them follows. Several methodologies are developed to target usually one type of relations. For example, there are approaches focusing in the retrieval of protein protein interactions (Hakenberg et al., 2010; Jelier et al., 2005; Cohen et al., 2005), or point protein mutations related to a specific disease (Burger et al., 2014; Doughty et al., 2011). Other studies focus on gene-phenotype (Paik et al., 2014), drug-disease (Cheung et al., 2013), disease-phenotype (Xu et al., 2013) or protein-ligand associations (Chang et al., 2012).

Types of relations

Towards the establishment of relations between biomedical entities, there are two basic approaches. The former is to retrieve relations based on the terms co-occurrence statistics. This method builds on the assumption that two entities found together in the same abstract or sentence/phrase are likely to be related. In an early study, Jenssen et al. (2001) build a gene-gene relationship network by weighting the co-occurrences of gene-gene pairs in *MEDLINE* abstracts. Garten et al. (2010) learn drug-gene

Co-occurrence

⁶<http://www.iupac.org/>

associations from a co-occurrence based network of drugs and genes also built from *MEDLINE* abstracts. Paik et al. (2014) extract disease co-occurrences from medical reports and show that there is a correlation between disease comorbidity and overlapping disease-related protein-protein interactions. *FACTA+*, *Pescador*, *Pharmspresso* and *GoGene* are just a few examples of tools that generate biomedical relations based on co-occurrence data (Tsuruoka et al., 2011; Barbosa-Silva et al., 2011; Garten and Altman, 2009; Plake et al., 2009).

The latter method to extract biomedical relations from text is to apply predefined or automatically generated patterns/rules. *AutoBind* is a pattern-based method for the automated extraction of protein-ligand associations (Chang et al., 2012). *RelEx* and *OpenDMAP* apply rules on dependency parse trees to extract protein-protein interactions (Hunter et al., 2008; Fundel et al., 2007). Cou (2010) also use dependency graph rules to detect pharmacogenomic relations. *EventMine* learn predicate-argument structures to extract biomolecular events from text (Miwa et al., 2012). *SemRep* extracts semantic predications (subject-predicate-object triples) from text between *UMLS* concepts (Rindflesch and Fiszman, 2003).

Evidently, these strategies are extensively applied in relation extraction and consequently, in *Literature Based Discovery* (see Table 2.1). However, co-occurrence based statistics tend to be slightly more popular compared to the pattern based strategies. This has several explanations. Unlike syntactic patterns, co-occurrence based statistics are relatively easier to implement and can be applied on the large scale without text pre-processing requirements (Zweigenbaum et al., 2007). Additionally, they are domain-independent. For example, it is impossible to use the same syntactic patterns to retrieve a drug-disease relationship and a protein-protein interaction. However, that very specificity of pattern-based approaches provide the user with the exact type of relation between the entities and thus opt for high precision. Ideally, the two methodologies combined produce high quality results. Xu and Wang (2014) learn syntactic patterns over automatically recovered occurrences of known drug-side effect pairs from literature. In an earlier study Bunescu et al. (2006) use both strategies to recover protein-protein interactions from text. Table 2.2 summarizes the advantages and disadvantages of each approach.

Patterns/rules

Pros & cons

Table 2.2: Co-occurrence vs. Patterns

	Co-occurrences	Patterns/Rules
Pros	Fast High recall Straightforward implementation	Type of relation High precision
Cons	No relation type Low precision	Text-preprocessing Domain-dependent Low recall Laborious generation

The table reports the advantages and disadvantages of co-occurrence based and pattern-based relation extraction.

2.5 Drug gene association prediction

To assist computational drug repurposing, many methodologies have been established towards the prediction of target proteins for drugs. Popular strategies involve the use of side effect similarity (Campillos et al., 2008), chemical structural similarity (Keiser et al., 2007) and protein structural similarity (Kinnings et al., 2009) for the identification of drug repositioning candidates. Other methodologies apply large scale molecular docking analysis of known drugs against known targets to identify off-target proteins with novel scaffolds or proteins structurally dissimilar to known targets (Li et al., 2011).

Lately, many studies combine chemical structural and protein sequence similarity to predict drug target interactions (van Laarhoven and Marchiori, 2013; Mei et al., 2013; Fakhraei et al., 2013; Perlman et al., 2011; Bleakley and Yamanishi, 2009). Towards the same direction, several works use on top of that pharmacological effects similarity (e.g., side or therapeutic indication), as well (Kim et al., 2013a; Yu et al., 2012; Yamanishi et al., 2010). Notably, Takarabe et al. (2012) show that the use of pharmacological effects similarity of drugs in tandem with the genomic similarity of targets in a pairwise kernel regression model achieves a better performance than the use of chemical similarity and genomic similarity combined.

Type of data

Protein-protein interactions (PPI) networks have also come to play. Hansen et al. (2009) rank human genes to a query drug by building on a local network of known interactions and learning on the similarity of the query drug (by both structure and indication) with drugs that interact with gene products in the local network. Emig et al. (2013) combine a PPI network with disease microarray data and learn both global and local features to rank a disease signature against a set of drug targets.

Generally, there are very few unsupervised methods towards the prediction of drug-gene associations. Chen et al. (2012) build a network of so-called semantically linked entities to drugs based on publicly available repositories which comprise drug-related information (i.e., pathway, side effect, disease data). Based on the topology and semantics of the neighborhood, they build a statistical model to infer drug-gene associations (edges) in the network. Of course, this method depends on the information completeness of the network; the more information (links) is known for a drug, the better is the ability of the method to successfully predict its target. Wu et al. (2012) harness the biomedical literature. They annotate drugs and genes on a subset of *MEDLINE* abstracts and examine the performance of the *Latent Dirichlet Allocation* towards the ranking of drug-gene associations on different levels of co-occurrence.

*Few
unsupervised
methods*

On the other hand, the supervised techniques for the prediction of drug-gene associations are numerous (Alaimo et al., 2013; Chen and Zhang, 2013; Mei et al., 2013; Yu et al., 2012; Perlman et al., 2011; Yamanishi et al., 2010). The majority of them views the set of drug target interactions as a bipartite graph, i.e., a graph where edges are only allowed to pass between one class of nodes (drugs) and the other (targets). Bleakley and Yamanishi (2009) were the first to use such a representation to predict drug-target interactions (edges in the bipartite graph) via learned local models from chemical and genomic data. In the same context, Fakhraei et al. (2013) and Gönen (2012) train a probabilistic model to predict edges on the bipartite drug-target interaction (DTI) network.

*Supervised
bipartite graph
methods*

Laarhoven et al. (2011) construct gaussian kernel functions from binary interaction profile vectors for drugs and targets and show that the topology of the bipartite DTI network is on its own a substantial source for predicting drug-target interactions. Cheng et al. (2012a) also learn the topological

DTI network similarity and demonstrate that *Network Based Inference* performs better than *Drug or Target based Inference* towards the prediction of new targets for known drugs. The latter is a limitation of the two aforementioned approaches; when a drug lacks known target information, it is not possible to predict new targets.

Other supervised methodologies apply the *Nearest Neighbor* algorithm to predict new targets for drugs of unknown interaction information (van Laarhoven and Marchiori, 2013; He et al., 2010). Wang and Zeng (2013) train a two-layer graphical model to predict the type of interaction between a drug and a target. Kim et al. (2013a) demonstrate that drug-drug interaction data is a contributing feature towards the prediction of drug-target interactions. Gao et al. (2013) assign drugs to target groups based on the associations of their ontological *ChEBI* terms. Other works learn from chemogenomic and structural activity features (Cheng et al., 2012b).

Notably, literature-based methods are limited. Zhu et al. (2005) learn from gene-gene, compound-compound and gene-compound co-occurrence data in *MEDLINE* abstracts and detect implicit gene-compound associations. Garten et al. (2010) replaced the drug-gene network in Hansen et al. (2009) by a gene-drug network derived from the sentence level co-occurrence of drugs and genes in full-text articles. They show that the logistic regression classifier trained on this network is as good as (and sometimes better than) the one trained on the network built from manually curated knowledge bases (i.e., the case in Hansen et al. (2009)). Plake (2010) presents an early stage approach that relates drugs to genes via concepts derived from *MEDLINE*. However, this work suffers from rudimentary evaluation; there is no filtering of the concepts applied when establishing drug gene relations and the dataset used for the evaluation fails to demonstrate the real efficacy of the proposed method. Still, this work motivates the use of literature towards drug gene association prediction.

*Literature
features
unexplored*

To conclude, few are the literature based approaches towards the prediction of drug gene associations. Moreover, the majority follows machine learning techniques that integrate features from highly diverse data; chemical structures, target aminoacid sequences, pharmacogenomic or chemogenomic information, protein-protein interaction data or even disease microarray data

(e.g., as in Emig et al. (2013)). Additionally, quite some of them are limited to predictions that consist of known targeted drugs and druggable proteins (e.g., Laarhoven et al. (2011); Alaimo et al. (2013)). Table 2.3 provides the respective overview. In a recent study, Pahikkala et al. (2014) revise the supervised methods towards drug-target interaction prediction and pinpoint the drawbacks. Among others, they state that these models are often being constructed and evaluated under overly simplified settings that do not reflect the real-life problem in practical applications.

Table 2.3: Drug gene association prediction methodologies

Tools	Unsupervised	Large-scale	<i>only-DTI</i> Drugs & Genes	<i>MEDLINE</i> mining	Ontologies	Phenotype	PPIs	Structure/ Sequence
Hansen et al.							✓	
Garten et al.				✓			✓	
Yamanishi et al.			✓			✓		✓
Laarhoven et al.			✓					
Cheng et al.			✓					
Gönen								✓
Alaimo et al.			✓					✓
Mei et al.								✓
Emig et al.		✓				✓	✓	
Chen et al.	✓	✓			✓	✓	✓	✓
Zhu et al.				✓				
Perlman et al.								✓
Wu et al.	✓							
Gao et al.					✓			

Network

The table provides an overview between the computational methodologies for drug gene association prediction. Very few utilize literature data. The vast majority are supervised learning and network-based approaches that exploit different drug and gene properties as features. Notably, quite some of them are limited to predictions of drug gene associations wherein the drugs and genes are already involved in known drug target interactions.

Chapter 3

Materials and Methods

CASSANDRA utilizes the biomedical literature to identify latent relations between drugs and genes by creating their ontological profiles and measuring their relatedness. This chapter reports the materials, text mining methods and mathematical formulas that were utilized to implement and evaluate CASSANDRA's efficacy.

Regarding the implementation, the following questions are answered

- Which text is annotated? Which terms are searched in text?
- Which annotators are utilized for the recognition of terms in text?
- How are the drug and gene profiles generated?
- How is the semantic relatedness between the profiles estimated?

With respect to the algorithm's evaluation:

- Which datasets are used?
- How are these datasets generated?
- Which alternative metrics of semantic similarity are compared against the semantic relatedness metric utilized by CASSANDRA?

3.1 Materials and Methods - Implementation

3.1.1 Resources - Text and terms

Biomedical text

CASSANDRA mines the abstracts and titles of *MEDLINE* indexed articles to establish latent drug gene associations. The *MEDLINE* database is a freely available bibliographic repository which contains journal citations, abstracts and full-body articles of biomedical literature from around the world. Currently, *MEDLINE* comprises around 24 million records. However, only $\sim 2\%$ of *MEDLINE* entries have open-access full-text articles available for text mining (Thomas et al., 2012). For that reason, CASSANDRA utilizes only the abstracts and titles of approximately 23 million biomedical articles that were available at the time of the algorithm's implementation (March 2013).

Biomedical terms

CASSANDRA searches for hidden indirect associations between drugs and genes. Each drug and each gene are assigned an ontological profile that consists of terms derived from *Gene Ontology* (*GO*) and *Medical Subject Headings Diseases* (*MeSH Diseases*).

As far as the drug terms are concerned, CASSANDRA utilizes an *in-house* dictionary of drugs derived from the *DrugBank* database (Wishart et al., 2008). The *DrugBank* database is both a bioinformatics and cheminformatics resource which combines detailed data descriptions and comprehensive target information for an extensive list of drugs. The database currently contains over 7,000 drug entries. Experimental drugs occupy around 70% of the database. *DrugBank* also includes small molecule and biotech (protein/peptide) drugs that are approved by the *U.S. Food and Drug Administration* (*FDA*). Each drug record (*DrugCard*) contains more

*Drugs from
DrugBank*

than 200 data fields with half of the information being devoted to drug/chemical data and the other half devoted to drug target or protein data (enzymes, transporters, carriers).

Regarding the gene names, these are derived from *UniProt* Knowledgebase (*UniProtKB*)¹. *UniProtKB* constitutes a central protein resource. It captures the core data of a protein (i.e., the amino acid sequence, protein name or description, taxonomic data and citation information) and any annotation information available, such as ontological terms, cross-references and classifications. *UniProtKB* consists of *UniProtKB/TrEMBL* and *UniProtKB/SwissProt*. The former contains unreviewed, automatically annotated records. The latter, i.e., *UniProtKB/SwissProt*, contains manually curated records and is utilized by CASSANDRA.

*Genes from
UniProtKB*

As far as the ontologies are concerned, *GO*² and *MeSH*³ constitute the terms that form the profiles and are used towards the characterization of drugs and genes. *GO* is the major and freely available controlled vocabulary of genes and gene products (Ashburner et al., 2000). *GO* explicitly focuses on the unification of genes's representation across all species (Consortium, 2008). *GO* is a dynamic ontology that is maintained and enriched regularly. It consists of three subontologies; *Biological Process*, *Molecular Function* and *Cellular Component*. *GO* is structured as a *Direct Acyclic Graph*, meaning a graph with no directed cycles. Each term has defined relationships with other terms in *GO*.

Gene Ontology

Medical Subject Headings (*MeSH*) is a popular medical controlled vocabulary thesaurus. It is freely available and maintained by the *U.S. National Library of Medicine* (*NLM*). *MeSH* consists of sets of terms (categories) that are hierarchically structured. It is primarily used for indexing *MEDLINE* articles and it is continually maintained. CASSANDRA focuses on the *Disease* terms (descriptors) to characterize genes and drugs. Only *MeSH Disease* terms found in text are used. *MeSH* terms provided by *MEDLINE* are ignored so that CASSANDRA remains independent of any manual annotation and hence generally applicable.

MeSH

¹<http://www.uniprot.org/>

²<http://geneontology.org/>

³<http://www.nlm.nih.gov/mesh/>

3.1.2 Recognition of terms in text

The recognition of terms in text involves the identification of drug names and their synonyms, gene names and *GO* and *MeSH* terms. The recognition of genes and drugs in biomedical text is analytically described in Plake (2010). For the sake of coherence, the details of the annotation process are once more explained.

Regarding the gene annotation process, *GNAT* was utilized (Hakenberg et al., 2008a). *GNAT* is a publicly available system that handles inter-species gene mention normalisation. Unlike traditional gene annotators, *GNAT* uses background knowledge on genes to assign ambiguous gene names to the correct *Entrez Gene* identifiers with a reported *F*-measure of 81.4% (90.8% precision at 73.8% recall). On the single species task considering only human genes, *GNAT* achieved an *F*-measure of 85.4%. Briefly, gene annotation with *GNAT* is divided in four stages. First, it searches for different species mentioned in text. Then, for all the species detected, dictionaries are loaded and the names of genes are annotated. The third step applies filters to remove false positive gene names, such as names of gene families, diseases or names that are ambiguous with common English words (e.g., white). In the last step of the gene annotation, the remaining candidate genes are ranked to the respective gene mention using context profiles built from *Entrez Gene* and *UniProt* annotations. Figure 3.1 summarises the key idea behind the gene mention normalisation with *GNAT*.

*Gene
annotation*




For the task at hand, an *in-house* drug dictionary was utilized (Plake, 2010). A list of drugs and their synonyms was drawn from *DrugBank* and their identification in text is conducted with the use of regular expressions. Each drug along with its synonyms is represented by a regular expression that captures its occurrence in text, taking into consideration slight spelling or naming modifications, e.g., capitalization, different spellings of their chemical *IUPAC* (*International Union of Pure and Applied Chemistry*) name⁴. Table 3.1 provides some relevant examples. All regular expressions are compiled to a single *Labeled Deterministic Finite State Automaton* (*LDFA*). Each end state in the automaton stores the corresponding

*Drug
recognition*

⁴<http://www.iupac.org/>

Figure 3.1: Interspecies gene mention normalisation with *GNAT*

A gene encoding a putative human RNA helicase, *p54*, has been cloned and mapped to the band *q23.3* of chromosome 11. The predicted amino acid sequence shares a striking homology (75% identical) with the female germline-specific RNA helicase ME31B gene of *Drosophila*. Unlike ME31B, however, the new gene expresses an abundant transcript in a large number of adult tissues and its 5' non-coding region was found split in a *t(11;14)(q23.3;q32.3)* cell line from a diffuse large B-cell lymphoma.

		
EntrezGene ID: 1656 P54; RCK; HLR2 Species: <i>H. sapiens</i> Chromosome: 11q23.3 GO: RNA Helicase	EntrezGene ID: 2289 P54; FKBP51; PPlase Species: <i>H. sapiens</i> Chromosome: 6p21.3-2 GO: isomerase activity	EntrezGene ID: 42828 S4; dRpt2; p54; p56 Species: <i>D. melanogaster</i> Chromosome: 3R;95C13 GO: proteolysis

Gene *p54* has a profile of four contexts. Only the first context from the left reports the concepts *Human*, *RNA helicase* and *q23.3* chromosomal band, which are also found in text (from Hakenberg et al. (2008b)).

identifiers of all drug names that potentially end at this state. When parsing a text, a match with the *LDFA* immediately triggers the annotation of the matching phrase with all identifiers associated with the corresponding accept state. To deal with the false positives that result from ambiguous abbreviations (e.g., *ACC* for *Acetylcysteine* or *Adenoid Cystic Carcinoma*), the abbreviations are mapped to their long forms with the use of the algorithm introduced by Schwartz and Hearst (2003). To assess the efficacy of this dictionary towards the implementation of *CASSANDRA*, a random set of 60 *MEDLINE* records was generated out of the set of 22 million references utilized by the suggested methodology. The corresponding titles and abstracts were manually annotated. The respective dictionary achieves a precision of 88% and a recall of 93% on the identification of drug names from *DrugBank*.

As far as the recognition of *MeSH Disease* and *GO* terms in text is concerned, this was made through the usage of *GoPubMed* (Doms and Schroeder, 2005), a knowledge-based search engine that organizes *MEDLINE* references with *MeSH* and *GO* annotations. *GoPubMed* exploits the hierarchical structure of the ontologies and their word composition. It

*MeSH and GO
terms
identification*

Table 3.1: Drug representation with regular expressions - Examples

Drug	Synonyms	Representation
Fluocinonide	Fluocinonide Fluocinonido Fluocinonidum	<code>\b(Ff)luocinonid\w+</code>
2-Phosphoglycolic Acid		<code>(II 2 ii)[-]?[Pp]hosphoglycolic[-]?[Aa]cid</code>

The table shows how the regular expressions are formulated so as to catch both synonyms and the official name of the drug.

first finds matching seed ontological terms in text and then, it iteratively extends this set of terms to provide a full annotation for the respective *MEDLINE* abstract (Delfs et al., 2004). Figure 3.2 demonstrates how an abstract is annotated with *GoPubMed*. With respect to *MeSH*, only disease terms were considered. On the other hand, *GO* was fully used.

3.1.3 Profile generation for drugs and genes

Following the recognition of drugs, genes and ontological terms in text, we proceed with the automatic generation of context profiles for drugs and genes. This is a step of particular importance, since the ontological profiles constitute the means for the estimation of the relatedness between a drug and a gene and hence, the establishment of putative drug gene associations.

As it has been already mentioned, the context profiles consist of ontological terms derived from *GO* and *MeSH*. The context profiles are literature-based, meaning that they rely on the ontological terms that co-occur with drugs or genes in *MEDLINE* indexed abstracts and titles.

The profile generation for drugs (or genes) is divided in two separate steps;

- Quantification of the strength of the associations between drugs (or genes) and ontological terms,
- Exclusion of probable incidental associations and generation of the final context profiles for genes and drugs.

Figure 3.2: Annotation of MEDLINE abstracts with *GoPubMed*

powered by TRANSINGHT Enterprise Semantic Intelligence Server

help login

find

Cellular tumor antigen p53[protein] and biological_process[go]

show abstracts documents statistics top author clipboard share export

3,954 documents found

Smurf2 regulates the degradation of YY1.

Authors: Jeong, Hyung Min, Lee, Sung Ho, Yum, Jimah, Yeo, Chang-Yeol, Lee, Kwang Youl
Journal: Biochimica et biophysica acta (Biochim Biophys Acta), 2014

UNLABELLED: Transcription factor YY1 plays important roles in cell proliferation and differentiation. For example, YY1 represses the expression of muscle-specific genes and the degradation of YY1 is required for myocyte differentiation. The activity of YY1 can be regulated by various post-translational modifications, however, little is known about the regulatory mechanisms for YY1 degradation. In this report, we attempted to identify potential E3 ubiquitin ligases for YY1, and found that Smurf2 E3 ubiquitin ligase can negatively regulate YY1 protein level, but not mRNA level. Smurf2 interacted with YY1, induced the poly-ubiquitination of YY1 and shortened the half-life of YY1 protein. Conversely, an E3 ubiquitin ligase-defective mutant form of Smurf2 or knockdown of Smurf2 increased YY1 protein level. PPXY motif is a typical target recognition site for Smurf2, and the PPXY motif in YY1 was important for Smurf2 interaction and Smurf2-induced degradation of YY1 protein. In addition, Smurf2 reduced the YY1-mediated activation of an YY1-responsive reporter whereas Smurf2 knockdown increased it. Finally, Smurf2 relieved the suppression of p53 activity by YY1. Taken together, our results suggest a novel regulatory mechanism for YY1 function by Smurf2 in which the protein stability and transcriptional activity of YY1 is regulated by Smurf2 through the ubiquitin-proteasome-mediated degradation of YY1.

PubMed-4 24803334 Related Articles Read Full Text

Affiliation: College of Pharmacy and Research Institute of Drug Development, Chonnam National University, Gwangju 500-757, South Korea ...

Wikipedia: Gene activation, YY1 transcription factor, Ubiquitination, Ubiquitylation, E3 ligase, Ubiquitin-protein ligase, E3 Ubiquitin Ligase, 1 ...

refine search

- biological_process 3,954
- cell cycle arrest 591
- senescence 123
- recognition of apoptotic cell 114
- pathogenesis 230
- negative regulation of cyclin-dependent protein kinase activity 62
- DNA fragmentation during apoptosis 62
- stem cell development 86
- transduction 93
- negative regulation of caspase activity 4
- caspase activation via phosphorylation 4
- All of biological_process 3,954
- cellular process 3,047
- biological regulation 2,713
- developmental process 2,859
- metabolic process 1,791
- response to stimulus 1,061
- multicellular organismal process 1,4
- multi-organism process 1,797
- localization 511
- establishment of localization 336
- reproduction 3,954
- reproductive process 3,954
- sexual reproduction 402
- multicellular organism reproduction 336

The query “*Cellular tumor antigen p53[protein] AND biological_process[go]*” is submitted to *GoPubMed*. The platform returns all *MEDLINE* abstracts annotated with all *GOBiological Process* terms and with the protein *p53*. The query terms are highlighted in green color. The user can exploit the *MeSH* and *GO* hierarchy on the left and browse the results with additional terms, e.g., *reproductive process*.

Regarding the first step, the strength of each association between the drug (or gene) and the respective ontological term is computed based on the *Pointwise Mutual Information (PMI)*. *PMI* is a probabilistic measure used to assess the strength of word collocations in a text corpus (Manning and Schütze, 1999). In corpus linguistics, a collocation is a sequence of words that co-occur more often than it would be expected by chance. Analogously to the collocation definition, we extend the application of *PMI* to co-occurring pairs of drugs (genes) and ontological terms in *MEDLINE* abstracts and titles. Thus, the higher the *PMI* score, the lower the probability that the drug (gene) and ontological term co-occur by chance.

Compute the strength of associations

Let E represent a drug or a gene entity term and C represent a *MeSH Disease* or a *GO* term. We denote with n_E the number of documents where E occurs, n_C the number of documents where C occurs, and $n_{E,C}$ the number of documents where E and C co-occur. N denotes the number of documents that any E is found to co-occur with any C . *PMI* between any given E and C is then defined as shown in Equation 3.1. The higher the *PMI* score of the two terms E and C is, the more probable it becomes to observe these two terms together in the same document.

$$pmi(E, C) = \log \frac{N \times n_{E,C}}{n_E \times n_C} \quad (3.1)$$

However, the values that the *PMI* score can receive are not fragmented and can rather take any real value. For this reason, we adopt the *normalised PMI* (Bouma, 2009) (*nPMI*) that takes values between $[-1, +1]$. Equation 3.2 shows the definition of *nPMI* given any two terms E and C . If *nPMI* equals -1 , this means that there is no co-occurrence between E and C in the corpus. A negative value signifies that E and C co-occur less frequently than one would expect by chance. Conversely, a positive value indicates that the two terms co-occur more frequently than it would have been expected by chance, and a value of 1 shows complete co-occurrence between E and C . An *nPMI* score of 0.0 shows independence between E and C , meaning that the two terms co-occur exactly as frequently as it would have been expected by chance. For that reason, only associations assigned an *nPMI* score greater than 0.0 were considered meaningful and thus, the respective concepts C were included in the profiles of the entities E .

$$npmi(E, C) = \frac{pmi(E, C)}{-\log \frac{n_{E,C}}{N}} \quad (3.2)$$

To enhance the quality of the automatically generated ontological profiles, a further refinement of the terms that participate in a profile is applied. Hence, an ontological term is included in the context profile of a drug (or gene), if its *nPMI* score with the respective drug (gene) meets the following requirements.

*Associations
refinement*

Let A represent the set of ontological terms C that co-occur with an entity E . For all terms C that co-occur with the entity E , we compute the *nPMI* score. Then, we calculate the arithmetic *mean* of the respective *nPMI* distribution. We retain for the profile the terms C , where

$$npmi(E, C) \geq mean_A(npmi(E, C)) \quad (3.3)$$

This step constitutes a refinement of the ontological terms inside the profile. Besides this, an external refinement is applied with respect to the overall distribution of *nPMI* scores between any drug (gene) and any ontological term, where

$$p\text{-value}(npmi(E, C)) \leq 0.05 \quad (3.4)$$

$$npmi(E, C) \geq mean(npmi(E, C)) \quad (3.5)$$

among all the $npmi(E, C)$ scores between any entity E and any concept C .

Analytically, all the ontological terms which participate in a context profile, must have a statistically significant *nPMI* score with the respective drug or gene. Additionally, the *nPMI* score between a drug (or a gene) and an ontological term must be greater than the arithmetic *mean* of the *nPMI* scores distribution between any drug (or gene) and any ontological term.

3.1.4 Computation of semantic relatedness between the profiles

After having generated the context profiles for all drugs and genes found in *MEDLINE* abstracts and titles, in this step, the relatedness between drugs and genes is computed based on their profiles. Unlike the previous step, wherein *PMI* qualified the selection of the ontological terms that will participate in a profile, herein, the respective formula is used to estimate the statistical semantic similarity between a drug and a gene based on their context profiles.

An ontological profile can be viewed either as a consolidated set of ontological terms from both *GO* and *MeSH Disease*, or as a set of two individual subprofiles, each one including terms from one single ontology. To compute the relatedness between a drug and a gene profile, three scores of statistical semantic similarity are calculated and then combined into an overall score; one score corresponds to the similarity between the *MeSH Disease* profiles, one score corresponds to the similarity between the *GO* profiles and the third score corresponds to the similarity between the consolidated profiles that consist of both types of ontological terms.

*MeSH Disease,
GO and joined
profile
similarity*

Given one type of ontological profile, meaning *MeSH Disease*, *GO* or a consolidated profile, the computation of the relatedness between a drug and a gene is based on the *nPMI* values between all possible pairs of the ontological terms comprising the drug and gene profiles. Thus, for each drug-gene pair all the possible combinations between their profile terms are generated and the *nPMI* score for each such combination is computed. The computation is based on Equation 3.2.

More formally, let P_d the set of the profile terms for a drug d and P_g the set of the profile terms for a gene g . For every term pair $(C_d \in P_d, C_g \in P_g)$, the $npmi(C_d, C_g)$ is computed as shown in Equation 3.2.

Once all of the $npmi(C_d, C_g)$ scores between all possible pairs of the drug and the gene profile terms are computed, the scores are combined to produce the overall score between the drug and the gene. We compute the combination of scores following the methodology described in the work of Varlamis et al. (2004). The proposed measure has been used in the past

*Subprofile
similarity*

to estimate the similarity between two sets of ontological terms (Halkidi et al., 2003).

In detail, given P_d and P_g the drug and gene profile terms respectively, for each $C_d \in P_d$ the maximum $npmi(C_d, C_g)$ score is detected, and the average of all such maximum scores is computed. This is shown in Equation 3.6 as $S_1(d, g)$.

$$S_1(d, g) = \frac{1}{|P_d|} \sum_{C_d \in P_d} \max_{C_g \in P_g} npm_i(C_d, C_g) \quad (3.6)$$

Similarly, $S_2(g, d)$ is computed for all $C_g \in P_g$, the way it is shown in Equation 3.7.

$$S_2(g, d) = \frac{1}{|P_g|} \sum_{C_g \in P_g} \max_{C_d \in P_d} npm_i(C_g, C_d) \quad (3.7)$$

Finally, the two scores $S_1(d, g)$ and $S_2(g, d)$ are combined as shown in Equation 3.8 to produce the overall score between a drug d and a gene g .

$$Score(d, g) = \frac{1}{2}(S_1(d, g), S_2(g, d)) \quad (3.8)$$

Hence, to estimate the semantic relatedness between a drug and a gene, each ontological term from the drug's profile is paired with the ontological term from the gene's profile which gives the highest $nPMI$ score. At the end, only the combinations of terms that produce the highest scoring pairs are kept.

After computing the semantic relatedness for each type of subprofiles, the overall score of semantic relatedness between a drug d and a gene g , i.e., $SemRel(d, g)$, is calculated as follows:

*Overall score
computation*

$$SemRel(d, g) = 1 - [(1 - Score(d, g)_{go}) * (1 - Score(d, g)_{ds}) * (1 - Score(d, g)_{go, ds})] \quad (3.9)$$

where $Score(d, g)_{go}$, $Score(d, g)_{ds}$ and $Score(d, g)_{go, ds}$ is the semantic relatedness between the *GO*, *MeSH Disease* and consolidated profiles of a drug d and a gene g respectively.

The computation of the overall score is based on the *Noisy-OR* gate model. This distribution belongs to the family of models which is often referred to as *Independence of Causal Influences (ICI)*. It is used when there are several possible causes for an event, any of which can cause the event by itself with a certain probability (Zagorecki and Druzdzal, 2004). Correspondingly, each subprofile similarity score is viewed as a probability score that independently can cause the event of similarity between a drug and a gene.

Given a collection of drug-gene pairs, the suggested methodology assigns a score to every drug-gene pair. These pairs are then, in turn, ranked to suggest putative drug-gene associations. The higher the score is, the higher the semantic relatedness between the respective drug and gene.

3.2 Materials and Methods - Evaluation

3.2.1 Evaluation datasets

In the current section, the datasets utilized for the evaluation of the algorithm's efficacy are analysed. The datasets serve two different evaluation purposes:

- drug gene association prediction and,
- drug repositioning.

3.2.1.1 Evaluation datasets for drug gene association prediction

For the evaluation of the proposed algorithm towards drug gene association prediction, a series of datasets is necessary. The datasets should consist of true and false drug gene associations. The suggested algorithm assigns a semantic relatedness score to every drug-gene pair included in the dataset.

Once the scores are assigned, the drug-gene pairs are ranked and the performance of the algorithm is evaluated based on the prioritization of true over false drug gene associations.

In the current work, the performance of the algorithm is examined based on two main aspects of evaluation. One aspect pertains to the type of drug and gene profiles, and another one pertains to the computation of their semantic relatedness. The former involves the potential of the proposed methodology to utilize either co-occurrence based profiles retrieved from the biomedical literature or manually curated profiles. The question that arises is what is the impact of each profile type in the algorithm's performance. The latter involves the comparison of *nPMI* against traditional measures of semantic similarity. Is *nPMI* indeed the most efficient measure of semantic similarity for the task at hand? Both tasks account for the use of datasets that allow the efficient demonstration of the algorithms's performance differentiations.

*Two
evaluation
aspects -
(1)Profile type
(2)Semantic
similarity
metric*

Resources

Obtaining datasets of drug gene associations is a rather demanding task due to their limited number. Notably, there exists only one *Gold Standard* set of drug target interactions which was introduced in the work of Yamanishi et al. (2008). In this work, the authors formalized the drug target interaction inference as a supervised learning problem by combining chemical structure and genomic sequence information. For the evaluation of their approach, they characterised four classes of drug target interactions in humans involving *enzymes*, *ion channels*, *G-protein-coupled receptors (GPCRs)* and *nuclear receptors*. Then, they utilized that information to examine the performance of their methodology towards the prediction of drug target interactions. Their dataset has been, since then, extensively used as a benchmark by several supervised methodologies towards drug target interaction prediction (Pahikkala et al., 2014). In the current work, this dataset is also included in the performance evaluation of the proposed algorithm and we shall refer to it as the *Yamanishi* dataset.

*Few
benchmark
datasets*

To expand the evaluation of the suggested methodology, the absence of benchmark drug gene associations sets has to be tackled. On that account, a series of additional datasets are compiled based on sets of true

*DrugBank
dataset*

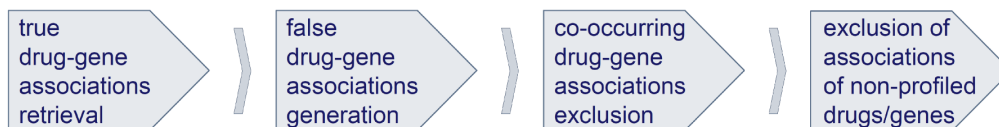
drug gene associations available in public pharmacogenomic databases. For that purpose, two popular drug related repositories were considered; *DrugBank* and the *Comparative Toxicogenomics Database (CTD)* (Davis et al., 2012). Both databases constitute freely accessible repositories of high-quality pharmacological data. As it has been mentioned before, *DrugBank* is both a bioinformatics and cheminformatics resource which combines detailed data descriptions and comprehensive target information for an extensive list of drugs. The fact that the dictionary used during the annotation of drugs in *MEDLINE* titles and abstracts was built based on this database, makes *DrugBank*'s drug target information ideal for the generation of a benchmark dataset.

CTD, on the other hand, takes one step further and, unlike traditional drug-related repositories which comprise mainly structural and chemical information of drugs along with a list of their physically binding targets, it includes formalised associations between drugs, genes and diseases. These associations are either curated or inferred. Notably, the curated associations are derived from *MEDLINE* articles and the respective textual evidence is also included in the database. Each set of diseases which are reported to be associated with either a drug or a gene in *CTD* can be considered as a manually curated profile that characterises the respective drug (or gene). This makes *CTD* an ideal candidate for the generation of a benchmark dataset on which the performance of the algorithm can be analysed when manually curated profiles are used. Most importantly, the intersection of drug gene associations provided by the suggested methodology and *CTD* form an additional dataset which enables the performance comparison of the co-occurrence based profiles against the manually curated profiles.

*CTD dataset -
manually
curated profiles
test*

Compilation of datasets

The compilation of the datasets is conducted as described in Figure 3.3. The datasets are compiled similarly to the so far unique benchmark dataset provided by Yamanishi et al. (2008). Indicatively, for each of the sets there exist both true and false examples of drug-gene interactions. The first step is to extract the positive pairs (true interactions) listed in each

Figure 3.3: Drug gene association datasets compilation

The pipeline for the compilation of the datasets used to evaluate CASSANDRA is shown. The set of *true* drug gene associations is retrieved from each resource. All drugs and genes are combined to generate the *false* drug gene pairs set. Then, all drug gene pairs wherein the drug and the gene co-occur in at least one *MEDLINE* reference are removed. Lastly, all drug gene pairs containing a non-profiled drug or gene are also excluded from each dataset.

database. These pairs are then, in turn, used for the generation of the negative examples (false interactions).

More precisely, each drug is paired with every gene and the resulting pairwise combinations constitute the so-called *false* drug gene associations. These associations are considered false based on the fact that they are not reported by the respective pharmacogenomic resource, although in actual fact they may as well constitute true drug gene associations, which have not yet been confirmed experimentally. However, since there is no information regarding non-interacting drug targets the above convention is necessary for the compilation of the evaluation datasets. Still, any *false* drug gene association scored highly by the proposed algorithm constitutes an *in-silico* prediction of a putative drug target interaction.

Unknown non-interacting drug-gene pairs

Two additional steps follow the generation of false drug gene associations. The former is to exclude from the evaluation dataset all these drug gene pairs for which there is at least one co-occurrence in a *MEDLINE* abstract or title. This is due to the fact that the goal of this study is to present a methodology that is able to indirectly pinpoint the similarity between a drug and a gene. On the premise that their co-occurrence is itself a signal of putative relation, these drug gene pairs are excluded, even in the case their *nPMI* score is negative. Indeed, including drug gene pairs of co-occurring drugs and genes in the evaluation is shown to introduce a positive bias in the algorithm's performance and hence, conceal the algorithm's true efficacy (see Chapter 4, Section 4.3.1.1).

Positive bias of co-occurring drugs and genes

Given the fact that not all drugs (genes) have profiles generated, the final step is to exclude from the datasets the drug gene pairs wherein either the drug or the gene has an empty profile. Two are the reasons for which a drug or a gene is assigned an empty profile; either the annotators reported no occurrence of the respective entity in *MEDLINE* abstracts and titles, or the ontological terms occurring with the entity do not participate in the profile due to their unfitting *nPMI* score as this has been defined in Section 3.1.3. Pairs of empty profiles result in zero scored drug gene associations that when included in the evaluation falsely enhance the performance of the suggested algorithm. Markedly, such is the case of the *DrugBank* dataset; 41 % of the false drug gene pairs generated after the exclusion of non-co-occurring drugs and genes, consist of non-profiled entities. When scored with zero, these pairs boost the performance of the suggested algorithm but, on the other hand, they obfuscate its true discriminative power. Excluding them eliminates the bias that they introduce (see Table 4.6).

*Positive bias
of non-profiled
drugs and
genes*

3.2.1.2 Evaluation sets for drug repositioning

Towards demonstrating the application of the method in identifying candidate drug repositioning cases, a set of all known drugs that have been repositioned has to be compiled. Ideally, the set should include along with the drug identifiers, the old indication of each drug and the new indication. However, there is no such dataset publicly available. Drug repositioning cases are scattered across the literature. For that reason, the literature is systematically mined for the manual generation of a dataset comprising drug repositioning data. The set includes drugs for which there exists a *DrugBank* identifier. Apart from *MEDLINE* abstracts, information is extracted from the *FDA*, *Wikipedia* and other web resources comprising drug

*No available
drug
repositioning
dataset*

related data ⁵. Old and new indications are reported along with the year of approval of each drug's new indication.

For the task at hand, we focus on drugs that were *FDA* approved within the last 5 years with their new indications and for which the proposed algorithm has generated co-occurrence based profiles out of biomedical literature. Thus, all *MEDLINE* data from 2009 and on was excluded from the application of the suggested methodology. The application of the method in identifying candidate drug repositioning cases is conducted as follows: the drug's profile is generated, and the *MeSH disease* terms that participate in the profile are examined. The efficacy of the approach can then be assessed on whether the new indication is included in the drug's profile, and if so, whether it is ranked high in the list of the drug's profile terms.

3.2.2 Alternative semantic similarity metrics

CASSANDRA utilizes the statistical measure *nPMI* to quantify the semantic similarity between a drug and a gene profile. To assess the *nPMI*'s efficacy, the suggested measure is compared with two traditional metrics of semantic similarity; the *Wu-Palmer* (1994) and *Lin* (1998) measures.

Wu and Palmer

For two concepts C_1 and C_2 and their *LCA Least Common Ancestor* C_3 , the *Wu-Palmer* semantic similarity is defined as follows:

⁵

<http://www.cancer.gov/>

<http://www.centerwatch.com/>

<http://www.drugs.com>

<http://www.medicalnewstoday.com/>

<http://www.medscape.com/>

<http://www.webmd.com/>

$$WP(C_1, C_2) = \frac{2 \times \text{depth}(C_3)}{2 \times \text{depth}(C_3) + \text{distance}(C_1, C_3) + \text{distance}(C_2, C_3)} \quad (3.10)$$

, where $\text{depth}(C_3) = \text{distance}(C_3, \text{root})$

For two ontological concepts, the *Wu-Palmer* metric estimates their semantic relatedness by calculating the distance of the shortest path that connects them in the ontology. The lower the distance between the concepts and their *LCA* and the higher the distance of *LCA* from the root of the ontology, the greater the similarity of these concepts.

Lin

Let two concepts C_1 and C_2 and their *LCA Least Common Ancestor* C_3 . The *Information Content* of C_1 is defined as:

$$IC(C_1) = -\log P(C_1) \quad (3.11)$$

, where $P(C_1)$ is the probability of occurrence of the concept C_1 in the corpus. The higher the probability of a concept's occurrence, the lower the *IC* of the concept is.

The *Lin* semantic similarity is, then, calculated as follows:

$$\text{Lin}(C_1, C_2) = \frac{1}{IC(C_1) + IC(C_2) - 2 \times IC(C_3)} \quad (3.12)$$

As shown by the equation above, the *Lin* metric takes into account the hierarchy of the ontology, but also considers the *Information Content (IC)* of the concepts in the calculation of semantic relatedness. Thus, it is both a probabilistic and structural metric of semantic similarity.

Chapter 4

Results

This Chapter describes the tasks that were conducted to evaluate the efficiency of CASSANDRA towards drug gene association prediction. More specifically, this chapter reports

- The occurrence and co-occurrence statistics of terms in abstracts and titles of *MEDLINE* indexed articles.
- The statistics of the drug gene association datasets used for the algorithm's evaluation.
- Analysis of the algorithm's performance towards drug gene association prediction.
 - How does the type of ontological profiles affect the algorithm's prediction efficacy? Are manually curated profiles better than co-occurrence based profiles?
 - Is *nPMI* the most appropriate measure of semantic similarity?
- Analysis of the algorithm's performance towards drug repurposing. Do the profiles include the new indications?
- Manual analysis of three drug gene associations proposed by CASSANDRA.

4.1 *MEDLINE* statistics

Overall, 23,487,871 *MEDLINE* abstracts and titles that were available at the time of the algorithm’s implementation are considered. Table 4.1 shows the number of unique drugs, genes, *MeSH Diseases* and *GO* terms found in these *MEDLINE* abstracts and titles. Additionally, the table shows the number of documents that contain at least one term of interest. Notably, the numbers reported indicate the excess of information available in *MEDLINE* abstracts and titles and motivates the exploitation of the respective data.

Table 4.1: Statistics of term annotations in *MEDLINE*

Term	Number	Documents	
drug	2,909	4,599,847	
gene	58,261	3,512,899	
<i>Gene Ontology</i>	20,255	<i>Biological Process</i>	8,602,996
		<i>Molecular Function</i>	4,550,929
		<i>Cellular Component</i>	3,036,912
<i>MeSH Disease</i>	4,194	13,242,432	

Table shows the considered term types, along with the number of the unique terms recognised. The final column shows the number of the *MEDLINE* titles and abstracts containing at least one annotation of the respective term type. Evidently, the number of terms and documents indicates the excessive information available.

The same holds for the co-occurrences of drugs and genes with *MeSH Diseases* and *GO* terms. As shown in Table 4.2, the co-occurrences are plentiful and hence they can be used for the contextual description of drugs and genes. In particular, the vast majority of drugs and genes co-occur with up to 100 *MeSH Disease* and 300 *GO* terms (~ 100 terms of each subontology) (see Figure 4.1). Limited are the drugs and genes that co-occur with more than 1000 terms. Evidently, the necessity to quantify the strength of each co-occurrence rises. Not every *MeSH Disease* or *GO* term co-occurring with either a drug or a gene can be included in the respective profile, and thus this motivates the use of *nPMI*. As it has been mentioned in Section 3.1.3, the higher the *nPMI* score, the lower the probability that a drug (or

gene) and an ontological term co-occur by chance. The additional filtering steps enhance the quality of the profiles.

Table 4.2: Co-occurrences for drugs, genes and ontological terms in *MEDLINE*

Entity	Ontological Term	Documents
drug	<i>MeSH Disease</i>	2, 828, 259
	<i>GO biological process</i>	2, 644, 296
	<i>GO molecular function</i>	1, 747, 743
	<i>GO cellular component</i>	1, 001, 228
gene	<i>MeSH Disease</i>	2, 360, 673
	<i>GO biological process</i>	2, 439, 335
	<i>GO molecular function</i>	2, 030, 772
	<i>GO cellular component</i>	1, 121, 913

The table shows the number of *MEDLINE* abstracts containing at least one co-occurrence of a drug or a gene with a *MeSH disease* or a *GO* term.

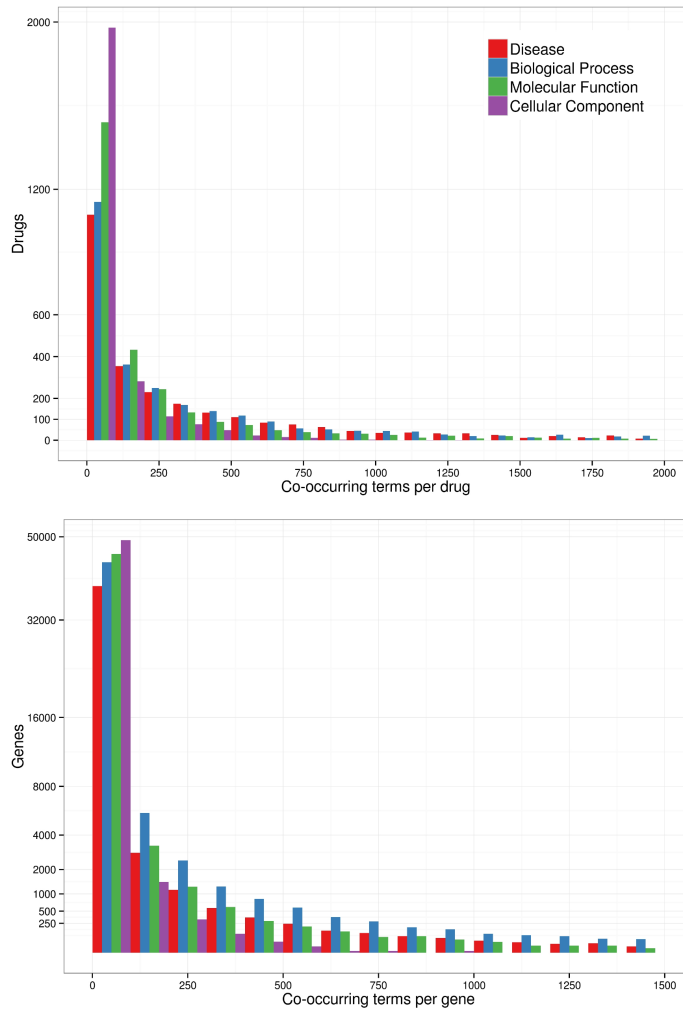
The co-occurrences between ontological terms in *MEDLINE* abstracts are also quantified. The results are shown in Table 4.3. Evidently, the number of documents containing at least one co-occurrence of two distinct *MeSH Disease* or *GO* concepts is significantly large. This suggests that *nPMI* can be used as a measure of semantic similarity between two ontological terms.

Table 4.3: Ontological terms co-occurrences in *MEDLINE*

Ontological Term	Ontological Term	Documents
<i>MeSH Disease</i>	<i>MeSH disease</i>	9, 065, 044
<i>Gene Ontology</i>	<i>Gene Ontology</i>	6, 885, 166
<i>MeSH Disease</i>	<i>Gene Ontology</i>	6, 320, 168

MEDLINE abstracts and titles containing at least one co-occurrence between two *MeSH Disease* or *GO* concepts. The number of co-occurrences is significantly large and this motivates the use of *nPMI* as a measure of semantic similarity between the ontological terms.

Figure 4.1: Graphical representation of co-occurrences for drugs and genes



The vast majority of drugs and genes co-occur with up to 100 *MeSH Disease* and 300 *Gene Ontology* terms (~ 100 terms of each subontology). Limited are the drugs and genes that co-occur with more than 1000 terms. Notably, not every term can be included in a drug/gene profile. The strength of co-occurrences has to be quantified and the meaningless co-occurrences have to be excluded.

4.2 Evaluation datasets statistics

In this subsection, the statistics of the datasets used for the evaluation of the algorithm's performance are provided. The resources and compilation pipeline utilized for the generation of the evaluation datasets are analytically described in Section 3.2.1.1. Table 4.4 gives an overview of the resulting datasets and reports the number of drugs, genes and associations (true and false) per dataset. It also reports the evaluation aspect for which each dataset is considered.

The *DrugBank* dataset splits in two subsets; the *Approved* and the *Experimental*. The former includes drugs that have been already approved and entered the pharmaceutical market, whilst the latter includes experimental compounds. The discrimination between these two datasets demonstrates whether the type of a drug affects the performance of the suggested algorithm or not. The datasets *DrugBank_{Approved}*, *DrugBank_{Experimental}* and *Yamanishi* are used for the evaluation of the proposed methodology when co-occurrence based profiles are considered.

Approved &
Experimental

Analogously, the *CTD_{Binding}* and *CTD_{Related}* sets are used to demonstrate the potential of the suggested methodology to propose drug gene associations when manually curated profiles are considered. As shown in Table 4.4 *CTD* also splits in two subsets. This is due to the content of the database itself. *CTD* includes two types of associations between drugs and genes; the associations wherein the drug physically binds to a product of the respective gene and the associations wherein the drug affects the regulatory processes of a gene or one of its products. Hence, the *CTD* datasets are accordingly compiled and form the *Binding* and the *Related* subsets respectively. In the case of *CTD*, the drugs are stored in the repository with their *MeSH* identifier from the *MeSH* tree *Chemicals and Drugs*. Seeing that not all *MeSH* identifiers correspond to *DrugBank* identifiers and, hence, mapping them would result to the loss of drug gene associations provided by *CTD*, the *CTD* identifiers are retained.

Binding &
Related

Finally, the *CTD_{cb&mc}* set is the intersection between the drug gene associations that have been identified and scored by the suggested algorithm in *MEDLINE* abstracts and titles and the drug gene associations provided by *CTD*. Consequently, all drug gene associations included in *CTD_{cb&mc}*

The
comparison
dataset

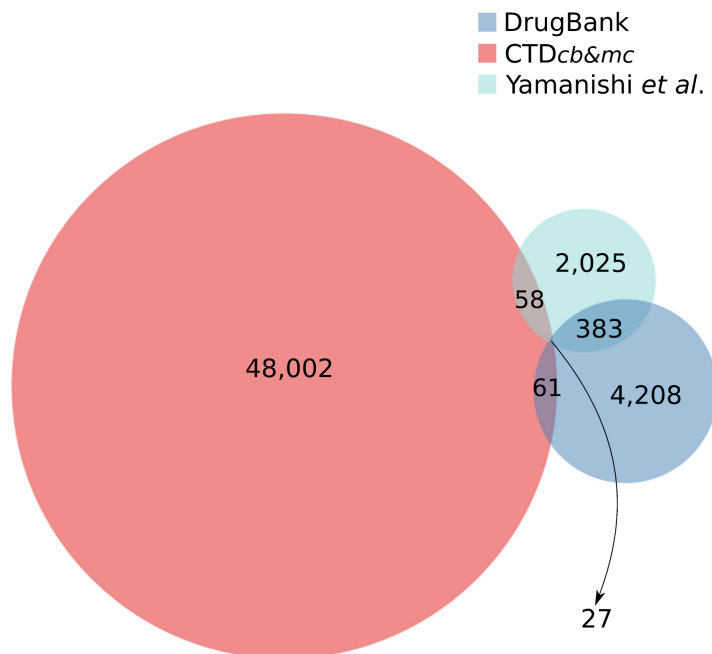
Table 4.4: Evaluation dataset statistics

	drugs	genes	true pairs	false pairs	evaluation type
<i>DrugBank_{Approved}</i>	751	659	1, 836		
<i>DrugBank_{Experimental}</i>	790	1, 732	2, 843	7, 660, 034	<i>Co-occurrence based profiles</i>
<i>Yamanishi</i>	302	657	2, 493	486, 904	
<i>CTD_{binding}</i>	955	633	2, 260		<i>Manually curated profiles</i>
<i>CTD_{related}</i>	4, 854	5, 983	203, 949	29, 195, 261	
<i>CTD_{cb&mc}</i>	915	5, 465	48, 148	5, 936, 324	<i>Co-occurrence based vs. manually curated profiles</i>

The table shows the number of drugs, genes and associations considered in each evaluation dataset. *DrugBank* true drug gene associations split into *Approved* and *Experimental* subsets based on the type of the drug. *CTD* true drug gene associations split into *Binding* (i.e., physically interacting) or *Related* associations. For the validation of the method when co-occurrence based profiles are considered, the *DrugBank* and *Yamanishi* datasets are utilized. For manually curated profiles, the *CTD Binding* and *Related* subsets are used. To compare the algorithm's performance when either manually curated or co-occurrence based profiles are considered, the dataset *CTD_{cb&mc}* is used.

consist of drugs and genes that they have been assigned both manually curated profiles by the *CTD* curators and co-occurrence based profiles by the proposed algorithm. To compile this dataset, for all *CTD* drug gene associations the drugs were, this time, mapped to their *DrugBank* identifiers. The resulting drug gene associations were then compared with the drug gene associations generated by the suggested algorithm after traversing the literature and then, the intersection of these sets was retained. This dataset is of particular importance as it enables the comparison between co-occurrence based profiles that are derived from literature and manually curated profiles. The subscript *cb&mc* stands for the *Co-occurrence Based & Manually Curated* and helps to discriminate $CTD_{cb\&mc}$ from the rest of the datasets derived from *CTD*.

Figure 4.2: True drug gene associations *Venn* Diagram

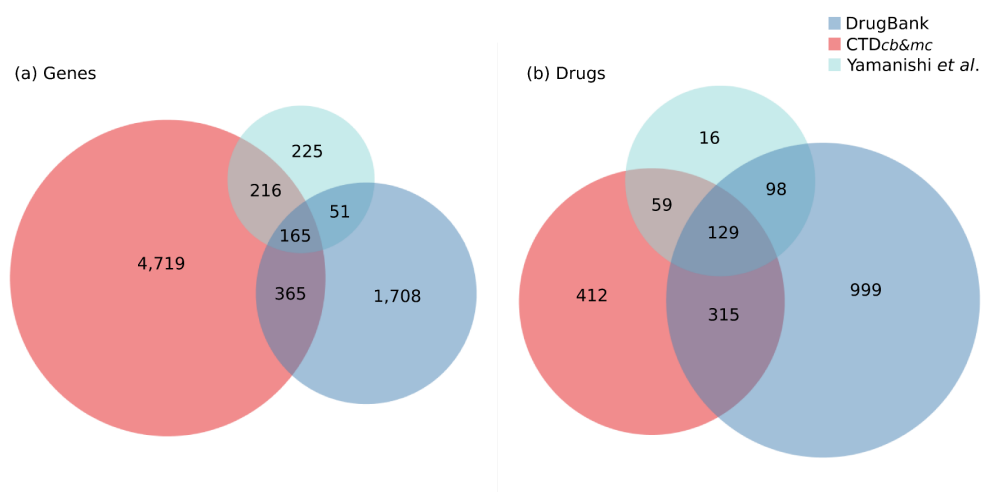


An overview of the datasets used for the different evaluation aspects of the suggested methodology. The *Venn* Diagram illustrates the overlap between the true drug gene associations included in the three datasets. The significantly small overlap indicates that the datasets are substantially differentiated.

Figures 4.2 and 4.3 illustrate an overview of the datasets used for the evaluation. The *Venn* Diagrams show the overlap between the *DrugBank*, *Yamanishi* and *CTD_{cb&mc}* datasets. As shown in Figure 4.2 the datasets are significantly differentiated in terms of the true drug gene associations they comprise. Indeed, all three of them have an overlap of only 27 drug gene associations. They also differ in size. The explanation for that is straightforward. The vast majority of drug gene associations in *CTD_{cb&mc}* are rather *related* than directly *binding* associations. On the other hand, the *DrugBank* and *Yamanishi* sets contain drug gene associations where the drug is proven to interact with the protein coded by the respective gene. Figure 4.3 shows the overlap of the datasets in terms of drugs and genes. As shown, the overlap in this case is again relatively small and the size of the sets again varies. The high differentiation degree of the datasets poses two advantages; first, it allows the comprehensive comparison of the suggested algorithm with alternative implementations of the same methodology and second, it demonstrates the robustness of the algorithms when inferring putative drug gene associations.

Differentiated datasets

Figure 4.3: *Venn* Diagrams for drugs and genes in the evaluation datasets



An overview of the datasets used for the different evaluation aspects of the suggested methodology. The drug and gene *Venn* Diagrams illustrate the overlap between the three datasets with respect to the drugs and genes they include. Similarly to the *Venn* Diagram of *true* drug gene associations, the datasets show a substantial differentiation.

4.3 Drug gene association prediction

In this section, the results of the algorithm's efficacy towards drug gene association prediction are reported. As it has been mentioned before, the presented method utilizes the co-occurrences of *GO* and *MeSH Disease* concepts with drugs and genes in *MEDLINE* titles and abstracts. Based on that co-occurrence information, profiles of ontological terms are created for both drugs and genes. Then, with the help of a corpus-based statistical measure, *nPMI*, the drugs are associated to genes by assessing the semantic relatedness of their profiles. Then the proposed drug gene associations are prioritized based on their relatedness score. For details regarding the steps of the algorithm, please refer to Chapter 2 (*Background*).

The suggested algorithm was applied on a series of evaluation datasets (for details regarding the datasets compilation and statistics, please refer to Sections 3.2.1.1 and 4.2 respectively). As shown, the datasets consist of true and false drug gene pairs. The proposed algorithm assigns a score of semantic relatedness to every drug gene pair included in these datasets and the drug gene pairs are, afterwards, ranked according to that score. To demonstrate the algorithm's efficacy towards the prioritization of true drug gene associations, the respective *Receiver Operating Characteristic (ROC)* curves, *Area Under the Curve (AUC)* values and *Precision-Recall (PR)* curves are provided.

ROC curves have been extensively used towards the evaluation of binary decision algorithms (Bandos et al., 2010). In a binary decision problem, the classifier labels input examples as either positive or negative. A *ROC* curve plots the *True Positive Rate (TPR)* on the *x*-axis and the *False Positive Rate (FPR)* on the *y*-axis for different cut-off points. Each point on the *ROC* curve, meaning a *TPR-FPR* pair represents the fraction of positive examples that have been correctly labeled with respect to the fraction of negative examples that are misclassified as positive, corresponding to a particular decision threshold. An algorithm with no overlap between the two distributions, hence with a *TPR* of 1.0 and a *FPR* of 0.0, succeeds a perfect discrimination between the class of the positive and the class of the negative examples. Therefore, the closer the *ROC* curve is to the upper

ROC curves

left corner, the higher the overall accuracy of the algorithm (Zweig and Campbell, 1993).

To evaluate the performance of the proposed method, in tandem with the *ROC* curves, the respective *Area Under the Curve (AUC)* values are provided. The *AUC* of a classifier is basically the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance (Fawcett, 2006). In this work, *ROC* curves and *AUC* scores are used in the notion of the performance evaluation applied in the field of *Information Retrieval* (Manning et al., 2008). More precisely, the usage of *ROC* curves illustrates the ability of the method to prioritize the positive over the negative examples, having as an input only a ranked list of examples. Herein, these examples are simply the scored drug gene associations that correspond to each evaluation dataset. Given a particular dataset, the probability of a randomly chosen positive drug gene association to rank higher than a randomly chosen negative drug gene association is represented by the corresponding *AUC* value.

AUC values

Besides *ROC* curves which are very insightful performance representations in the case of binary classifications, *PR* curves are also utilized to illustrate the efficacy of the algorithm. Unlike *ROC* curves, *PR* curves show the ratio of true positives among all the predicted positives under a given recall rate. It has been shown that a classification method dominates the *ROC* space if and only if it dominates the *PR* space (Davis and Goadrich, 2006). *PR* curves are particularly informative and biologically meaningful in the case of imbalanced datasets (Chen et al., 2012). In an effort to give an overall picture of CASSANDRA's efficacy, the *PR* curves in varying degrees of imbalance between true and false examples (i.e., 1:1, 1:2, 1:4 and 1:8) are also provided.

*PR curves
contribution*

Another insightful performance metric is *Specificity* (or *True Negative Rate*) which measures the proportion of negatives that are correctly identified as such. Apparently, to measure *Specificity*, a clear indication of what a true negative example is becomes necessary. However, in the current evaluation the datasets contain clear signals of what are the positive examples, but it is not possible to accurately assess which are the true negative examples (and hence compute the *Specificity*) due to the absence of supporting experimental evidence (as shown in Figure 3.3, the false examples were

generated from the pairwise combination of all drugs and genes included in the respective drug related repositories). The computation of *Specificity* remains an open problem in the studies of drug-target prediction (Pahikkala et al., 2014).

The evaluation of the algorithm towards drug gene association prediction is divided in two main parts; the first part examines the performance of the algorithm with respect to the type of terms that constitute the drug and gene profiles. Both manually curated and co-occurrence based profiles are utilized and the respective *ROC* curves, *AUC* values and *PR* curves are provided. In the second part, alternative measures of semantic similarity are explored and compared to the statistical semantic similarity metric *nPMI*. Along with the *ROC* curves, *AUC* values and *PR* curves, the discriminative power of each measure is also considered to assess in what degree the positive drug gene associations differentiate from the negative drug gene associations.

*Two main
evaluation
parts*

4.3.1 Performance evaluation - Ontological profiles

As it has been mentioned before, the proposed methodology utilizes ontological profiles in order to assess the semantic similarity between a drug and a gene. The ontological profiles constitute the literature fingerprints of drugs and genes, hence assessing their quality is of primary importance towards the establishment of drug gene association predictions. In this part of the evaluation, the role of ontological profiles is thoroughly examined.

In principle, the suggested algorithm utilizes co-occurrence based profiles derived from biomedical literature. However, one main advantage of the proposed pipeline is that it can also compute the semantic relatedness between a drug and a gene in case the respective manually curated profiles are available. Consequently, the arising question is: Which profile type is the most suitable in terms of the algorithm's performance?

Thus, three evaluation aspects are considered for this task:

- Performance evaluation when co-occurrence based profiles are used.
- Performance evaluation when manually curated profiles are used.

- Comparison of performance between manually curated and co-occurrence based profiles.

4.3.1.1 Co-occurrence based profiles

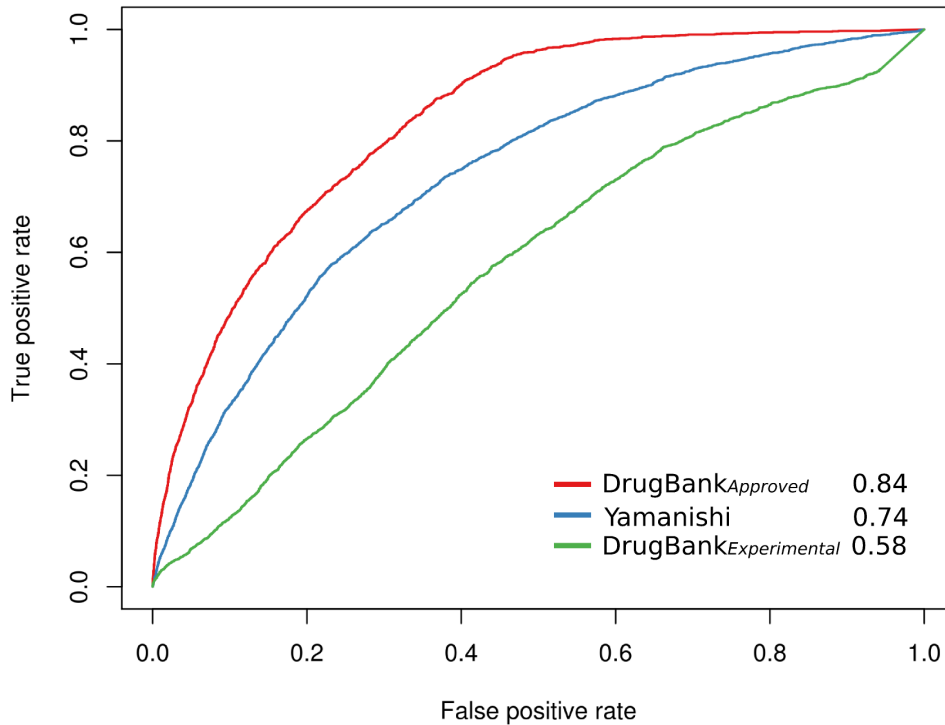
With regards to this evaluation part, the *DrugBank* and the *Yamanishi* dataset are utilized. Figure 4.4 demonstrates the performance of the algorithm. The suggested method obtained an *AUC* of 0.84 for the *DrugBank* drug gene associations in which *Approved* drugs participate and 0.58 when *Experimental* drugs participate. Considering the reported results for the *Yamanishi* dataset (an *AUC* of 0.74 was obtained), and taking into account its small overlap with the *DrugBank* datasets (as shown in Figure 4.2, the *DrugBank* and the *Yamanishi* datasets overlap only by 6% in terms of true drug gene associations), the value of the *AUC* (0.74) suggests the robustness of the suggested methodology.

As it is demonstrated in Figure 4.4, the reported results suggest that the *AUC* for the *Approved* drugs is higher than the *AUC* for the *Experimental* drugs. To understand the reason for this difference, the number of literature references for both types of drugs is examined. Approximately, 87% of the papers with at least one drug occurrence mention an *Approved* drug, while only 25% of the papers mention an *Experimental* drug. The underrepresentation of *Experimental* drugs in literature results in poor profiles for the respective type of drugs. Indeed, the average number of concepts in the profile of an *Experimental* drug is 273, while for an *Approved* drug is 699 (see Table 4.5).

Accordingly, when drug gene associations wherein *Human* genes participate are considered, the *AUC* values increase for both *Approved* and *Experimental* drugs to 0.88 and 0.77 respectively, as shown in Figure 4.5. This is because, *Human* genes are discussed more in literature than the genes that belong to other species. Altogether, genes from 31 species were annotated in *MEDLINE* abstracts and titles. Approximately, 70% of the papers with at least only one gene occurrence mention a *Human* gene, while 38% of the papers mention genes that belong to the rest of 30 other species. The average number of concepts in the profile of a *Human* gene is 451 and it is significantly higher than that of a gene which belongs to other species (i.e.,

Why better performance for Approved drugs?

Why better performance for Human genes?

Figure 4.4: ROC curves for co-occurrence based profiles

ROC curves for datasets when co-occurrence based profiles are utilized. The curves show that CASSANDRA is robust and predicts well if there is sufficient underlying data (*Approved* and *Yamanishi*). For *Experimental* drugs little is published and hence the method performs worse.

61). This explains the improvement in the performance of the proposed algorithm when applied on the respective subset.

The above results suggest that the method under evaluation successfully interrelates drugs and genes even when these are not co-mentioned in text. Clearly, the amount of literature references plays a significant role towards the establishment of reliable profiles for both drugs and genes and the computation of their semantic relatedness.

The impact of empty-profiled and co-occurring entities

At this point, two important steps towards the assessment of the algorithm's efficacy have to be pinpointed; the exclusion from the evaluation

Table 4.5: *Arithmetic mean* of statistically significant concepts in profiles

Entity	Type	<i>Disease</i>	<i>GO</i>	All	Documents (%)
drug	<i>Approved</i>	239	461	699	87
	<i>Experimental</i>	58	215	272	25
gene	<i>Human</i>	161	290	451	70
	<i>non-Human</i>	17	44	61	38

The table reports the average number of statistically significant concepts that are included in the profiles of *Approved/Experimental* drugs and *Human/non-Human* genes. The percentage of documents that include at least one occurrence of the each entity are also provided. The underrepresentation of the *Experimental* drugs and *non-Human* genes in literature results in poor profiles. This affects the algorithm’s performance on the respective datasets.

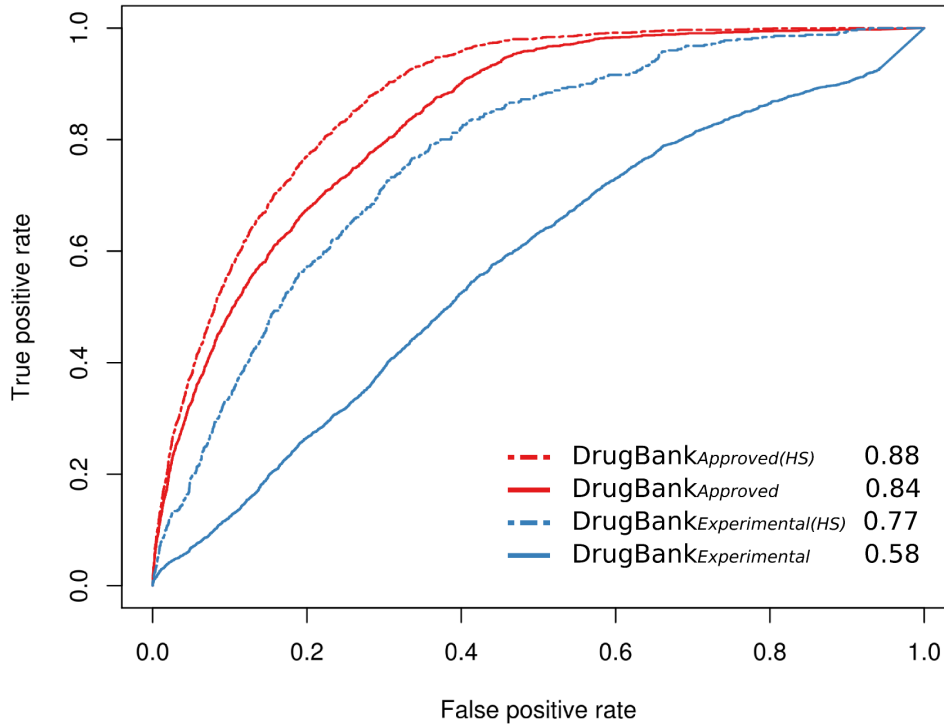
datasets of drug gene pairs that constitute by a drug or a gene with an empty profile and the exclusion of drug gene pairs wherein the drug is found to occur with the gene in at least one *MEDLINE* publication (see Figure 3.3). These steps have a significant impact in the demonstration of the algorithm’s performance.

More specifically, when drugs and genes with an empty profile are considered, the *AUC* values obtained for the datasets *DrugBank_{Approved}*, *DrugBank_{Experimental}* and *Yamanishi* are 0.89, 0.50 and 0.74 respectively. This effect can be explained when considering the statistics reported in Table 4.6. When a drug gene pair constitutes of an empty-profiled entity, the association score between the drug and the gene equals to 0.0. Thus, the inclusion of such pairs in the evaluation datasets signifies the increase in the number of zero-scored drug gene pairs.

As shown in the Table, 41% of the initial set of false drug gene associations consist of an empty-profiled drug or gene, hence 41% of the drug gene pairs have an association score of 0.0. On the other hand, the majority of the true drug gene pairs included in the *DrugBank_{Approved}* dataset are scored highly (see upper left corner of Figure 4.12). Consequently, the classification task in this case is facilitated and the algorithm obtains an *AUC* of 0.89 (instead of the previously reported 0.84).

The impact of empty-profiled entities

The better get better

Figure 4.5: ROC curves for *Human* and non-*Human* genes

Performance evaluation of associations between drugs and *Human* genes. The curves show that CASSANDRA performs better when drug gene pairs include *Human* genes. For *Human* genes there is a lot published and that has a beneficial impact on the method's performance.

However, the situation differs for the dataset that consists of *Experimental* drugs. The percentage of empty profiled pairs for both true and false drug gene associations is similar (46% and 41% respectively). At this point, it has to be noted that the score distributions between the true and the false drug gene associations contained in this dataset are already quite similar (see 1st row, 2nd column in Figure 4.12). When considering the empty-profiled pairs in the evaluation, the distributions become even more alike and this results in the drop of the algorithm's classification performance. The *AUC* obtained is 0.50 (i.e., random classification). On the other hand, the performance of the algorithm in the *Yamanishi* dataset remains stable to an *AUC* of 0.74. This is no surprise, if we take into account that less than 1.0% of the drug gene pairs constituting both the initial set of true

The worse get worse

Table 4.6: *AUC* values when including non-profiled drugs and genes

	% empty-profiled		<i>AUC</i>
	true pairs	false pairs	
<i>DrugBank_{Approved}</i>	0.01	41	0.89
<i>DrugBank_{Experimental}</i>	46		0.50
<i>Yamanishi</i>	0.02	0.4	0.74

The table reports the percentage of empty-profiled drug gene pairs contained in the initial evaluation datasets and the respective *AUC* values. For the *DrugBank_{Approved}* dataset, the performance improves due to the high association score of true drug gene pairs and the high percentage of empty-profiled false drug gene pairs. In the *DrugBank_{Experimental}* dataset, true and false drug gene pairs contain similar percentages of empty-profiled drug gene pairs, hence, the *AUC* value drops to 0.50. In the *Yamanishi* dataset the small amount of empty profiled drug gene pairs hardly affects the classification task.

and false drug gene pairs of the *Yamanishi* dataset consist of empty profiled drug gene pairs.

Similar is the case when the evaluation datasets include drug gene pairs that consist of drugs and genes which co-occur in at least one *MEDLINE* record. For example, in the case of the *DrugBank_{Approved}* dataset, such pairs further boost the algorithm’s performance by 4.0% (this time the algorithm obtains an *AUC* value of 0.91). However, the goal of this study is to present a methodology that is able to indirectly pinpoint the similarity between a drug and a gene.

Consequently, to elucidate CASSANDRA’s true classification efficacy, the drug gene pairs that consist of empty-profiled or co-occurring drugs and genes are excluded from the evaluation datasets. This way the method remains free from positive bias that would render the proposed drug gene association prediction method overoptimistic.

*The impact of
co-occurring
entities*

4.3.1.2 Manually curated profiles

The suggested algorithm utilizes co-occurrence based profiles of ontological concepts derived from the biomedical literature towards the assessment of latent associations between drugs and genes. These profiles can be viewed as feature vectors for drugs and genes. The methodology can incorporate any such feature vectors as long as they consist of *MeSH Disease* and *GO* terms. Ideal is the case of acquiring a set of manually curated ontological terms that are found and suggested by curators to be associated with a drug or a gene. On that premise, could CASSANDRA successfully predict drug gene associations? To answer this question, *CTD* is utilized.

As it has been aforementioned, *CTD* is a publicly available biomedical repository, which unlike to other resources that comprise pharmacological information, it also contains formalized relations between drugs, diseases and genes. These relations are being manually curated in a regular basis, so that *CTD* constitutes a reliable and up-to-date pharmacological repository. For this evaluation step, the drug-*MeSH Disease* and the gene-*MeSH Disease* relations provided by *CTD* are utilized to define the manually curated profiles for drugs and genes respectively. The relations between a *MeSH Disease* and a drug or a gene can be either of therapeutic nature (the application of the drug or targetting the respective gene or its products has a beneficial effect towards the treatment of the respective disease) or of a causal/metabolism-pertaining nature.

Similarly, the evaluation datasets comprising true and false drug gene pairs are also derived from *CTD*. As it has been mentioned in Section 3.2.1, *CTD* comprises two types of drug gene associations. Those that represent a physical binding between a drug and a gene (or its products) and those that represent the impact of the drug on the gene's regulatory processes.

For instance, an example of a *Related* association would be:

Thalidomide results in **increased activity** of *ABCB1* protein.

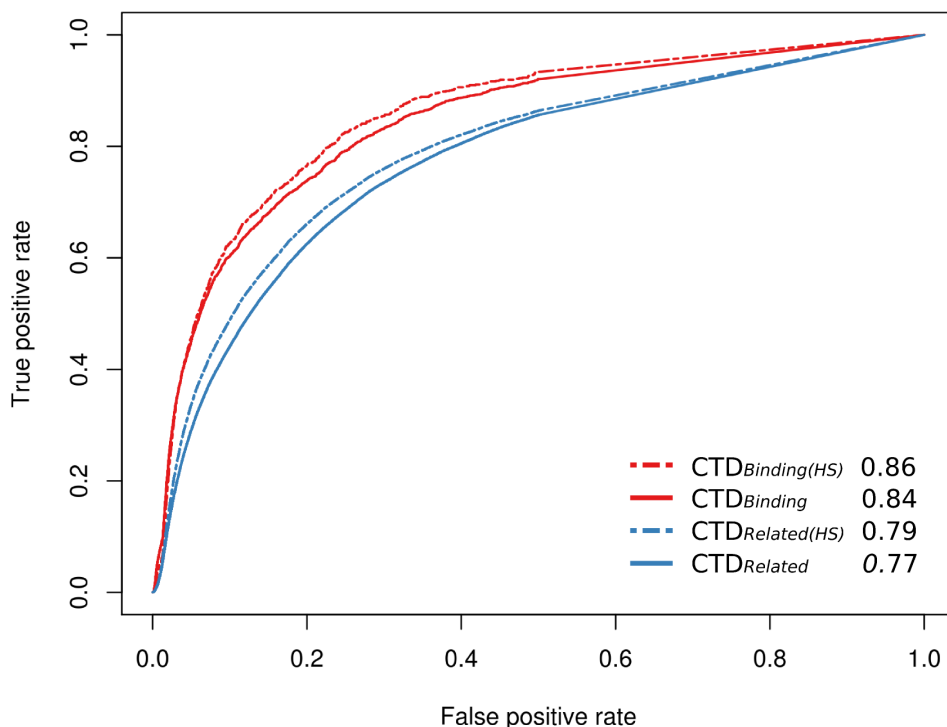
while,

Diclofenac **binds to** *ALB* protein.

is considered as a *Binding* association.

Why CTD?

Binding and Related associations

Figure 4.6: ROC curves for manually curated profiles

The curves show that CASSANDRA produces comparable prediction performance when manually curated profiles are utilized. The algorithm successfully prioritizes direct (i.e., physically interacting) drug gene associations over the indirect ones. Once more, the performance improves when associations containing *Human* genes are considered.

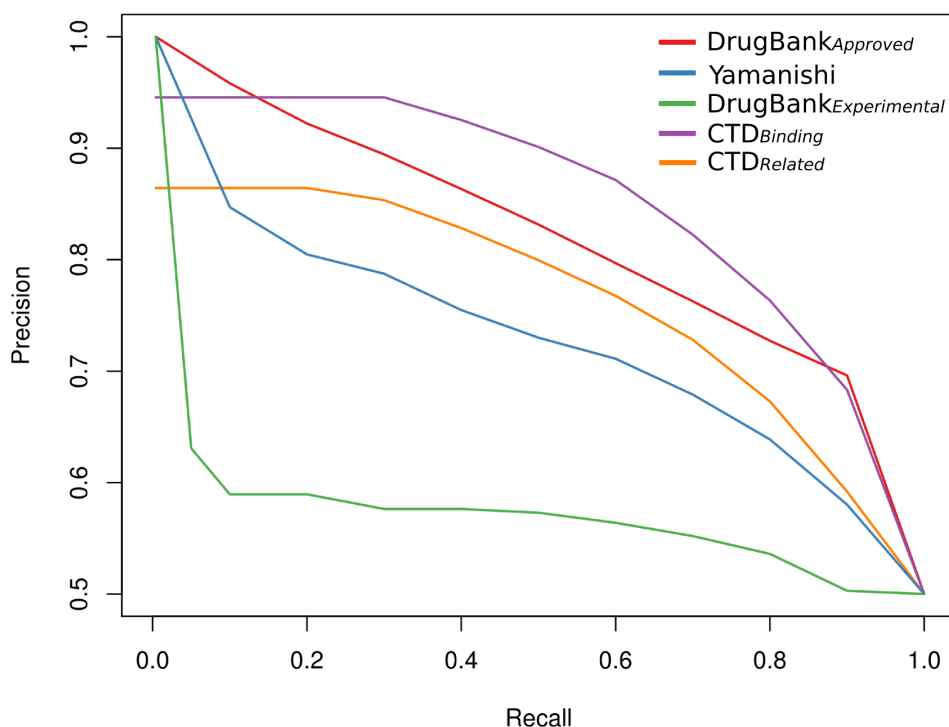
As shown in Figure 4.6, the suggested method obtains an *AUC* value of 0.84 for the *Binding* subset. When considering only *Human* genes, the *AUC* value rises to 0.86. With respect to the *Related* drug gene associations dataset, CASSANDRA achieves an *AUC* of 0.77. This value increases to 0.79 when taking into account only drug-*Human* gene pairs. The arithmetic mean of *MeSH Diseases* related to *Human* genes in the *CTD* subsets *Binding* and *Related* is 8 and 4 respectively. In the case of non-*Human* genes, the respective means decrease to 6 and 2. This finding corroborates the fact that the more terms are included in the profile, the better the predictive efficiency of the algorithm. Another interesting finding is that the proposed method demonstrates in general a better performance on the *Binding* dataset than on the *Related* dataset. This justifies the

*Improved
performance
for Binding
associations*

algorithm's potential to successfully prioritize the drug-target interactions over drug gene associations.

The above results clearly demonstrate that the proposed methodology can utilize either co-occurrence based ontological profiles or manually curated profiles with a comparable performance. In both cases, the estimation of the statistical semantic relatedness between these profiles can produce a meaningful ranking of drug gene associations. Notably, the algorithm promotes in ranking direct drug gene associations (i.e., physically interacting drug protein associations) over indirect drug gene associations.

Figure 4.7: *PR* curves for all datasets



All *PR* curves are plotted for an 1:1 ratio of imbalance between the set of true and false drug gene associations. As shown, CASSANDRA's classification efficacy on the *PR* space is consistent to the one demonstrated in the *ROC* space. Again, the performance of the algorithm is better on the datasets *DrugBank_{Approved}* and *CTD_{Binding}*, when these are compared to the performance of *DrugBank_{Experimental}* and *CTD_{Related}* respectively.

To confirm the algorithm's efficacy towards drug gene association prediction, the *PR* curves for the datasets utilized so far are provided. Given a

specific dataset, *Precision* is plotted against *Recall* for different ratios of imbalance between the set of true drug gene associations and a subset of randomly selected false drug gene associations.

Figure 4.7 shows the respective *PR* curves on an 1:1 ratio of imbalance for the datasets *DrugBank_{Approved}*, *DrugBank_{Experimental}*, *Yamanishi*, *CTD_{Binding}* and *CTD_{Related}*. As shown, CASSANDRA dominates both the *ROC* and the *PR* space. The results are consistent to the ones suggested by the *ROC* curves. Given the same recall space, the algorithm achieves higher precision for the datasets *DrugBank_{Approved}* and *CTD_{Binding}* when these are compared to *DrugBank_{Experimental}* and *CTD_{Related}* respectively. Same is the algorithm's behavior when different ratios of imbalance are considered (see *Supplementary Material*, Figure 7.1). The case of *DrugBank_{Experimental}* dataset is particular. As shown in Figure 4.12 (1st row, 2nd column), the score distributions for true and false drug gene associations included in this dataset are very similar. Therefore, when increasing the ratio of imbalance the performance is expected to drop faster in contrast with the rest of the datasets.

PR curves are consistent

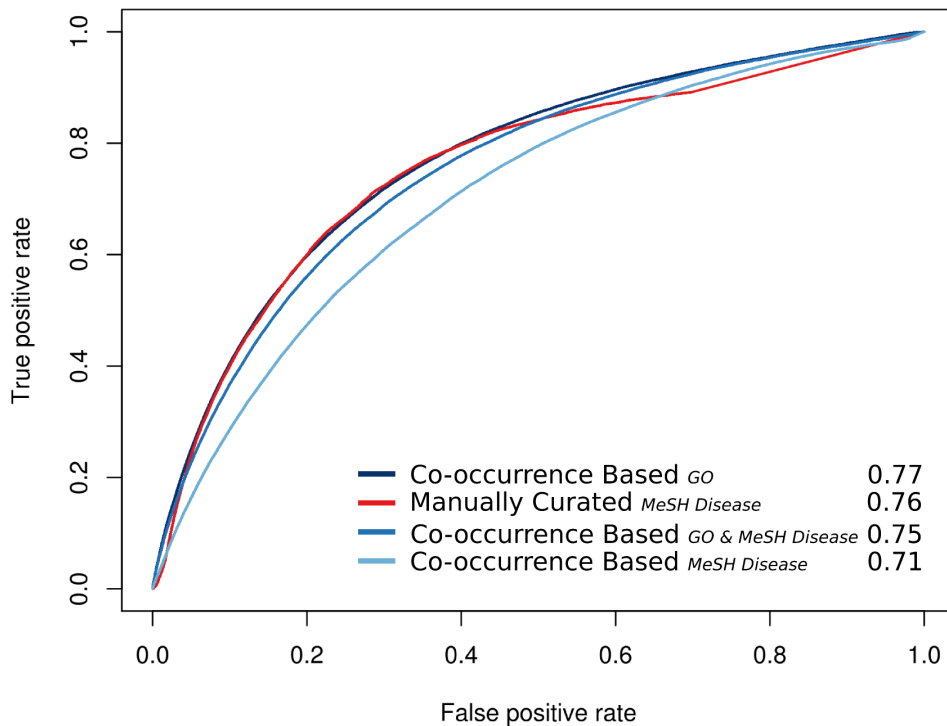
4.3.1.3 Co-occurrence based vs. manually curated profiles

The suggested methodology can utilize either co-occurrence based profiles or manually curated profiles and successfully prioritize known drug gene associations. The arising question is which profile type performs better towards the establishment of latent drug gene associations. Manually curated profiles are expected to outperform due to their high quality and strongly established relation to drugs or genes compared to the co-occurrence based profiles. In this section, the comparison between utilizing manually curated and co-occurrence based profiles towards the discovery of drug gene associations is described.

To compare the two profile types, all *CTD* true drug gene associations are considered. The set of false drug gene associations is compiled as described in Section 3.2.1. Then, the drug gene associations wherein the drug and the gene have both manually curated and co-occurrence based profiles are maintained. These associations basically constitute the intersection of *CTD* drug gene associations and the drug gene associations provided by the

suggested methodology. Of course, the drug gene pairs wherein the drug and gene are co-mentioned in literature were excluded. The resulting dataset i.e., $CTD_{cb\&mc}$, is then used for the comparison between the manually curated and the co-occurrence based ontological profiles.

Figure 4.8: ROC curves - Manually curated vs. co-occurrence based profiles

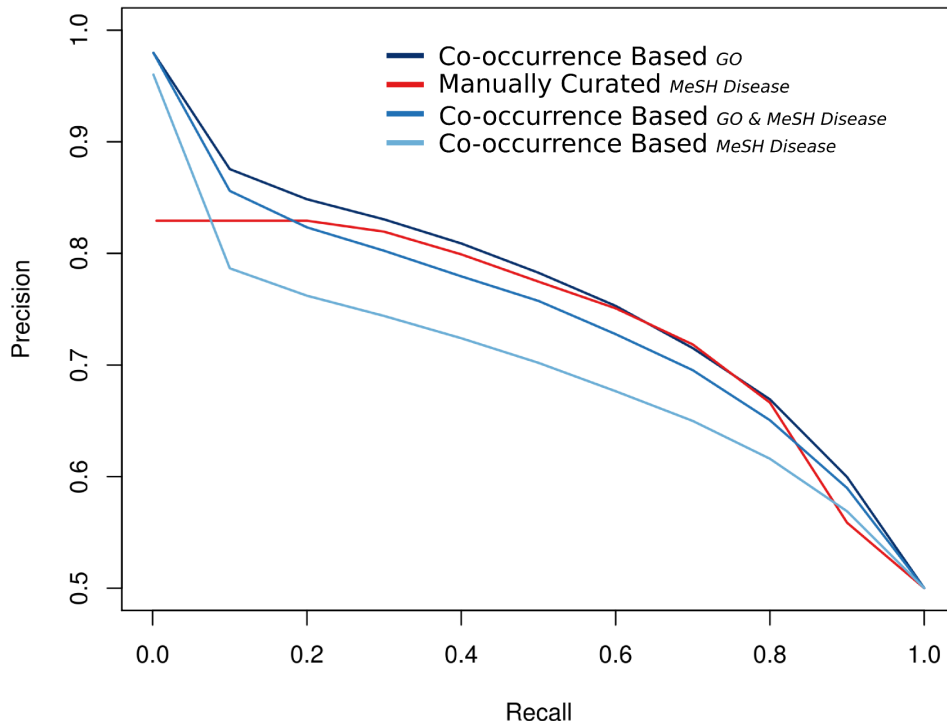


The curves show that CASSANDRA performs better in the case of co-occurrence based *GO* profiles than in the case of manually curated *MeSH Disease* profiles. When co-occurrence based profiles of both ontologies are used, the *AUC* value equals the one obtained by manually curated profiles. The performance drops by 6% in the case of co-occurrence based *MeSH Disease* profiles.

As shown in Figure 4.8, co-occurrence based profiles that constitute of *GO* terms outperform manually curated *MeSH Disease* profiles. The former achieves an *AUC* of 0.77 in comparison to the latter that managed an *AUC* of 0.76. Surprisingly, when using both *GO* and *MeSH Disease* co-occurrence based profiles, the performance drops to an *AUC* of 0.75. Comparing manually curated to co-occurrence based *MeSH Disease* profiles, the former outperformed the latter only by 6%.

*Co-occurrence
GO profiles
outperform in
ROC space*

Figure 4.9: *PR* curves - Manually curated vs. co-occurrence based profiles



The *PR* curves are consistent to the respective *ROC* curves. The co-occurrence based *GO* profiles achieve a higher *Precision* to *Recall* performance compared to the manually curated *MeSH Disease* profiles. Combined co-occurrence based profiles that consist of both ontologies also manage higher precision than the manually curated ones within certain ranges of *Recall*.

To insure the aforementioned results, the *PR* curves are also provided (see Figure 4.9). The performance of the algorithm is consistent to the one illustrated by the *ROC* curves. As shown, the co-occurrence based *GO* profiles achieve higher precision for most of the *Recall* values. Within the *Recall* ranges (0.0, 0.2) and (0.85, 1) combined co-occurrence based profiles of *GO* and *MeSH Disease* terms also managed higher *Precision*. When introducing different ratios of imbalance the algorithm's efficacy is not substantially affected (see *Supplementary Material*, Figure 7.2). Co-occurrence based profiles including *GO* terms continue to outperform the rest. When focusing on *MeSH Disease* terms, the co-occurrence based profiles are quite close in performance to the manually curated ones. This

*Co-occurrence
GO profiles
outperform in
PR space*

suggests that co-occurrence information encompasses a sufficient and reliable signal of association between two entities.

The impact of *GO* terms

At this point, the differentiations in the algorithm’s performance pertaining to the type of ontological terms included in a profile, are further explored. All datasets for which there exist co-occurrence based ontological profiles corresponding to drugs and genes are utilized. The evaluation process is repeated and the algorithm’s efficacy is estimated considering only one type of ontological terms in each experiment; *MeSH Disease* or *GO* terms. The results are compared to the *AUC* values obtained when using combined ontological profiles. Table 4.7 reports the respective findings.

Table 4.7: *AUC* values for different type of ontological profiles

	Combined profile	<i>Gene Ontology</i>	<i>MeSH Disease</i>
<i>DrugBank_{Approved}</i>	0.837	0.842	0.77
<i>DrugBank_{Approved(HS)}</i>	0.875	0.846	0.855
<i>DrugBank_{Experimental}</i>	0.577	0.675	0.393
<i>DrugBank_{Experimental(HS)}</i>	0.772	0.79	0.71
<i>Yamanishi</i>	0.735	0.775	0.655
<i>CTD_{cb&mc}</i>	0.756	0.771	0.71

The datasets comprising drugs and genes with co-occurrence based profiles were included in the analysis. The *AUC* is estimated when solely *GO* or solely *MeSH Disease* terms are utilized. The results are compared to the performance achieved for combined ontological profiles. *GO* terms boost the performance of the algorithm and have the most intense impact in successfully discriminating true from false drug gene associations.

GO terms were observed to have a significantly larger impact in the prediction performance compared to the *MeSH Disease* terms. More precisely, considering only *GO* terms, the *AUC* values increase in almost all datasets. Notable is the case of the *DrugBank_{Experimental}* dataset wherein *GO* profiles performed better by 14.5% when compared to the combined profiles,

*GO terms are
the most
beneficial*

and by 42% when compared to *MeSH Disease* profiles. As it has been reported before, *Experimental* drugs are underrepresented in literature and the average number of ontological concepts is significantly lower than the number of concepts included in the profiles of *Approved* drugs. Evidently, *GO* terms are proved to be more efficient in establishing latent associations between drugs and genes, even when the literature information provided is limited. These findings are in complete accordance with the results reported in Figures 4.8 and 4.9, wherein the co-occurrence based profiles are compared to the manually curated ones.

The above analysis suggests that the use of co-occurrence based *GO* profiles can successfully prioritize true from false drug gene associations. But why do *GO* terms boost the algorithm's performance? This is most likely due to specific nature of *GO* terms compared to that of *MeSH Disease* terms. More specifically, *GO* constitutes a controlled vocabulary entirely dedicated to the attributes of genes and their products. These terms may sufficiently characterise the latent processes that a drug is involved in, in case the drug and the term share significant mutual information. On the other hand, diseases are usually a combination of several conditions, functions, processes and symptoms that each one of them can be connected to several drugs and genes at the same time. Hence, the associations proposed by the semantic similarity of *MeSH Disease* terms are not as precise as the ones derived from the semantic similarity between *GO* terms.

To conclude, the results provided in this evaluation part, demonstrate the potential of co-occurrence based profiles compared to manually curated profiles towards the prediction of putative drug gene associations. Most importantly, it is shown, that if co-occurrence based profiles consist of terms highly precise and descriptive (such as *GO* terms), they can outperform manually curated profiles. The contribution of *GO* terms towards drug gene association prediction is significantly higher than the contribution of the *MeSH Disease* terms. *GO* terms succeed in efficiently prioritizing true over false drug gene associations, even when little literature information is provided. In the end, to establish a relation between two terms, the type and quality of terms play the same or even more important role than the way this relation is established in the first place (manually or statistically).

Why are GO terms more predictive?

The type of terms is important

4.3.2 Performance evaluation - Semantic similarity

In the previous Section (4.3.1), the impact of ontological profiles is examined based on the way the profiles are generated (manually curated or co-occurrence based) and the type of ontological terms they comprise (*GO* or *MeSH Disease* terms). In this Section the efficacy of *nPMI* as a statistical measure of semantic similarity between the profiles will be investigated.

To demonstrate the results of the metrics comparison the datasets *Yamanishi*, *DrugBank_{Approved}* and *DrugBank_{Experimental}* are utilized. The statistical semantic similarity *nPMI* is replaced by the metrics *Wu-Palmer* and *Lin* and the efficacy of the algorithm is estimated respectively (see Section 3.2.2). Both *ROC* and *PR* curves are provided (Figures 4.10 and 4.11 respectively). Additionally, we plot density distributions between true and false drug gene associations produced by the algorithm when each one of the metrics is considered.

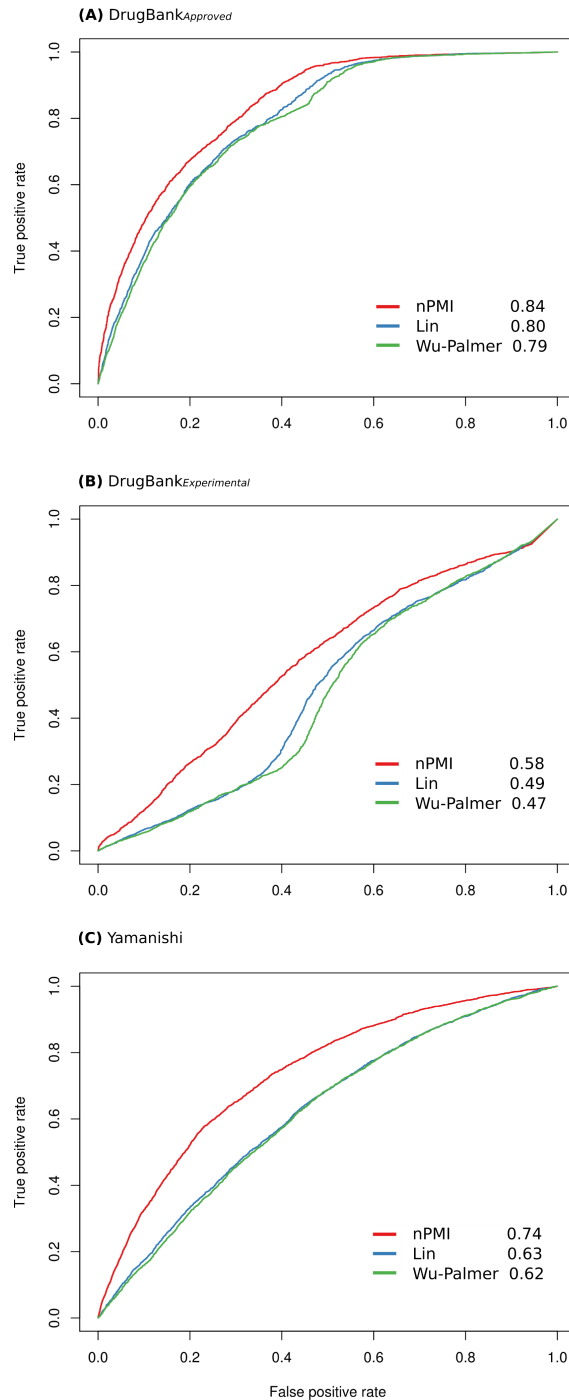
As shown in Figure 4.10, *nPMI* succeeds the highest *AUC* value in all datasets under evaluation. Notably, the metrics *Wu-Palmer* and *Lin* when applied on the *Experimental* dataset perform worse than the random classifier. Clearly *nPMI* constitutes the metric of choice when the provided literature information is limited and sparse, as in the case of the *Experimental* dataset. In all graphs, *Lin* presents slight improvement when compared to *Wu-Palmer* most likely due the fact that as a measure it incorporates the signal each concept carries in the corpus (i.e., *Information Content*).

nPMI
outperforms
Wu-Palmer
and *Lin*

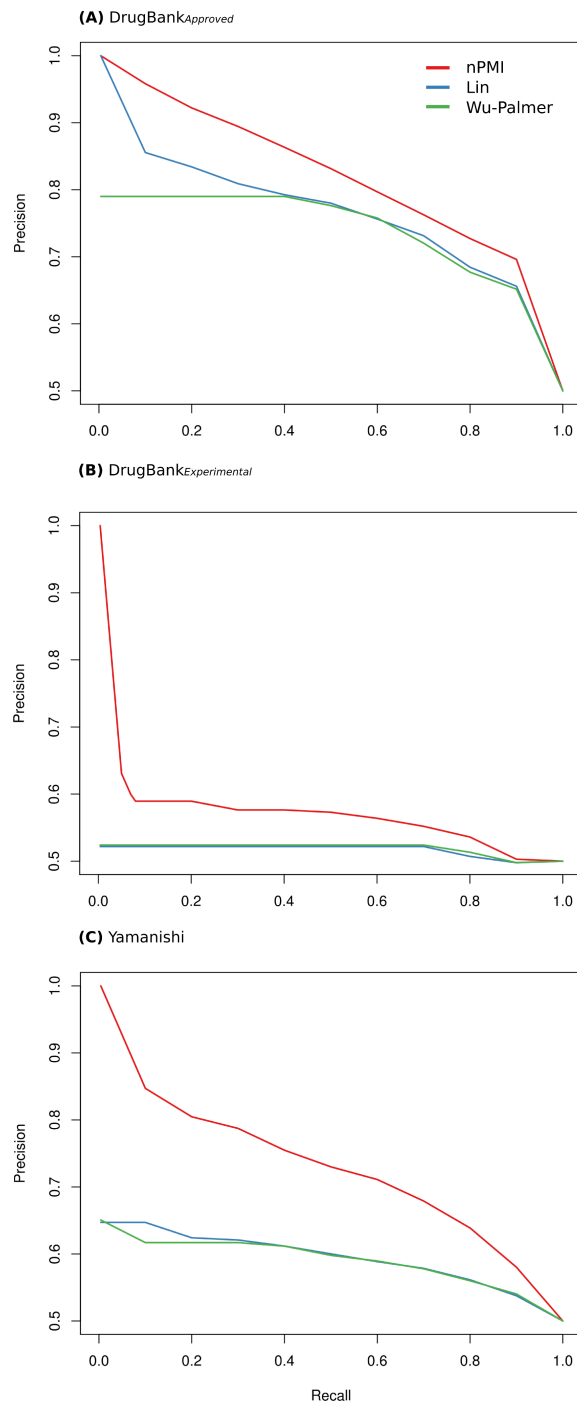
The *PR* graphs shown in 4.11 are in accordance with the behavior of the metrics as that displayed in the *ROC* curves. In all datasets, *nPMI* statistical semantic similarity interrelates drugs to genes and discriminates true to false drug gene associations with significantly higher *Precision* for the whole range of *Recall* values.

To examine the discriminative power of *nPMI* versus the metrics *Lin* and *Wu-Palmer*, the *Probability Density Functions (PDFs)* are designed. For each one of the datasets, given the number of true drug gene pairs, an equal number of false drug gene pairs is randomly selected from the respective subset of false drug gene associations. As shown in Figure 4.12, *nPMI* succeeds the highest degree of discrimination between the true and false drug gene pairs in all datasets.

nPMI has the
highest
discriminative
power

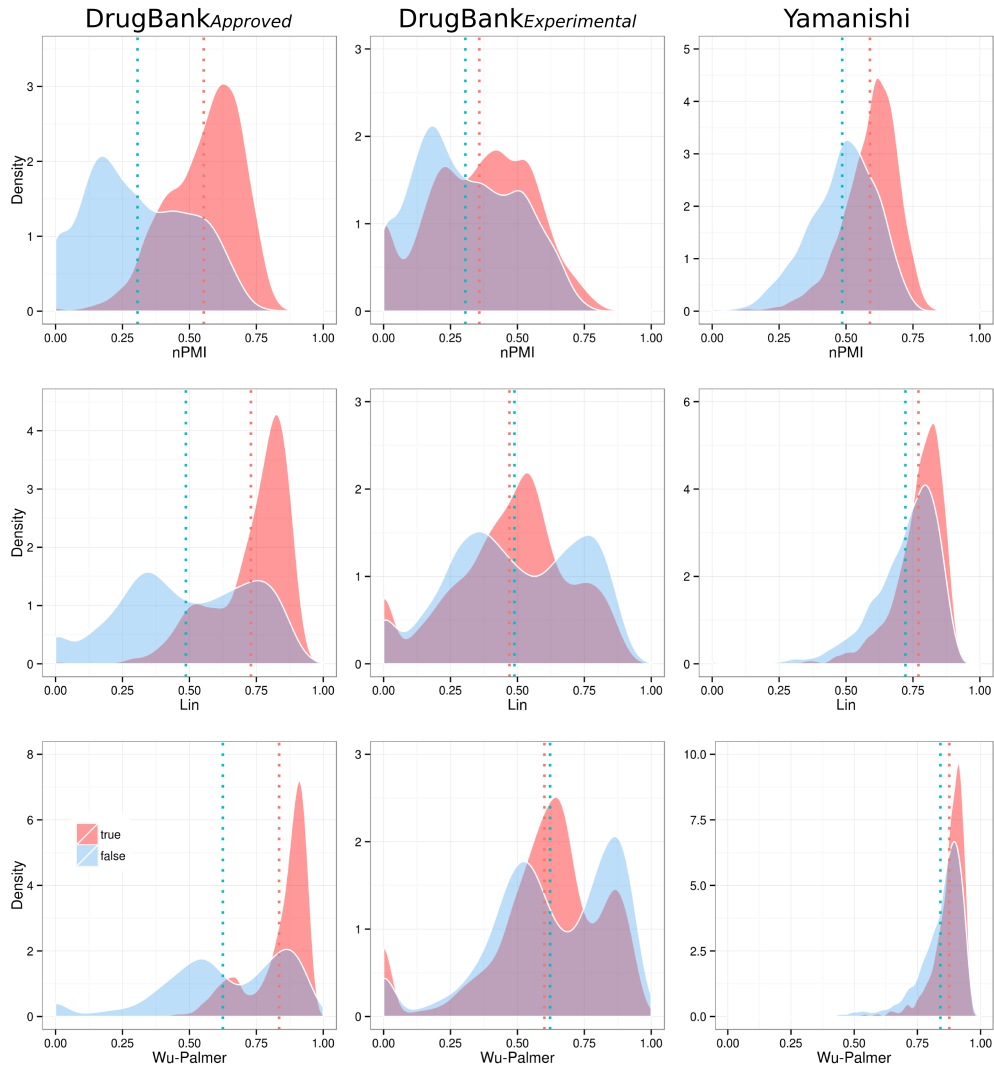
Figure 4.10: ROC curves for different metrics of semantic similarity

ROC curves plotting the true positive rate against the false positive rate of each semantic similarity measure on the datasets *DrugBankApproved*, *DrugBankExperimental* and *Yamanishi*. The curves show that *nPMI* outperforms both *Wu-Palmer* and *Lin*. *Lin* demonstrates slightly better performance than *Wu-Palmer* due to the incorporation of the concepts' *Information Content*.

Figure 4.11: *PR* curves for different metrics of semantic similarity

The *PR* curves each metric achieves on the datasets *DrugBank Approved*, *DrugBank Experimental* and *Yamanishi* are shown. The curves show that *nPMI* achieves higher *Precision* for all *Recall* values compared to *Wu Palmer* and *Lin* metrics. Again, *Lin* demonstrates slightly better performance than *Wu Palmer* due to the incorporation of the *Information Content* of the concepts towards the estimation of semantic similarity.

Figure 4.12: Density distributions for different measures of semantic similarity



The *Density* distributions of the similarity scores assigned to true and false drug gene pairs produced by each one of the metrics, are provided for the datasets *DrugBankApproved*, *DrugBankExperimental* and *Yamanishi*. The vertical red and blue lines represent the arithmetic *mean* of the scores for the true and false drug gene pairs respectively. Clearly, *nPMI* is the most efficient measure to discriminate true from false drug gene associations in all datasets. Notably, in the case of the *DrugBankExperimental* dataset, *Wu-Palmer* and *Lin* completely fail the task; the *mean* of false drug gene pairs is greater than the *mean* of the scores assigned to true drug gene pairs.

More specifically, the arithmetic *mean* of the scores assigned to true drug gene pairs by *nPMI* is always greater than the *mean* assigned to the false drug gene pairs scores. Moreover, when calculating the difference of the true to false *mean* values again *nPMI* outperforms. Interestingly, when applied on the *DrugBank_{Experimental}* dataset, the metrics *Lin* and *Wu-Palmer* completely fail to discriminate true from false drug gene pairs. As shown in Figure, the *mean* values for the density distributions of the false drug gene pairs are greater than the *mean* values for the density distributions of the false drug gene associations (for analytical *mean* values, see Table 7.1 in *Supplementary Material*).

To further assess the discriminative power of each metric of semantic relatedness, the statistical *Student's t-test* is performed. The respective statistical test computes the probability of the *null* hypothesis, meaning the probability that two sets of scores come from the same distribution. A probability less than 0.05 signifies that the respective metric can discriminate with statistical significance the two distributions. Additionally, the *Kolmogorov-Smirnoff (KS)* statistical test is applied. For both true and false drug gene association scores, the distance between the respective *Empirical Cumulative Distribution Function (ECDF)* distributions is computed, according to the *KS* test. Briefly, the greater the distance between two *ECDF* distributions, the higher the discrimination degree between them. Tables 4.8 and 4.9 collectively report the probability, *ECDF* distances and *AUC* scores for every dataset constituted by drugs and genes with co-occurrence based or manually curated profiles respectively.

According to the values reported in Tables 4.8 and 4.9, the statistical similarity measure *nPMI* systematically achieves the highest *AUC* values in all datasets, whether co-occurrence based profiles (see Table 4.8) or manually curated profiles (see Table 4.9) are used. Moreover, *nPMI* manages to efficiently discriminate true from false drug gene pairs with statistical significance. Even in the case of the datasets *DrugBank_{Experimental}* and *Yamanishi*, *nPMI* achieves the lowest *p-value*. Additionally, when *nPMI* is utilized, the distance between the *ECDF* distributions of the scores for true and false drug gene pairs is the highest.

What is the explanation behind *nPMI*'s efficacy towards the estimation of

*KS test and
t-test are
concordant*

*nPMI
outperforms*

Table 4.8: Overall performance scores - Datasets for co-occurrence based profiles

Datasets	AUC			p-value			KS distance		
	nPMI	WP	Lin	nPMI	WP	Lin	nPMI	WP	Lin
<i>DrugBank Approved</i>	0.837	0.788	0.798	0.0	0.0	0.0	0.5083	0.4308	0.437
<i>DrugBank Approved(HS)</i>	0.875	0.84	0.854	0.0	0.0	0.0	0.597	0.5501	0.5731
<i>DrugBank Experimental</i>	0.577	0.471	0.485	$4.75 \times e^{-39}$	$2.38 \times e^{-9}$	0.0024	0.1352	0.1516	0.125
<i>DrugBank Experimental(HS)</i>	0.772	0.725	0.74	$4.86 \times e^{-100}$	$2.34 \times e^{-103}$	e^{-92}	0.4319	0.3836	0.4083
<i>Yamanishi</i>	0.735	0.619	0.625	0.0	$4.3 \times e^{-108}$	$7.79 \times e^{-123}$	0.3566	0.1911	0.192
<i>CTD_{cb&mc}</i>	0.755	0.722	0.725	0.0	0.0	0.0	0.3931	0.3293	0.3393

For all the datasets that consist of drugs and genes with co-occurrence based profiles, the performance of *nPMI* against traditional measures of semantic similarity, i.e., *Wu-Palmer* and *Lin*, is evaluated. The *AUC* scores are provided. To assess the discriminative power of the suggested method over true and false drug gene associations, the statistical *Student's t-test* and the *KS test* are applied. *nPMI* outperforms even in the case of *DrugBank Experimental* and *Yamanishi* datasets, wherein the discrimination task is more demanding.

Table 4.9: Overall performance scores - Datasets for manually curated profiles

Datasets	AUC		p-value		KS distance				
	nPMI	WP	Lin	nPMI	WP	Lin			
<i>CTD_{Binding}</i>	0.844	0.821	0.816	0.0	0.0	0.0	0.5456	0.5128	0.5019
<i>CTD_{Binding(HS)}</i>	0.856	0.837	0.836	$1.24 \times e^{-203}$	$3.96 \times e^{-250}$	$8.11 \times e^{-236}$	0.5775	0.5406	0.5392
<i>CTD_{Related}</i>	0.774	0.749	0.749	0.0	0.0	0.0	0.4395	0.3959	0.3868
<i>CTD_{Related(HS)}</i>	0.789	0.771	0.772	0.0	0.0	0.0	0.4676	0.4322	0.4277
<i>CTD_{cb&mc}</i>	0.758	0.742	0.743	0.0	0.0	0.0	0.4297	0.3889	0.3895

For all the datasets that consist of drugs and genes with manually curated profiles, the performance of the proposed measure against the metrics *Wu-Palmer* and *Lin*, is evaluated. The *AUC* scores are provided. Again the discriminative power of CASSANDRA is assessed by applying the statistical *Student's t-test* and the *KS test*. *nPMI* outperforms in every dataset.

semantic similarity between drugs and genes? Two are the basic characteristics that differentiate *nPMI* from the *Wu-Palmer* and *Lin* metrics.

First, *nPMI* is a statistical and fully corpus-based metric which ignores the structure of the ontology, meaning the relationships between the terms. When applying *nPMI*, the concepts that constitute the respective ontology, are simply treated as a set of terms, i.e., a lexicon, wherein the similarity between them is estimated solely based on the degree of their co-occurrence in *MEDLINE* indexed titles and abstracts. This accounts for the general applicability of the method across different ontologies and has a positive impact on the method's performance. More precisely, the statistical measure *nPMI* spots similarities between concepts which are not detected by any other measure that is based on the hierarchy of the ontology. With the use of *nPMI* these pairs of concepts participate in the calculation of the association score between a drug and a gene. Consequently, this has a beneficial effect on the recall of the method, since additional drug gene associations can be suggested.

nPMI doesn't depend on the structure of the ontology

In particular, *nPMI* is able to assign a similarity score between *GO* terms which belong to different subontologies, e.g., a term from the *Biological Process* and a term of the *Molecular Function* subontology if there is substantial co-occurrence data. Accordingly, similarity can be computed between a *MeSH Disease* and a *GO* concept. In the cases described above, the measures *Wu-Palmer* and *Lin* would assign a zero similarity score. *Wu-Palmer* and *Lin* would also fail to compute any semantic relatedness between two *MeSH Disease* terms wherein the latter is a symptom of the former. More precisely, let us consider *Prader-Willi Syndrome*, a congenital disease affecting many parts of the body. According to the *MeSH* definition, the symptoms of this disease include *Hypogonadism*. A quick look up in the *MeSH* hierarchy shows that the *Lowest Common Ancestor (LCA)* of *Prader-Willi Syndrome* and *Hypogonadism* is the root. Consequently, both *Wu-Palmer* and *Lin* would assign a 0.0 similarity to these concepts, while *nPMI* assigns a score of 0.42.

nPMI uncovers hidden similarities

The second characteristic of *nPMI* is the ability to discriminate concept pairs based on their frequency. Pairs composed of low-frequency terms receive a higher score compared to the ones composed of high-frequency terms (Manning and Schütze, 1999). More precisely, let us consider two

nPMI prioritizes hidden associations

concept pairs that constitute of concepts that are semantically close, meaning concepts that are close in the ontology tree. If one pair is frequent (and hence general) and the other pair is rarely found in text, then the latter is more informative than the former. With the use of $nPMI$ this difference is captured and represented in the association score between a drug and a gene. Highly frequent concept pairs contribute less to a drug gene association score than pairs of lower frequency. This explains why $nPMI$ performs better than the traditional measures of semantic similarity. Table 4.10 shows an example of this phenomenon.

Table 4.10: $nPMI$ computations - Examples

	Pair A		Pair B	
	<i>Coronary Disease</i>	<i>Myocardial Ischemia</i>	<i>Kearns-Sayre Syndrome</i>	<i>Ophthalmoparesis</i>
$n_{C_d/g}$	164,596	55,757	514	846
n_{C_d,C_g}	18,136		134	
distance	1		2	
$nPMI$	0.46		0.71	
<i>Wu-Palmer</i>	0.89		0.83	

The table reports the difference between the similarity scores assigned by $nPMI$ and *Wu-Palmer* on two different concepts pairs. The distance between the concepts which constitute the pairs is reported along with the occurrence and co-occurrence values of the terms. $nPMI$ prioritizes the less frequent and hence, more informative concept pairs.

Assume $C_d = \textit{Coronary Disease}$ and $C_g = \textit{Myocardial Ischemia}$ two terms that participate in the profile of a drug d and a gene g respectively. Table 4.10 reports the number of *MEDLINE* documents where C_d occurs (n_{C_d}), the number of *MEDLINE* documents where C_g occurs (n_{C_g}), and the number of *MEDLINE* documents where C_d and C_g co-occur (n_{C_d,C_g}). It additionally reports the distance of the terms in the ontology tree and two values of semantic similarity. The $nPMI$ and the *Wu-Palmer* semantic similarity. Next, assume the pair $C_d = \textit{Kearns-Sayre Syndrome}$ and $C_g = \textit{Ophthalmoparesis}$, for which the respective numbers are also reported. This example shows that for two pairs of concepts that are semantically

close, though the number of occurrences and co-occurrences of the first pair is significantly higher than the respective values of the second pair, the second pair receives much higher *nPMI* score. Through this example we can observe that the application of *nPMI* enables the identification of latent relations between ontological terms that do not necessarily occur very frequently, as in the case between *Kearns-Sayre Syndrome* and *Ophthalmoparesis* where the former is a syndromic variant of the latter. *nPMI* prioritizes these pairs in comparison to other frequent and hence less informative concept pairs.

Conclusively, *nPMI* is an efficient measure of semantic relatedness between two ontological concepts. Comparison against the traditional metrics *Wu-Palmer* and *Lin* demonstrates that *nPMI* has the best performance and the highest discriminative power. The suggested measure of semantic similarity spots associations between two concepts that the other metrics fail to reveal. Even when suggesting the same concept pairs, *nPMI* prioritizes the more informative relations.

4.4 Drug repositioning

We manually mine the literature and compile a set of drugs repositioning cases (see Section 3.2.1.2). Table 4.11 shows the analysis of the application of the suggested method in identifying new indications for existing drugs. We focus on the last 5 years and collect the drug repositioning cases that were approved by *FDA* and that correspond to drugs for which we have profiles. The drug profiles are generated based on literature data before the year of approval of each repositioning case. The table illustrates the old and new indications for each of the examined drugs along with their positions in the list of the drugs' profile terms.

As the table suggests, in almost all cases the old indications appear in the top 3 associated disease terms of the drug. In parallel, the new indications are always included in the co-occurrence based profiles of drugs among the top 30 associated disease terms of the respective profiles. To assess the association between a drug and the respective new indication, 2 types of *z*-scores are computed. The first is the *z*-score the new indication achieves

*Literature
before
repositioning*

*New
indications
rank high in
profiles*

inside the profile of the drug. The second z -score corresponds to the overall distribution of drug-*MeSH Disease* associations. As shown, the *mean* of the z -score of an old indication both in the profile of the drug and in the overall distribution (3.94 and 4.19 respectively) is higher than that of the new indication (2.02 and 2.41 respectively). This is expectable if we consider that the old indications of the drugs are more discussed in literature than the new indications. The results of this analysis suggest CASSANDRA can be utilized towards the identification of new indications for already existing drugs.

4.5 Case studies

In the following CASSANDRA's potential towards drug gene association prediction and drug repositioning is illustrated via three case studies. In all three cases, the findings in the scientific literature that support the proposed associations are provided along with the respective graphical representations.

4.5.1 Cathine-*GHRL* association

In Figure 4.13, all the hypotheses suggesting an association between *Cathine* (*DrugBank*: DB01486) and *GHRL* (*EntrezGene*: 51738) are shown. *Cathine*, which is a psychotropic compound, is selected because it is among the *DrugBank* compounds which do not have any target information. Altogether, 11,302 human genes were ranked against *Cathine*. *GHRL*, a gene coding for the growth hormone-releasing peptide *ghrelin*, ranks among the top 0.4% of these genes (position 54) with a z -score of 2.83 and a p -value < 0.05.

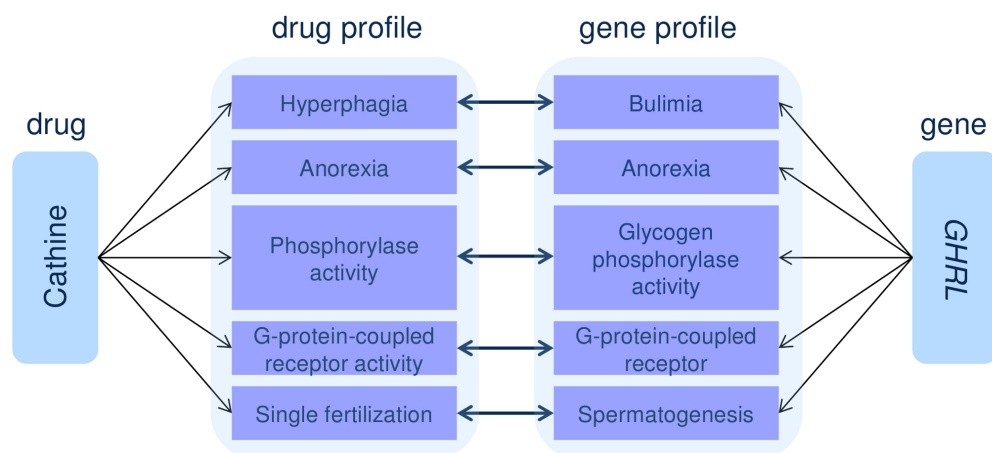
Table 4.12 shows representative textual pieces of evidence suggesting this association which is discovered by CASSANDRA. *Cathine* and *GHRL* are interconnected via the concepts *Hyperphagia* (*MeSH*: D006963) and *Bulimia* (*MeSH*: D002032). According to *MeSH*, *Bulimia* is a form of *Hyperphagia*. The association emerges as follows: *Cathine* is isolated from the plants *Catha Edulis* and *Ephedra Sinica*, and acts as a stimulant. It is a *Phenylpropanolamine* (*PPA*) isomer, along with *Norephedrine*. There exist

*Hyperphagia-
Bulimia*

Table 4.11: Examples of drug repositioning potential

Drug	Old Indications	z-score		New Indications	z-score		Year
		profile	overall		profile	overall	
<i>Milnacipran</i>	<i>Depression (N/A)</i>	N/A	N/A	<i>Fibromyalgia (1)</i>	3.44	4.29	2009
<i>Tadalafil</i>	<i>Impotence (1)</i>	4.78	5.19	<i>Hypertension, Pulmonary (11)</i>	2.07	2.42	2009
<i>Doxepin</i>	<i>Depression (N/A)</i>	N/A	N/A	<i>Insomnia (8)</i>	1.90	2.08	2010
<i>Duloxetine</i>	<i>Diabetic (3) Neuropathies</i>	2.77	3.49	<i>Shoulder Pain (5)</i>	2.05	2.71	2010
<i>Duloxetine</i>	<i>Diabetic (3) Neuropathies</i>	2.77	3.49	<i>Back Pain (15)</i>	1.60	2.22	2010
<i>Duloxetine</i>	<i>Diabetic (3) Neuropathies</i>	2.77	3.49	<i>Osteoarthritis, Knee (64)</i>	0.62	1.16	2010
<i>Tadalafil</i>	<i>Impotence (1)</i>	4.78	5.19	<i>Prostatic (12) Hyperplasia</i>	1.94	2.26	2011
<i>Mifepristone</i>	<i>Abortion, (1) Incomplete</i>	5.10	4.32	<i>Cushing (25) Syndrome</i>	2.09	1.67	2012
<i>Topiramate</i>	<i>Epilepsy (1)</i>	3.23	3.81	<i>Bulimia (15)</i>	2.33	2.80	2012
<i>Budesonide</i>	<i>Asthma (3)</i>	3.71	3.80	<i>Colitis, (29) Ulcerative</i>	1.87	2.18	2013
<i>Lenalidomide</i>	<i>Multiple (1) Myeloma</i>	4.05	4.51	<i>Lymphoma, (4) Mantle-Cell</i>	2.30	2.74	2013

The potential application of CASSANDRA on drug repositioning. For all FDA approved drugs repositioned over the last 5 years, the co-occurrence based profiles from publications before 2009 are generated. The table shows that the new indications always exist and rank high in the profiles.

Figure 4.13: *Cathine - GHRL*

The figure illustrates the intermediate connections of *Cathine* to *GHRL*. Clearly, the most surprising connection is established via the concepts *Single Fertilization* and *Spermatogenesis* that pertain to *Reproduction*. These concepts are semantically quite distinct from the *Eating Disorders* concepts to which *Cathine* and *GHRL* have known relations.

several studies reporting the appetite imminent suppressive role of *PPA*'s (PMID: 3703896; 7855211). Studies regarding the effects of *PPA* on different types of *Hyperphagias* conclude that *PPA* sufficiently suppresses appetite in hyperphagic rats (PMID: 3310024). All the above support the hypothesis that *Cathine* suppresses *Hyperphagia* as an phenylpropanolamine isomer. As a result, *Cathine* may also be effective in restraining *Bulimia*. In parallel, *ghrelin* is the only known hunger-stimulating hormone and is related to several eating disorders including *Bulimia Nervosa* (MeSH: D052018) (PMID: 21453750). It is reported that when increasing the levels of *ghrelin* via its direct injection into the brain ventricles, the consumption of rewarding foods in mice and rats increases, as well (PMID: 21354264). In the same paper it is stated that *ghrelin receptor (GHS-R1A)* antagonists show beneficial effects towards the suppression of food intake. In addition, it is also stated that variations in the *GHS-R1A* and pro-*ghrelin* genes have been associated with *Bulimia Nervosa* and obesity.

Cathine can also be connected to *GHRL* via *Anorexia* (MeSH: D000855). *Cathine*'s product information describes the drug as anorexic. It has also been stated that *ghrelin* in hypothalamic neurons controls *Anorexia* and

Anorexia

Cachexia (*MeSH*: D002100) (PMID: 22632865). The therapeutic applications of *ghrelin* towards these conditions have been also discussed (PMID: 21635929).

Moreover, *Cathine* is involved in *G-protein Coupled Receptor Activity* (*GO*: 0004930) (PMID: 17158213), and *ghrelin*'s receptor is also a G-protein coupled receptor (PMID: 16382107).

GPCR activity

In addition, an increase in the adrenal *Phosphorylase Activity* (*GO*: 0004645) has been observed after the administration of *Cathine* (PMID: 7903110). In the same study, it is also reported that the *glycogen* levels were decreased. Other studies in tundra vole (*Microtus oeconomus*) (PMID: 15302267) show that after the injection of intraperitoneal *ghrelin*, kidney *Glycogen Phosphorylase Activity* (*GO*: 0008184) increased, whilst kidney *glycogen* levels decreased. The above suggests similar responses after *ghrelin*'s or *Cathine*'s administration.

Phosphorylase Activity

The last connection is a surprising one, since it is formed via the concepts of *Single Fertilization* (*GO*: 0007338) and *Spermatogenesis* (*GO*: 0007283). Both concepts pertain to reproduction. Studies in incapacitated mouse spermatozoa, markedly demonstrate that *cathine* significantly accelerates capacitation (PMID: 15513978; 17158213). Additionally, observations in normal adult rats suggest *ghrelin*'s modulative role in *Spermatogenesis* (PMID: 22360851;22658447).

Unexpected relation to Reproduction

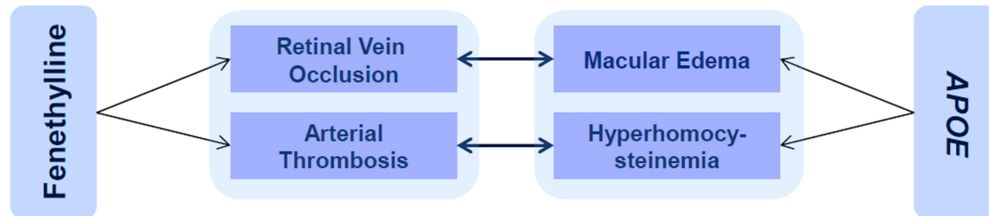
Conclusively, all the described links above account for the hypothesis that *Cathine* and *GHRL* are associated. Although, both the drug and the gene are known to be related to *Eating Disorders*, their association is intensified by the discovery of a new intermediate connection that exists between them; the connection pertaining to *Reproduction*.

4.5.2 Fenethylamine-*ApoE* association

The second example suggests a connection between the gene *ApoE* and the drug *Fenethylamine* (Figure 4.14 and Table 4.13 give the respective overview). *ApoE* ranks among the top 5 in the ranked list of 11, 302 *Human* genes with a *z*-score of 4.27 and a *p*-value < 0.05.

Table 4.12: *Cathine-GHRL* textual findings

Relation	Textual Evidence
Cathine is also called PPA	Cathine , ... is one of the optical isomers of phenylpropanolamine (PPA), <i>Wikipedia</i>
PPA suppresses hyperphagia	PPA is capable of suppressing appetite in rats made hyperphagic by various stimuli (PMID:3310024)
Bulimia is an Hyperphagia	<i>MeSH</i>
Ghrelin is involved in bulimia	Ghrelin increases food intake ... relevance in the regulation of bulimia nervosa (PMID:21453750)
Cathine is an anorexic drug	Product Information
Ghrelin controls cachexia	Ghrelin in concert with hypothalamic neurons control anorexia and cachexia (PMID:22632865)
Cathine affects phosphorylase activity	After the administration of cathine , an increase in the adrenal phosphorylase activity has been observed (PMID:7903110)
Glycogen phosphorylase activity is a phosphorylase activity	<i>Gene Ontology</i>
Ghrelin affects Glycogen phosphorylase activity	After the injection of intraperitoneal ghrelin , kidney glycogen phosphorylase activities increased (PMID:15302267)
Cathine affects adrenergic receptors	Regulation of adenylyl cyclase/ <i>cAMP</i> in a G protein -mediated fashion by cathine may possibly involve adrenergic receptors (PMID:15513978)
Adrenergic receptors are G-protein coupled receptors	<i>Gene Ontology</i>
Ghrelin 's receptor is a G-protein coupled receptor	Growth hormone secretagogue receptor is a G-protein coupled receptor that binds ghrelin (PMID:16382107)
Cathine boosts single fertilization	Cathine can enhance chances of fertilization in vivo (PMID:17158213)
Single fertilization and spermatogenesis pertain to reproduction	<i>Gene Ontology</i>
Ghrelin modulates spermatogenesis	Ghrelin may be considered as a modulator of spermatogenesis (PMID:22360851)

Figure 4.14: *Fenethylline-ApoE*

The figure shows the intermediate connections of *Fenethylline* to *ApoE*. Both the drug and the gene are strongly related to cardiovascular conditions, which are expressed either as conditions directly related to heart diseases or as conditions affecting the retinal area.

Fenethylline is a stimulant compound and has been used for the treatment of *Hyperkinesia* and depression (PMID: 23420919). When metabolised is forming the substances *Amphetamine* and *Theophylline* (PMID: 5496920). *Apolipoprotein E* is a mediator of liver endocytosis and it has been characterised as a major genetic risk factor for *Alzheimer's disease* (PMID: 22622580).

The first hypothesis is formed via the interrelated concepts *Retinal Vein Occlusion* (MeSH: D012170) and *Macular Edema* (MeSH: D008269). According to the respective MeSH definition *Retinal Vein Occlusion* is the condition describing the blockage of the retina. It is a high risk condition for patients with *Diabetes* or several cardiovascular diseases. Three cases of hemorrhagic central *Retinal Vein Occlusion* following continuous uses of *Fenethylline* have been reported (PMID: 20214057). In the same study is also stated that after the discontinuation of the drug, the symptoms markedly withdrew. Following the MeSH definitions, *Macular Edema* is the accumulation of fluid or protein around the macula of the eye and it is oftenly seen with retinal occlusive diseases (PMID: 23410812; 22823029). Besides, another study characterises the *APOE* gene polymorphism as a risk factor for the severity of *Macular Edema* (PMID: 11910554).

Secondly, two more concepts relate *Fenethylline* to *ApoE*. As it has been already mentioned, *Fenethylline* forms *Amphetamine* which is associated with *Arterial Thrombosis* (MeSH: D013927) (PMID: 20118172). Results

*Retinal Vein
Occlusion -
Macular
Edema*

*Thrombosis-
Hyperhomo-
cysteinemia*

prove that *Hyperhomocysteinemia* (*MeSH*: D020138) is the most common condition that is highly associated with both *Venous* and *Arterial Thrombosis* (PMID: 22933895). Generally, there has been a long-recognized connection between high levels of *Homocysteine* (*Hyperhomocysteinemia*) and *Thrombosis* (PMID: 22461473). Concomitantly, another study suggests that *ApoE4* (one of *ApoE*'s isoforms) is related to *Hyperhomocysteinemia* (PMID: 17158432). Moreover, in scientific literature null-*ApoE* mice are extensively used in the study of *Hyperhomocysteinemia* effects (PMID: 22704348; 23017835; 20696152).

Adding it together, the above information suggests a putative association between *Fenethylamine* and the processes wherein the gene *ApoE* is involved.

4.5.3 Milnacipran-*SLC6A4* association

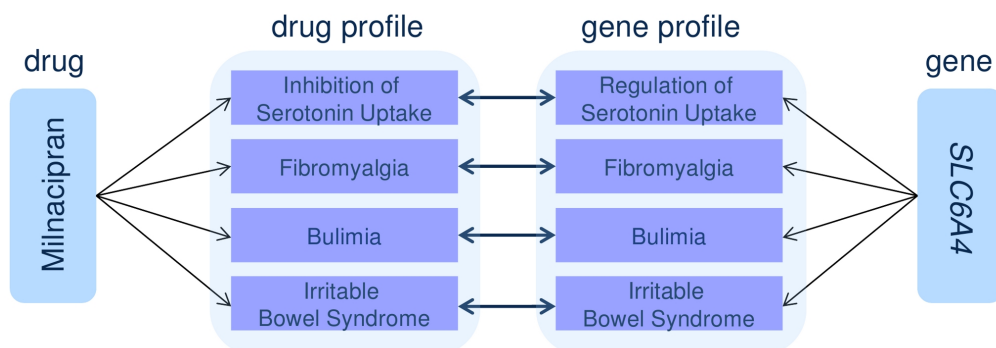
The following case study represents the repositioning potential of CAS-SANDRA. It describes the known association between the drug *Milnacipran* (*DrugBank*: DB04896) and the gene *SLC6A4* (*Entrez Gene*: 6532), which codes for *Milnacipran*'s known target, *serotonin transporter* (*SERT*). *SERT* ranks at the top (1st) of the list of 11,302 *Human* genes with a *z*-score of 4.48 and a *p*-value < 0.05.

Milnacipran is a serotonin-norepinephrine reuptake inhibitor (*SNRI*) initially approved for the treatment of *Depression* (*MeSH*: D003863) (1996). In January 2009 *Milnacipran* was also approved for the treatment of *Fibromyalgia* (*MeSH*: D005356). The *SLC6A4* gene codes for the *serotonin transporter*, which is the target protein of many antidepressant medications and whose polymorphic region is associated with a variety of anxiety-related traits and susceptibility for *Depression* (PMID: 17726476).

Figure 4.15 shows the suggested connections and Table 4.14 summarizes the textual pieces of evidence that support them. The connections are generated from *MEDLINE* abstracts and articles published before 2009, when *Milnacipran* was repositioned to *Fibromyalgia*.

The first connection is formed via the interrelated concepts of *Inhibition of Serotonin Uptake* (*GO*: 0051614) and *Regulation of Serotonin Uptake* (*GO*:0051611). *Milnacipran* belongs to the class of *SNRIs*. *SNRIs* increase

*Serotonin
Uptake
Regulation*

Figure 4.15: *Milnacipran- $SLC6A4$* 

The figure demonstrates the indirect connections for the known association *Milnacipran* to its target-coding gene *SLC6A4*. The connections are generated from data published before 2009. As shown, *Fibromyalgia*, a condition to which *Milnacipran* was repositioned after 2009, participates in the establishment of the respective association.

the levels of *serotonin*, by blocking *SERT* which is responsible for the *Regulation of Serotonin Uptake*.

The second concept relating *Milnacipran* to *SERT* is *Fibromyalgia* (MeSH: D005356). Several articles describe clinical trials and report the efficacy of *Milnacipran* in the treatment of *Fibromyalgia* more than 4 years before the compound has been approved for use against that condition (PMID: 15378666; 16206355). Other reports confirm that the polymorphic region of *SLC6A4* is associated to *Fibromyalgia* (PMID:11920428;10555044).

Fibromyalgia

Moreover, *Milnacipran* may have a beneficial effect in the treatment of *Bulimia nervosa* (PMID: 12650949; 18728825). Several articles also state the association of *SERT* polymorphisms to eating disorders and in particular to *Bulimia nervosa* (PMID: 20209488; 14987118; 12768277). The last connection is formed via the concept *Irritable Bowel Syndrome* (*IBS*, MeSH: D043183) which is a condition co-morbid with *Fibromyalgia*. Experiments conducted in rodents show that *Milnacipran* has a potential in the treatment of *IBS* (PMID: 21996314). Other studies suggest that *SLC6A4* is a candidate gene potentially involved in the pathogenesis of *IBS* (PMID: 22457857; 23594334).

Bulimia & IBS

The above pieces of evidence confirm that the proposed methodology includes in the prediction of drug gene associations medical conditions that can be considered as repositioning candidates.

Table 4.13: *Fenethylamine-ApoE* textual findings

Relation	Textual Evidence
Fenethylamine causes retinal vein occlusion	...3 cases of hemorrhagic central textbfretina textbfvein textbfocclusion following continuous use of textbfphenethylamine (PMID:20214057)
Macular edema is associated with retinal vein occlusion	...treatment of macular edema associated with central retinal vein occlusion... (PMID:22823029)
Macular edema is affected by ApoE	...allele of apolipoprotein E gene is a potential risk factor for the severity of macular edema... (PMID:11910554)
Fenethylamine forms amphetamine	Fenethylamine , when metabolised is forming the substances theophylline and amphetamine , (PMID:5496920)
Amphetamines increase arterial thrombosis incidence	Amphetamines induce tissue factor ... arterial thrombosis is in turn triggered by tissue factor (PMID:20118172)
Arterial thrombosis is responsible for hyperhomocysteinemia	Hyperhomocysteinemia proves to be the most common condition highly associated with both venous and arterial thrombosis (PMID:22933895)
Hyperhomocysteinemia is related to ApoE	Apolipoprotein E e4 allele affects risk of hyperhomocysteinemia (PMID:17158432)

Table 4.14: *Milnacipran- $SLC6A4$* textual findings

Relation	Textual Evidence
Milnacipran is a serotonin-norepinephrine reuptake inhibitor	<i>Wikipedia</i>
Inhibition of serotonin uptake is a regulation of serotonin uptake	<i>Gene Ontology</i>
<i>SERT</i> is responsible for the regulation of serotonin uptake	<i>Wikipedia</i>
Milnacipran cures fibromyalgia	In this Phase II study, milnacipran led to statistically significant improvements in pain and other symptoms of fibromyalgia (PMID:16206355)
<i>SLC6A4</i> polymorphism is related to fibromyalgia	Confirmation of an association between fibromyalgia and serotonin transporter promoter region polymorphism (PMID:11920428)
Milnacipran treats bulimia nervosa	Milnacipran in the treatment of bulimia nervosa : a report of 16 cases. (PMID:12650949)
<i>SLC6A4</i> polymorphism is related to bulimia nervosa	The serotonin transporter , encoded by the <i>SLC6A4</i> gene, may also have an important role in eating disorders, as its availability is decreased in patients with bulimia nervosa ... (PMID:14987118)
Milnacipran treats Irritable Bowel Syndrome	... milnacipran has potential clinical application in the treatment of visceral pain, such as in irritable bowel syndrome... (PMID:21996314)
<i>SLC6A4</i> is a biomarker of Irritable Bowel Syndrome	...suggesting that <i>SLC6A4</i> is a potential candidate gene involved in the pathogenesis of Irritable Bowel Syndrome . (PMID:22457857)

Chapter 5

Discussion

This Chapter discusses the outcome of the current study. The implementation decisions taken for the realization of the task at hand are explained. CASSANDRA is compared against other methodologies of *Literature Based Discovery* and drug gene association prediction. More specifically, the following are discussed

- How does CASSANDRA differentiate from the rest of the methodologies implemented towards drug gene association prediction?
 - Why focusing on *Literature Based Discovery*? How is CASSANDRA contributing to the field?
 - Why utilizing titles and abstracts is sufficient?
 - What are the advantages of co-occurrence?
 - The role of ontologies; why *GO* and *MeSH Disease* terms?
 - What is the impact of the statistical semantic similarity measure? Which measure is the most appropriate?
-

In this work we introduce CASSANDRA; an algorithm for the automated extraction of candidate drug gene associations from biomedical text on the large scale. CASSANDRA combines standardized text mining techniques and biomedical ontologies. It constitutes an unsupervised approach that predicts new drug gene associations solely by systematically analysing the co-occurrence of biomedical terms in the scientific publications indexed by *MEDLINE*.

*Overview of
CASSANDRA*

More specifically, drug and gene names are obtained from the popular and well-established repositories *DrugBank* and *UniProtKB* respectively. The ontological terms belong to the widely used terminologies *Gene Ontology (GO)* and *Medical Subject Headings (MeSH)*. The proposed algorithm utilizes the *Pointwise Mutual Information (PMI)* to rank the most related *GO* and *MeSH Disease* concepts to the drug and the gene respectively. These concepts form an individual profile for each drug and gene. Then, by quantifying the statistical semantic relatedness between these profiles, CASSANDRA assesses and prioritizes the associations between drugs and genes. The degree of semantic similarity between a drug and a gene profile signifies the strength of their association.

CASSANDRA successfully identifies direct drug gene associations with high precision and prioritizes them over indirect associations (i.e., associations wherein the drug affects a certain gene product without necessarily binding physically to it). Validation shows that the algorithm achieves an *Area Under the Curve (AUC)* up to 0.88 for a dataset consisting of *Approved* drugs and *Human* genes. Additionally, the statistical analysis demonstrates that the proposed semantic similarity metric is more efficient compared to traditional measures towards the discrimination of *true* from *false* drug gene associations.

*Results
overview*

The use of co-occurrence based profiles doesn't at all affect the performance of the algorithm. On the contrary, the results show that profiles of ontological terms generated from co-occurrence based statistics have a comparable and, at times better, performance than the manually curated profiles. Evidently, the generated profiles can provide an insight into biomedical properties for drugs and genes and contribute to the inference of associations that might not have been included in a database nor explicitly reported in the literature.

*Co-occurrence
based profiles
are efficient*

Notably, CASSANDRA can support drug repurposing not only in a target based fashion (meaning to propose new genes related to drugs), but also in a disease-based fashion. Indicatively, for all *FDA* approved drugs repositioned over the last 5 years, co-occurrence based profiles were generated from publications before 2009. The analysis shows that the new therapeutic indications are always included in the profiles and rank relatively higher than the rest of the conditions.

*Disease-based
drug
repurposing*

5.1 Drug gene association prediction

In Table 2.3 we provide an overview of methodologies implemented for drug gene association prediction. Evidently, CASSANDRA significantly deviates from traditional drug gene association prediction. The proposed algorithm is among the very few unsupervised methodologies utilized for this task. As shown, it is also among the very few approaches that apply drug gene association prediction on the large scale. Instead of learning from structural and sequence similarity of drugs and genes respectively, CASSANDRA utilizes ontologies and literature data. One major advantage is that, unlike other methods, CASSANDRA doesn't require existing drug target information to predict a new associated gene for a drug.

*Significant
differences
from
traditional
methods*

Due to the unsupervised nature of CASSANDRA, the method needs no training data. This is particularly advantageous in the case of drug gene association prediction, wherein the lack of benchmark datasets poses a significant problem. In fact, there is only one benchmark dataset introduced by Yamanishi et al. on 2008. Since then, the respective dataset has been the predominant means to cross-compare the mainly supervised methods implemented towards drug-target interaction prediction. The works van Laarhoven and Marchiori (2013), Fakhraei et al. (2013) and Gönen (2012) are just a few examples.

*Training
independence*

However, the limitation of such dataset is that it contains only true-positive drug target interactions. The negative interactions are generated by the pairwise combination of all drugs and targets contained in the true-positive dataset, as it has been also the case for the evaluation datasets generated within this study. That is the Achilles' heel in machine learning approaches.

*Dataset
independence*

In the end, the models learn only from positive data, since there is no way to obtain and hence train a classification model over true-negative drug target interactions (Pahikkala et al., 2014; Ding et al., 2013). As a matter of fact, it is such the dependence of some methods on the training dataset that they are unable to predict a drug target interaction between a drug and target that do not have already known interaction information with other targets and drugs respectively. Alaimo et al. (2013), Cheng et al. (2012a), Laarhoven et al. (2011) and Yamanishi et al. (2010) constitute representative examples of such methodologies. CASSANDRA, on the other hand, remains independent of the dataset features; apart from its unsupervised implementation, the proposed algorithm focuses on the prioritization of drug gene associations rather than their binary classification to true and false drug-target interactions. All drug gene pairs are assigned a score which represents the strength of the association, regardless of whether the drug or the gene already participate in known drug target interactions.

An additional advantage stems from CASSANDRA's unsupervised implementation. The algorithm is able to process massive literature data and predict drug gene associations on the large scale. On the other hand, supervised large scale classification is an increasingly *Big Data* problem and so far little has been published towards the practical resolution of this issue (Sun et al., 2014). It has been shown that the performance of *Support Vector Machine* classification on large-scale taxonomies is "far from satisfactory" (Liu et al., 2005). In methods applying the *Bipartite Local Model (BLM)* (e.g., Alaimo et al. 2013, Yu et al. 2012, Perlman et al. 2011) or the *Pairwise Kernel Method (PKM)* (e.g., Takarabe et al. 2012, Jacob and Vert 2008) this issue is intensified (Ding et al., 2013). That explains why the majority of the supervised methods addressing drug target prediction focus on the readily available yet size-restricted dataset of Yamanishi et al. (2008).

The outbreak of machine learning approaches towards drug target prediction received the notice of Pahikkala et al. (2014), who in a recent study examine the quality of the respective methodologies. Pahikkala et al. claim that the striking performance of these methods is unrealistic. They suggest that the problem of drug target interaction prediction should be formulated as a prioritization problem rather than a binary classification problem. They also state that the currently used evaluation datasets are not

Scalability

Current supervised methods are overoptimistic

appropriate for the task at hand and they propose the use of biochemical selectivity assays. Moreover, they experimentally demonstrate 2 more factors that dramatically affect the prediction results of several supervised learning studies (e.g., Mei et al. 2013, Laarhoven et al. 2011). The former pertains to the evaluation procedure; they show that simple cross-validation leads to overoptimistic performance. Lastly, they pinpoint the bias between the training and test sets, meaning the common shared drugs, targets or drug target interactions.

Even so, such an overoptimistic viewpoint towards the evaluation procedure can be found in unsupervised methodologies, as well. Although elementary and on an early stage, the approach of Flake (2010) constitutes a representative example. The idea shares common ground with CASSANDRA but fails to produce meaningful drug gene associations mainly because it applies no filtering on the concepts that form the associations between drugs and genes. Most importantly, in this work, drug gene association prediction is evaluated only on an *in-house* built dataset which includes a 70% percent of zero-scored drug gene associations. Consequently, the proposed method shows an exceptional performance. However, when applied on other datasets the method's efficacy drops dramatically. Already, in Section 4.3.1.1 we show how influencing is the inclusion of zero-scored associations when assessing the algorithm's efficacy. The quality of the evaluation dataset is indeed an issue that requires particular consideration.

Taking all the above into account, it is evident that CASSANDRA constitutes a robust method towards drug gene association prediction. This is also corroborated in Section 4.3.1.1, wherein the efficacy of the algorithm is assessed by its application on a series of differentiated datasets. The formulation of the problem as a prioritization task proves to be more realistic towards drug gene association prediction. CASSANDRA doesn't depend on training and test data, hence it remains bias free. Lastly, following an unsupervised methodology enables to perform drug gene association prediction on the large scale.

5.2 Exploring the literature

Biomedical literature constitutes a valuable source of information. However, it remains unexplored mainly due to its vast volume. With more than 400,000 articles published every year, the task to stay current with the literature could easily occupy 75% of a scientist's working day (Cheng et al., 2008). Investigating hidden associations in the plethora of relationships reported in scientific articles could be considered easily as searching a needle in the haystack. Still, it is highly significant to leverage from that information that is usually not yet included in biomedical repositories. The latter is exactly acknowledged by the very concept of *Literature Based Discovery* and by the variety of tools developed towards the automation of this procedure (see Table 2.1).

*Hidden
information*

CASSANDRA focuses on *Literature Based Discovery*, the (semi)-automatic inference of implicit knowledge out of literature (Weeber et al., 2005). The serendipitous discovery of *Swanson* that related *Fish Oil* to *Raynaud's Syndrome* (Swanson, 1986) remains the core motivation basis of CASSANDRA. The algorithm projects this approach to drug gene association prediction and attempts the systematic and automated retrieval of relevant hypotheses from the biomedical literature.

Unlike existing tools in the domain of *Literature Based Discovery* (see Table 2.1), CASSANDRA takes a step further and deviates from the conventional *ABC* model by incorporating the notion of two intermediate, yet similar concepts *B*. As stated in Cameron et al. (2013), relevant information may exist in longer chains of concepts semantically connected. CASSANDRA generates such longer chained hypotheses and ranks them towards the identification of indirectly connected drugs and genes that would be difficult to uncover without computational assistance or prior knowledge.

*Expanding the
ABC model*

Additionally, CASSANDRA is one of the *Literature Based Discovery* methods that fully utilize ontologies. As shown in Table 2.1, ontologies, although constituting useful and controlled vocabularies, are not extensively used towards the establishment of hypotheses. Most of the tools implemented use their *in-house* defined terminologies and this hinders their general applicability. CASSANDRA uses ontologies to explore the literature and more precisely, it systematically applies the whole range of *GO*

*Harnessing
ontologies*

and *MeSH Disease* terms towards the establishment of hypotheses pertaining to drug and gene relations. An additional feature that differentiates CASSANDRA from the majority of *Literature Based Discovery* tools is the statistical refinement of the associations that establish the drug gene predictions. Besides using a probabilistic tool (i.e., *nPMI*) to establish these associations, CASSANDRA applies a further filtering to insure the quality of the generated hypotheses.

Moreover, CASSANDRA constitutes one of the few methods that solely utilize the biomedical literature to generate drug gene association predictions (see Table 2.3). Zhu et al. (2005) learn from gene-gene, compound-compound and gene-compound co-occurrence data in a preselected set of *MEDLINE* abstracts and suggest implicit compound-gene associations. Wu et al. (2012) collect drug gene co-occurrence data on a subset of *MEDLINE* abstracts and utilize *Latent Dirichlet Allocation* to prioritise them. However, both methodologies do not attempt a large scale drug gene association prediction. Indeed, only a small subset of *MEDLINE* records is used in each case ($\sim 0.4\%$ and 1.0% respectively). In the study of Zhu et al. (2005) this is somewhat explainable, since supervised approaches are generally difficult to scale (Wang et al., 2008). Additionally, Wu et al. (2012) focus on the prioritization of directly co-occurring drugs and genes, while CASSANDRA excludes such pairs and focuses on the prediction of indirectly related drugs and genes.

*Few
literature-based
approaches*

Utilizing the biomedical literature lends CASSANDRA an additional important feature. Unlike several mainly supervised strategies towards drug target interaction prediction, CASSANDRA is able to provide ranked lists of associated genes for drugs with no known targets and ranked lists of associated drugs for non-coding (or at least reported as such) target genes. Of course, the only prerequisite is that the existing literature enables the generation of statistical significant co-occurrence based profiles for these drugs and genes.

*Prediction for
all drugs and
genes*

5.2.1 Focusing on abstracts and titles

When it comes to text mining applications the arising question is: "Which text to mine?" CASSANDRA utilizes co-occurrence data from *MEDLINE*

indexed abstracts and titles. Notably, there are many tools utilizing the same resource towards the extraction of biomedical information (Gijón-Correas et al., 2014; Kim et al., 2013b; Fontaine et al., 2011; Frijters et al., 2010). The main reason for that is their public accessibility (Vincze et al., 2008). Only $\sim 2\%$ of *MEDLINE* entries have open-access full-text articles available for text mining (Thomas et al., 2012).

Another reason is that the experimental procedures or results described in full-text articles are hard for researchers to re-use (Névéol et al., 2011). In a recent study, Fontelo et al. (2013) also state that abstracts appear to be equally informative as full-text articles. Most importantly, it has been demonstrated that text mining tools perform better in abstracts than on article bodies (Cohen et al., 2010b). This can be possibly attributed to the fact that, unlike full-text articles, abstracts contain less hedgy sentences (Fontelo et al., 2013), and state clearly the respective research findings (Jenssen et al., 2001). All the arguments stated above opt for the use of abstracts over full-text articles. Still, this doesn't necessarily rule out the additional use of full-text articles by the suggested methodology in the future.

*Abstracts
deliver
information*

5.2.2 Using co-occurrence

Co-occurrence is often applied in biomedical relation extraction and particularly in *Literature Based Discovery* (Paik et al., 2014; Cheung et al., 2013; Tsuruoka et al., 2011; Frijters et al., 2010; Yildiz and Pratt, 2006). Another way to establish such relations would be the use of predefined or automatically derived rule/patterns (Chang et al., 2012; Cou, 2010). Currently, both strategies are used towards the extraction of biomedical relations, either individually or combined (Xu and Wang, 2014). The proposed algorithm utilizes co-occurrence based statistics to establish relations between the biomedical terms on the large scale.

Usually, the main argument against the use of co-occurrence based statistics is the quality or even existence of the returned associations in the first place (Zweigenbaum et al., 2007). However, the analysis conducted within this study suggests otherwise. The quality of the relations generated is comparable to the quality of manually curated associations. As it has

*Co-occurrence
is meaningful*

been shown (Chapter 4, Section 4.3.1.3) ontological profiles for genes and drugs derived from textual co-occurrence demonstrate equal or at times better performance than manually curated profiles. Similar is the finding of Garten et al. (2010). The authors replace a manually curated drug gene association network with a co-occurrence derived network and show that the performance of their algorithm remains the same or even improves at certain cases.

In an early study, Jenssen et al. (2001) build a gene-gene relationship network by weighting the co-occurrences of gene-gene pairs in *MEDLINE* abstracts. They state that the co-occurrences of biological entities in scientific abstracts do reflect meaningful relationships, due to the condensed information and clear statements of the research findings they contain. A recent work also demonstrates that abstract level co-occurrence strongly correlates with sentence level co-occurrence (Niu et al., 2010). Zhu et al. (2005) attribute any false positive co-occurrence relationships to the errors introduced by the *Name Entity Recognition (NER)* systems.

The co-occurrence degree is also proved to be correlated with the quality of the association (Jenssen et al., 2001). Moreover, another study shows that less than 10% of the sentences "contains a modifier that radically influences the semantic content of the sentence", meaning a hedgy expression or a negation (Vincze et al., 2008). These facts explain why the performance of CASSANDRA is hardly affected by any false positive relations detected. Given that the negative associations are underrepresented in literature (Pérez et al., 2004) and therefore would result to a low *nPMI* score, no relevant filtering is necessary.

Another reason for applying co-occurrence based statistics is their high recall (Zweigenbaum et al., 2007). That is why, when combined with other methodologies, there is a significant boost in the performance. Indicatively, Aubry et al. (2006) demonstrate that introducing co-occurrence statistics between genes and *GO* terms radically improves the gene functional annotation. Seki and Mostafa (2007) build a network of genes, *MeSH* and *GO* terms to discover indirect associations between genes and diseases. When integrating textual co-occurrence from *MEDLINE* abstracts, the system's

*Underepresented
negative
associations*

*Co-occurrence
has high recall*

predictive power improves by 4.6%. Thereafter, co-occurrence based statistics are at this point preferred over the supervised or not use of predefined patterns/rules for the extraction of relationships between biomedical terms.

Moreover, it is easier to apply co-occurrence methods on the large scale. Using a mathematical schema to model co-occurrences is definitely more straightforward than the application of pattern based strategies. Clearly, the generation of patterns (especially when this is conducted manually) is a laborious task (Huang et al., 2004). Additionally, it requires thorough study of the respective domain. For example, in this study, were for pattern-based approaches to be followed, 7 different syntactic patterns would have to be generated corresponding to the 7 different types of relations, i.e., drug/gene-*MeSH Disease*, drug/gene-*GO*, *MeSH Disease- MeSH Disease*, *MeSH Disease-GO* and *GO- GO* patterns. On the other hand, co-occurrences are generally applicable (domain-independent) and need no text-preprocessing (Zweigenbaum et al., 2007).

Scalability

Of course, co-occurrence based strategies do not provide the type of relation and that is another argument against their usage. Indeed, when focusing on *first-order* relation extraction (i.e., the relation between a term *A* and a term *B*), it might be interesting to know the exact type of relation. However, for *second-order* relation extraction (i.e., the *ABC* model) strategies, such as CASSANDRA, even if the relations are known, they still have to be concordant so as to generate a consistent hypothesis. In other words, in both cases the hypotheses have to be further analyzed. However, in the case of CASSANDRA *nPMI* already prioritizes the stronger hypotheses and this significantly facilitates the procedure.

Type of relation

To sum up, co-occurrence based statistics have high recall, they are relatively easy to apply on the large scale, they are domain-independent and the associations they suggest do reflect meaningful relations between the biomedical entities (Jenssen et al., 2001). The latter is also demonstrated by the results of this very study; co-occurrence based profiles show a comparable or at times better performance than the manually curated profiles. For these reasons, co-occurrence based statistics were at this point chosen over the generation of syntactic patterns or rules towards the extraction of relations. Still, we think that the two approaches are complementary

and consider the use of patterns towards the refinement of drug gene association hypotheses generated by CASSANDRA. Indicatively, there are several works combining the two methodologies towards an enhanced quality of results (Xu and Wang, 2014; Bunescu et al., 2006).

5.3 The role of ontologies

Ontologies have a critical role in the representation of knowledge and the dissemination of data in the biomedical domain. They are widely used in data indexing and information retrieval (Whetzel et al., 2011), but also in data integration and reasoning (Ashburner et al., 2000; Magka et al., 2014). Therefore, it is fitting to utilize ontologies for connecting islands of drug and gene data, as it is the task of this study. The arising question is which is the most appropriate ontology for this task?

There is a plethora of biomedical ontologies of varied granularity and dedicated to different biomedical subdomains, (e.g., Barton et al. (2014) or Zheng et al. (2013)). Still, for the task at hand the goal is to focus on the ontological terms that would be adequate to efficiently describe drugs and genes, capture their functional properties and hence, ensure the successful estimation of similarity between their profiles. In the case of genes the decision is quite straightforward regarding the use of *Gene Ontology* (Ashburner et al., 2000). *Gene Ontology* is an extensively used terminology that addresses the need for consistent descriptions of gene products across databases. It encompasses terms that represent biological processes, cellular components and molecular functions. Obviously, these terms can be used to build the context that encloses the processes and functionalities which relate to drugs.

Why GO?

Evidently, drugs and genes are strongly related to diseases. Hence, the use of terms stemming from a medical vocabulary, such as *Medical Subject Headings* (*MeSH*), is also suitable for their description. *MeSH* constitutes a widely accepted use case of a clinical controlled vocabulary. It is manually curated, regularly updated and it is used to index the articles stored in *MEDLINE*. Unlike other terminologies (e.g., *LOINC* or *GALEN*), *MeSH* is suitable for the task at hand because it focuses and has proven to be

Why MeSH?

particularly successful towards medical literature retrieval (Nelson, 2009). *SNOMED CT*, on the other hand, which is another popular terminology, focuses on the representation and encoding of clinical data for the *Electronic Health Records (EHR)*. Moreover, it contains concepts with subtle differences in meaning which are difficult to discriminate (Chiang et al., 2006). This can be partially attributed to post-coordination, meaning that a concept can be coded by a coordination of different codes (Stevens and Sattler, 2013). Although post-coordination confers a dynamic structure in *SNOMED CT*, it also results in many ambiguous and context-dependent concepts, the resolution of which requires the application of reasoning systems. *MeSH*, on the other hand, is a static terminology of readily available terms.

MeSH contains 16 trees overall and more than 27,000 terms (descriptors). Apart from the *Diseases* tree, *MeSH* includes categories such as *Phenomena and Processes*, *Information Science*, *Publication Characteristics*, *Geographicals* or *Disciplines and Occupations*. The majority of them is not relevant and, thus, constitutes a source of error for text mining. On the other hand, *GO* is also large, but entirely focused on the biological processes and functions. For that reason, *GO* was fully used, whilst from *MeSH* only the *Diseases* tree was utilized.

Clearly, the application of *MeSH Diseases* and *GO* terms has proven successful towards the drug gene association prediction from literature, as shown by the overall performance of CASSANDRA (Chapter 4). More specifically, *GO* successfully discriminates true from false drug gene associations even when the literature for the respective entities is limited (see Section 4.3.1.1). Co-occurrence based profiles containing *GO* terms also perform better than manually curated *MeSH Disease* profiles, illustrating that the type of terms plays a significant role towards the establishment of associations. But what is special about *GO* terms? *GO* is entirely dedicated to the sufficient and accurate representation of genes, gene products and their attributes. Hence, the terms included in *GO* are by nature highly descriptive and precise. *MeSH Diseases*, on the other hand, constitute concepts of a broader perspective, i.e, a disease encompasses a set of functions, processes, signs and symptoms to which many drugs and genes could be associated. That difference between the two terminologies is reflected on the algorithmic performance.

*Why only
Diseases?*

*GO
outperforms
MeSH
Diseases*

5.3.1 Estimating the semantic similarity

The means to estimate the semantic relatedness between a drug and a gene ontological profile is critical towards assessing the strength of the drug gene association. Hence, which is the most appropriate semantic similarity metric for the task at hand?

CASSANDRA utilizes the statistical semantic similarity, meaning the *normalized Pointwise Mutual Information (nPMI)*. *nPMI* is a fully corpus-based, probabilistic measure that treats the ontological concepts just as a set of terms. The similarity between two concepts is estimated solely based on the degree of their co-occurrence in *MEDLINE* indexed titles and abstracts.

One major advantage of *nPMI* over other similarity metrics is its general applicability. Several metrics are built to explicitly identify relations between terms of a specific ontology (e.g., *GO* as in Yang et al. 2012; Jain and Bader 2010). For example, Wang et al. (2007) propose a metric wherein they exploit the *GO is_part_of* and *is_a* relations. On the contrary, *nPMI* is ontology-independent and can be applied on both *MeSH Diseases* and *GO* terms or any other ontologies incorporated in the algorithm in the future.

*Ontology
independence*

To assess the efficiency of *nPMI*, we compared the metric to two traditional measures of semantic similarity, i.e., *Wu-Palmer* (1994) and *Lin* (1998). As shown in the results (Chapter 4, Section 4.3.2) *nPMI* demonstrated the best performance, in terms of *Area Under the Curve*, *Precision-Recall* and discriminative power. Unlike the *Wu-Palmer* and *Lin* metrics, *nPMI* can capture the similarity between terms that belong to different subontologies or even different ontologies, e.g., terms from *GO Biological Process* and *GO Molecular Function* or *GO* and *MeSH Disease* terms. Furthermore, the application of *nPMI* enables the identification of latent relations between ontological terms that do not necessarily occur very frequently in text. This is of particular importance, since the goal of CASSANDRA is to uncover the hidden information that lies behind unknown drug gene associations.

*Capturing
latent relations*

Chapter 6

Conclusion

This chapter discusses the contribution and future directions of CASSANDRA. It summarizes the important findings of this study and proposes certain steps towards the improvement and expansion of the suggested methodology.

6.1 Contribution

The current thesis introduces CASSANDRA. An unsupervised corpus-based algorithm that addresses the prediction of drug gene associations from literature. The algorithm is an extended version of traditional *Literature Based Discovery*. It explores the biomedical literature and interrelates drug to genes via intermediate ontological concepts.

In a nutshell,

- 23,487,871 *MEDLINE* abstracts and titles were annotated with drugs, genes, *GO* and *MeSH Disease* terms
- co-occurrence based profiles were precomputed for 2,837 drugs and 57,395 genes
- 5 new drug gene association datasets were compiled to systematically assess the algorithm's efficacy
- overall 37,603,728 distinct drug gene associations were scored

The method generates co-occurrence based concept profiles for both drugs and genes. *Normalized Pointwise Mutual Information (nPMI)* is used to create the profiles by finding statistical significant associated ontological terms in *MEDLINE* titles and abstracts. *nPMI* is also utilized as measure of semantic similarity to estimate the relatedness between drugs and genes via their profiles.

Use of nPMI

The application of CASSANDRA towards drug gene association prediction and the identification of drug repurposing cases has been demonstrated. Regarding the drug gene association prediction, the performance of the algorithm is evaluated on 6 datasets. It has to be pinpointed that all drug-gene pairs consisting of co-occurring drugs and genes are removed from the datasets. Results show that the suggested method is robust and its performance is independent from the size and type of the dataset. Notably, CASSANDRA achieves an *AUC* up to 0.88 in prioritizing true associations between *Approved* drugs and *Human* genes. Considering the overoptimistic results provided by supervised techniques which reach at

Good performance

times an *AUC* up to 0.93 (Pahikkala et al., 2014), CASSANDRA’s performance is quite striking.

The evaluation revealed further interesting properties of CASSANDRA. First, it can successfully prioritise direct (i.e., physically binding) from indirect drug gene associations, as shown by the *AUC* values obtained on the datasets *CTD_{Binding}* and *CTD_{Related}*. Second, the use of co-occurrence based profiles derived from literature doesn’t affect at all the performance of CASSANDRA. In particular, when using *GO* terms, the co-occurrence based profiles outperformed the manually curated *MeSH Disease* profiles. This suggests that the co-occurrence based profiles derived from the biomedical literature are indeed reliable for associating drugs to genes.

With regards to drug gene association prediction, three prediction cases were further analyzed. Two of them correspond to the drugs *Cathine* and *Fenethylamine* for which there are no known targets, so far. CASSANDRA predicts two strongly associated genes, one for each drug; *GHRL* and *ApoE* respectively. The predicted genes rank among the first 0.4 % of the ordered list of genes scored against each drug. To assess the quality of these predictions, we dugged into the literature and manually retrieved concrete pieces of evidence which indeed suggest that the respective associations are meaningful. An additional case study was investigated in the same fashion, this time between the known drug *Milnacipran* and its known associated gene *SLC6A4* that codes for its target. The results reveal that CASSANDRA did consider the new therapeutic indication *Fibromyalgia* for the establishment of the association, despite the fact that the drug’s efficacy on the respective disease was at the time unknown.

The efficacy of CASSANDRA towards drug repurposing was further evaluated. A dataset comprising all known drug repositioning cases that were *FDA* approved within the last 5 years was compiled. For these drugs the generated profiles are based on the literature data before the year of approval of each repositioning case. We demonstrated that the respective profiles always include the new indications, and that in the majority of the cases these indications are ranked high among drugs’ profile terms.

*Co-occurrence
produces
meaningful
predictions*

*Textual
evidence
confirm
predictions*

*New
therapeutic
indications*

6.2 Future directions

There are certain directions to improve further the results generated by CASSANDRA. The first is to alleviate the error introduced by the annotation methodologies which are responsible to a big extent for the existence of false positive drug gene associations (Zhu et al., 2005; Jenssen et al., 2001). Perhaps the use of additional annotators could further enhance the quality of results if the interagreement space of them were to be considered.

*Improve
annotation*

Towards the same direction, an interesting extension is to use more drug/-compound related dictionaries. CASSANDRA is currently limited only in drugs reported in the *DrugBank* database. Although *DrugBank* consists a high quality resource of drug target interaction data and the respective *in-house* dictionary demonstrates high performance, still there is the possibility to harness new chemical taggers that have been recently developed. For example, *ChemSpot* by Rocktäschel et al. (2012) or *OSCAR4* by Jessop et al. (2011) have shown encouraging results towards the annotation of chemicals in text. The tricky issue of the recognition of *IUPAC* drug names in text is not extensively studied but there are also some works that CASSANDRA could benefit from, such as the supervised approach proposed by Klinger et al. (2008).

*More chemical
dictionaries*

Regarding the profile generation, as shown in Chapter 4, the co-occurrence based profiles proved highly efficient. However, one could not deny that the information existing in public repositories is also of value and can contribute to the generation of high quality drug gene association hypotheses. Consequently, a further improvement would be to enhance the quality of the profiles with ontological concepts found in biomedical repositories. For example, it would be interesting to investigate whether the performance improves if the manually curated and co-occurrence based profiles are merged in the case of *MeSH Diseases*. Regarding *GO* terms, the *Gene Ontology Annotation (GOA)* could also enhance the profiles with interesting associations. Moreover, the profiles of drugs and genes could be also enriched with physicochemical properties and additional protein structural information respectively. Another suggestion would be to exploit pharmacogenomic data that is aggregated and represented in the form of networks (Daminelli et al., 2012).

*Integrate the
profiles*

Abstracts and titles from *MEDLINE* have proven a high quality resource for the identification of relationships between the biomedical entities. However, despite the limited space of available full-text articles (Thomas et al., 2012), it would be interesting to examine if they provide relations that can further improve the prioritization of drug gene associations. Nevertheless, this has to be taken with a grain of salt, since the percentage of "negation" and hedgy sentences increases in the article bodies (Fontelo et al., 2013). So far, CASSANDRA utilizes co-occurrences on the abstract-level. An interesting expansion is to take one step further and use sentence-level co-occurrence information. Although the majority of biomedical entities co-occurring in abstracts, also co-occur in sentences (Niu et al., 2010) and hence recall remains unaffected, still an increase in the precision would be expected. Of course, in such a case the whole application of the *nPMI* model would have to be reformulated.

*Mine a
different text*

The next step towards literature-based drug gene association prediction would be the use of patterns/rules that would interpret the hypotheses. A rough suggestion is to automatically retrieve the text in between the biomedical entities, which then in turn can be either inspected manually or mined to extract explicit associations. However, such an extension is far more feasible if automatic pattern generation is followed, since patterns are domain dependent and laborious to create. Nevertheless, the use of patterns isn't enough to consolidate a drug gene association derived from text. Rule based approaches to unravel the concordant relations could alleviate the problem, but also add to the complexity of the task significantly.

*Pattern/rule
application*

Is it worthy to seek such a laborious and complicated direction to further establish the hypotheses that CASSANDRA generates? The associations provided by CASSANDRA are derived from text. They are in fact the representation of a signal that a drug and a gene are potentially associated. Instead of dedicating hours of research in literature to discover hidden relations and common islands of data between drugs and genes, CASSANDRA speeds up the process and produces the respective signal. That is, by nature, complementary evidence to other approaches towards drug-target interaction prediction. The results of CASSANDRA can be furtherly examined via large scale molecular docking analysis as in Li et al. (2011). Lately, these approaches become more and more popular. In a recent study, it is suggested that the binding site similarity is

*Complementarity
with other
approaches*

the key to identify promiscuous drugs and boost drug repurposing. (Haupt et al., 2013). CASSANDRA could significantly benefit from such analyses towards the refinement of the proposed associations without necessarily having to further process the text that underlies them.

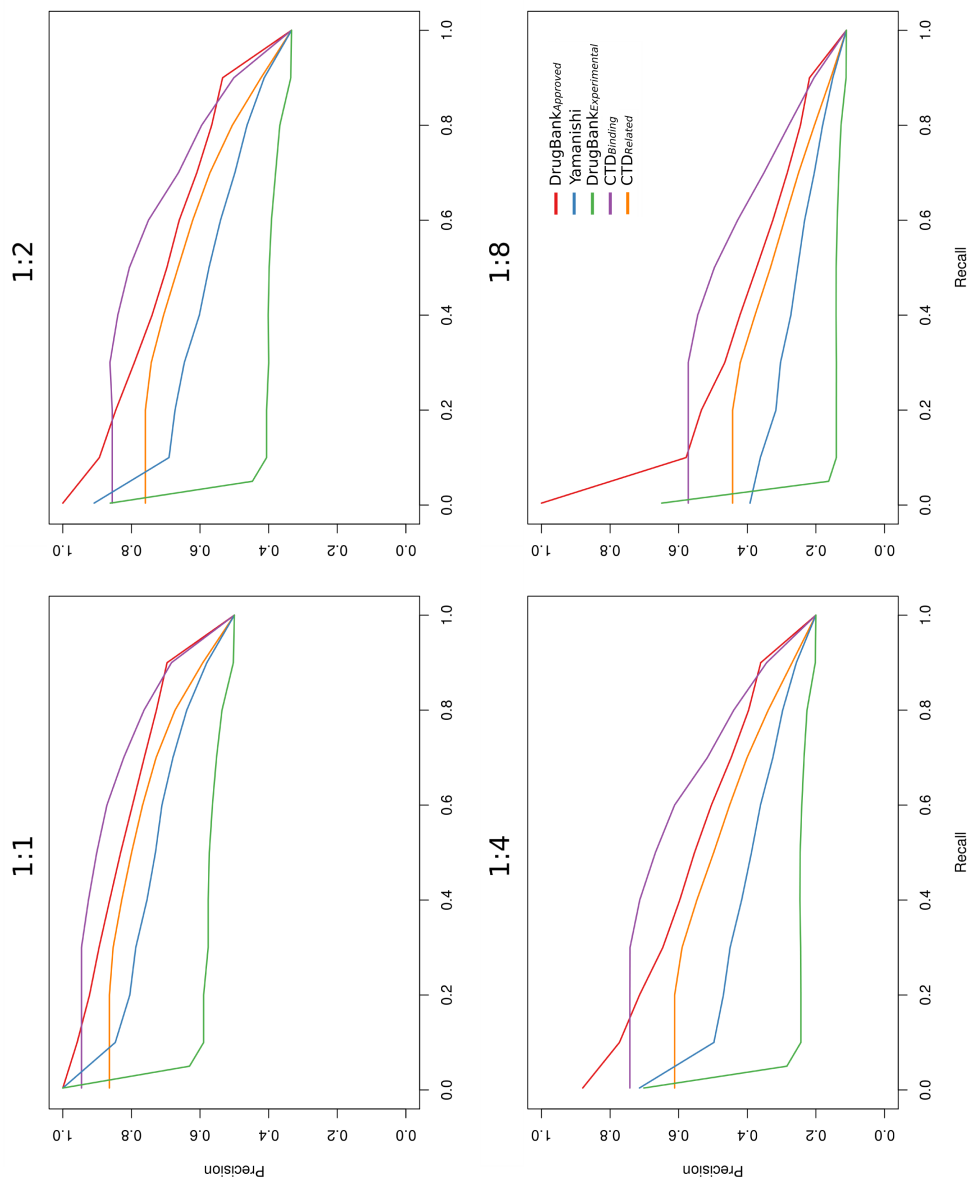
Chapter 7

Supplementary Material

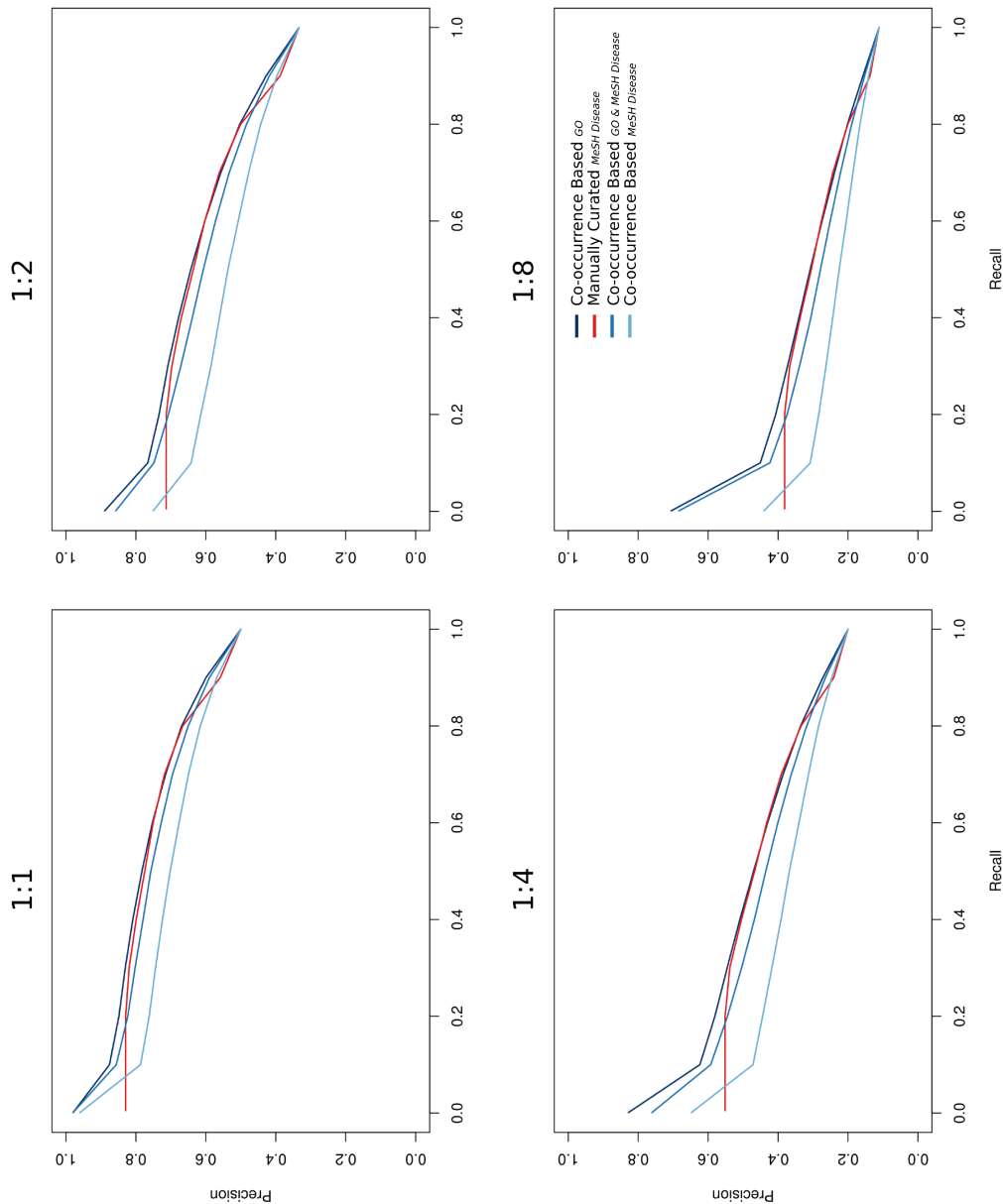
Table 7.1: *Arithmetic means* for different semantic similarity metrics

		<i>nPMI</i>	<i>Lin</i>	<i>Wu-Palmer</i>
<i>DrugBank_{Approved}</i>	true	0.553	0.730	0.835
	false	0.306	0.487	0.625
	distance	0.247	0, 243	0, 210
<i>DrugBank_{Experimental}</i>	true	0.358	0.470	0.601
	false	0.306	0.489	0.622
	distance	0, 052	-0, 019	-0, 021
<i>Yamanishi</i>	true	0.590	0.771	0.876
	false	0.486	0.722	0.841
	distance	0, 104	0, 049	0, 035

The table shows the *arithmetic means* for the score distributions of true and false drug gene associations that every metric of semantic similarity achieves on the respective dataset. The difference between the *means* of true and false distribution is provided. Clearly, *nPMI* achieves the highest difference on every dataset. When applied on the *DrugBank_{Experimental}* dataset, *Wu-Palmer* and *Lin* completely fail the discrimination task.

Figure 7.1: All ratios *PR* curves for all datasets

All *PR* curves for different ratios of imbalance between true and false drug gene associations contained in the evaluation datasets are shown. The curves are consistent with the respective *ROC* curves. Again, CASSANDRA performs better on the datasets *DrugBankApproved* and *CTDBinding* than the *DrugBankExperimental* and *CTDRelated* respectively. Increasing the number of false drug gene associations significantly affects the *DrugBankExperimental* due the highly similar score distributions that exist between true and false drug gene pairs.

Figure 7.2: All ratios *PR* curves - Manually curated vs. co-occurrence based profiles

All *PR* curves for different ratios of imbalance between true and false drug gene associations are shown. Co-occurrence based profiles of *GO* terms demonstrate the best performance across all ratios. Manually curated *MeSH Disease* profiles perform similar to the co-occurrence joined profiles. Doubling or quadrupling the size of false drug gene associations doesn't affect significantly the algorithm's efficacy.

Bibliography

- Steven M. Paul, Daniel S. Mytelka, Christopher T. Dunwiddie, Charles C. Persinger, Bernard H. Munos, Stacy R. Lindborg, and Aaron L. Schacht. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery*, 9(3):203–214, March 2010.
- M. R. Hurle, L. Yang, Q. Xie, D. K. Rajpal, P. Sanseau, and P. Agarwal. Computational drug repositioning: From data to therapeutics. *Clinical Pharmacology & Therapeutics*, 93(4):335–341, April 2013.
- Christos Andronis, Anuj Sharma, Vassilis Virvilis, Spyros Deftereos, and Aris Persidis. Literature mining, ontologies and information visualization for drug repositing. *Briefings in Bioinformatics*, 12(4):357–368, July 2011.
- D R Swanson. Fish oil, raynaud's syndrome, and undiscovered public knowledge. *Perspectives in biology and medicine*, 30(1):7–18, 1986.
- R A DiGiacomo, J M Kremer, and D M Shah. Fish-oil dietary supplementation in patients with raynaud's phenomenon: a double-blind, controlled, prospective study. *The American journal of medicine*, 86(2):158–164, February 1989.
- O. Bodenreider. Biomedical ontologies in action: Role in knowledge management, data integration and decision support. *Yearbook of medical informatics*, pages 67–79, 2008.
- Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, May 2000.
- A. Srinivas Reddy and Shuxing Zhang. Polypharmacology: drug discovery for the future. *Expert review of clinical pharmacology*, 6(1), January 2013.

- Ted T. Ashburn and Karl B. Thor. Drug repositioning: identifying and developing new uses for existing drugs. *Nature Reviews Drug Discovery*, 3(8):673–683, August 2004.
- Barri J. Fessler, Graciela S. Alarcón, Gerald McGwin, Jeffrey Roseman, Holly M. Bastian, Alan W. Friedman, Bruce A. Baethge, Luis Vilá, John D. Reveille, and for the LUMINA Study Group. Systemic lupus erythematosus in three ethnic groups: Xvi. association of hydroxychloroquine use with reduced risk of damage accrual. *Arthritis & Rheumatism*, 52(5):1473–1480, 2005.
- Jean-Paul Richalet, Pierre Gratadour, Paul Robach, Isabelle Pham, Michèle Déchaux, Aude Joncquiart-Latarjet, Pascal Mollard, Julien Brugniaux, and Jérémy Cornolo. Sildenafil inhibits altitude-induced hypoxemia and pulmonary hypertension. *American Journal of Respiratory and Critical Care Medicine*, 171(3):275–281, February 2005.
- Takumi Ito, Hideki Ando, Takayuki Suzuki, Toshihiko Ogura, Kentaro Hotta, Yoshimasa Imamura, Yuki Yamaguchi, and Hiroshi Handa. Identification of a primary target of thalidomide teratogenicity. *Science*, 327(5971):1345–1350, 2010.
- Iris Breitzkreutz and Kenneth C Anderson. Thalidomide in multiple myeloma – clinical trials and aspects of drug metabolism and toxicity. *Expert Opinion on Drug Metabolism & Toxicology*, 4(7):973–985, July 2008.
- Zaina P. Qureshi, Enrique Seoane-Vazquez, Rosa Rodriguez-Monguio, Kurt B. Stevenson, and Sheryl L. Szeinbach. Market withdrawal of new molecular entities approved in the united states from 1980 to 2009. *Pharmacoepidemiology and Drug Safety*, 20(7):772–777, July 2011.
- Psaty, Bruce M., Furberg, Curt D., Ray Wayne A., and Weiss Noel S. Potential for conflict of interest in the evaluation of suspected adverse drug reactions: Use of cerivastatin and risk of rhabdomyolysis. *JAMA*, 292(21):2622–2631, December 2004.
- Hyun Uk Kim, Soo Young Kim, Haeyoung Jeong, Tae Yong Kim, Jae Jong Kim, Hyon E Choy, Kyu Yang Yi, Joon Haeng Rhee, and Sang Yup Lee. Integrative genome-scale metabolic analysis of vibrio vulnificus for drug targeting and discovery. *Molecular Systems Biology*, 7(1), 2011.
- Aimee L. Jackson and Peter S. Linsley. Recognizing and avoiding siRNA off-target effects for target identification and therapeutic application. *Nat Rev Drug Discov*, 9(1):57–67, January 2010.

- Ulrich AK Betz, Ronald Farquhar, and Karl Ziegelbauer. Genomics: success or failure to deliver drug targets? *Current Opinion in Chemical Biology*, 9(4): 387 – 391, 2005. Next-generation therapeutics.
- Justin Lamb, Emily D. Crawford, David Peck, Joshua W. Modell, Irene C. Blat, Matthew J. Wrobel, Jim Lerner, Jean-Philippe Brunet, Aravind Subramanian, Kenneth N. Ross, Michael Reich, Haley Hieronymus, Guo Wei, Scott A. Armstrong, Stephen J. Haggarty, Paul A. Clemons, Ru Wei, Steven A. Carr, Eric S. Lander, and Todd R. Golub. The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313(5795):1929–1935, September 2006.
- Marina Sirota, Joel T. Dudley, Jeewon Kim, Annie P. Chiang, Alex A. Morgan, Alejandro Sweet-Cordero, Julien Sage, and Atul J. Butte. Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Science Translational Medicine*, 3(96):96ra77–96ra77, August 2011.
- Joel T. Dudley, Marina Sirota, Mohan Shenoy, Reetesh K. Pai, Silke Roedder, Annie P. Chiang, Alex A. Morgan, Minnie M. Sarwal, Pankaj Jay Pasricha, and Atul J. Butte. Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. *Science Translational Medicine*, 3(96): 96ra76, 2011a.
- Ling Jin, Jian Tu, Jianwei Jia, Wenbin An, Huanran Tan, Qinghua Cui, and Zhixin Li. Drug-repurposing identified the combination of trolox c and cytosine for the treatment of type 2 diabetes. *Journal of Translational Medicine*, 12: 153, May 2014.
- Michael J. Keiser, Bryan L. Roth, Blaine N. Armbruster, Paul Ernsberger, John J. Irwin, and Brian K. Shoichet. Relating protein pharmacology by ligand chemistry. *Nature Biotechnology*, 25(2):197–206, February 2007.
- Monica Campillos, Michael Kuhn, Anne-Claude Gavin, Lars Juhl Jensen, and Peer Bork. Drug target identification using side-effect similarity. *Science*, 321 (5886):263–266, July 2008.
- Eugen Lounkine, Michael J. Keiser, Steven Whitebread, Dmitri Mikhailov, Jacques Hamon, Jeremy L. Jenkins, Paul Lavan, Eckhard Weber, Allison K. Doak, Serge Côté, Brian K. Shoichet, and Laszlo Urban. Large-scale prediction and testing of drug activity on side-effect targets. *Nature*, 486(7403): 361–367, June 2012.

- Joel T. Dudley, Tarangini Deshpande, and Atul J. Butte. Exploiting drug-disease relationships for computational drug repositioning. *Briefings in Bioinformatics*, 12(4):303–311, January 2011b.
- Marc Weeber, Jan A. Kors, and Barend Mons. Online tools to support literature-based discovery in the life sciences. *Briefings in Bioinformatics*, 6(3):277–286, September 2005.
- GW Tietjen, S Chien, E Leroy, I Gavras, H Gavras, and FE Gump. Blood viscosity, plasma proteins, and raynaud syndrome. *Archives of Surgery*, 110(11):1343–1346, November 1975.
- B E Woodcock, E Smith, W H Lambert, W M Jones, J H Galloway, M Greaves, and F E Preston. Beneficial effect of fish oil on blood viscosity in peripheral vascular disease. *British Medical Journal (Clinical research ed.)*, 288(6417):592–594, February 1984.
- D R Swanson. Somatomedin c and arginine: implicit connections between mutually isolated literatures. *Perspectives in Biology and Medicine*, 33(2):157–186, 1990.
- D R Swanson and N R Smalheiser. Indomethacin and alzheimer’s disease. *Neurology*, 46(2), 1996.
- D R Swanson and N R Smalheiser. Calcium-independent phospholipase a2 and schizophrenia. *Archives of General Psychiatry*, 55(8), 1998.
- D R Swanson. Migraine and magnesium: eleven neglected connections. *Perspectives in Biology and Medicine*, 31(4):526–557, 1988.
- Alexander Mauskop and Jasmine Varughese. Why all migraine patients should be treated with magnesium. *Journal of Neural Transmission*, 119(5):575–579, 2012.
- Delroy Cameron, Olivier Bodenreider, Hima Yalamanchili, Tu Danh, Sreeram Vallabhaneni, Krishnaprasad Thirunarayan, Amit P. Sheth, and Thomas C. Rindflesch. A graph-based recovery and decomposition of swanson’s hypothesis using semantic predications. *Journal of Biomedical Informatics*, 46(2):238–251, 2013.
- Trevor Cohen, G. Kerr Whitfield, Roger W. Schvaneveldt, Kavitha Mukund, and Thomas Rindflesch. Epiphanet: An interactive tool to support biomedical discoveries. *Journal of biomedical discovery and collaboration*, 5:21–49, 2010a.

- Padmini Srinivasan. Text mining: Generating hypotheses from medline. *Journal of the American Society for Information Science and Technology*, 55(5):396–413, 2004.
- Marc Weeber, Henry Klein, Lolkje T. W. de Jong-van den Berg, and Rein Vos. Using concepts in literature-based discovery: Simulating swanson’s raynaud-fish oil and migraine-magnesium discoveries. *J. Am. Soc. Inf. Sci. Technol.*, 52(7):548–557, may 2001.
- Ruggero Gramatica, T. Di Matteo, Stefano Giorgetti, Massimo Barbiani, Dorian Bevec, and Tomaso Aste. Graph theory enables drug repurposing –how a mathematical model can drive the discovery of hidden mechanisms of action. *PLoS ONE*, 9(1):e84912, 01 2014.
- Weiwei Dong, Yixuan Liu, Weijie Zhu, Quan Mou, Jinliang Wang, and Yi Hu. Simulation of swanson’s literature-based discovery: Anandamide treatment inhibits growth of gastric cancer cells in vitro and in silico. *PLoS ONE*, 9: e100436, 06 2014.
- Nancy C. Baker and Bradley M. Hemminger. Mining connections between chemicals, proteins, and diseases extracted from medline annotations. *Journal of Biomedical Informatics*, 43(4):510 – 519, 2010.
- Caroline B. Ahlers, Dimitar Hristovski, Halil Kilicoglu, and Thomas C. Rindfleisch. Using the literature-based discovery paradigm to investigate drug mechanisms. In *AMIA 2007, American Medical Informatics Association Annual Symposium, Chicago, IL, USA, November 10-14, 2007*, 2007.
- Padmini Srinivasan and Bisharah Libbus. Mining medline for implicit links between dietary substances and diseases. *Bioinformatics*, 20:290–296, 2004.
- Jonathan D. Wren, Raffi Bekeredjian, Jelena A. Stewart, Ralph V. Shohet, and Harold R. Garner. Knowledge discovery by automated identification and ranking of implicit relationships. *Bioinformatics (Oxford, England)*, 20(3):389–398, feb 2004.
- Marc Weeber, Rein Vos, Henny Klein, Lolkje T. De Jong-Van Den Berg, Alan R. Aronson, and Grietje Molema. Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide. *Journal of the American Medical Informatics Association : JAMIA*, 10(3):252–259, 2003.

- Neil R Smalheiser and Don R Swanson. Using arrowsmith: a computer-assisted approach to formulating and assessing scientific hypotheses. *Computer Methods and Programs in Biomedicine*, 57(3):149 – 153, 1998.
- Y. Tsuruoka, M. Miwa, K. Hamamoto, J. Tsujii, and S. Ananiadou. Discovering and visualizing indirect associations between biomedical concepts. *Bioinformatics*, 27(13):i111–i119, 2011.
- Raoul Frijters, Marianne van Vugt, Ruben Smeets, René van Schaik, Jacob de Vlieg, and Wynand Alkema. Literature mining for the discovery of hidden connections between drugs, genes and diseases. *PLoS Comput Biol*, 6(9): e1000943, 09 2010.
- Dimitar Hristovski, Borut Peterlin, Joyce A. Mitchell, and Susanne M. Humphrey. Using literature-based discovery to identify disease candidate genes. *International Journal of Medical Informatics*, 74(2–4):289 – 298, 2005.
- Michael J. Cairelli, Christopher M. Miller, Marcelo Fiszman, Terri Workman, and Thomas C. Rindflesch. Semantic MEDLINE for discovery browsing: Using semantic predications and the literature-based discovery paradigm to elucidate a mechanism for the obesity paradox. In *AMIA 2013, American Medical Informatics Association Annual Symposium, Washington, DC, USA, November 16-20, 2013*, 2013.
- Trevor Cohen, Dominic Widdows, Roger W. Schvaneveldt, Peter Davies, and Thomas C. Rindflesch. Discovering discovery patterns with predication-based semantic indexing. *Journal of Biomedical Informatics*, 45(6):1049 – 1065, 2012.
- D. Hristovski, C. Friedman, T. C. Rindflesch, and B. Peterlin. Literature-based knowledge discovery using natural language processing. 15:133–152, 2008.
- Qian Zhu, Michael Lajiness, Ying Ding, and David Wild. Wendi: A tool for finding non-obvious relationships between compounds and biological properties, genes, diseases and scholarly publications. *Journal of Cheminformatics*, 2(1):6, aug 2010.
- Thomas R. Gruber. The role of common ontology in achieving sharable, reusable knowledge bases. In *Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning (KR'91). Cambridge, MA, USA, April 22-25, 1991.*, pages 601–602, 1991.

- Olivier Bodenreider and Robert Stevens. Bio-ontologies: current trends and future directions. *Briefings in bioinformatics*, 7(3):256–274, sep 2006.
- V. Maojo, J. Crespo, M. Garcia-Remesal, D. de la Iglesia, D. Perez-Rey, and C. Kulikowski. Biomedical ontologies: Toward scientific debate. *Methods of Information in Medicine*, 50(3):203–216, mar 2011.
- Nigam Shah, Clement Jonquet, Annie Chiang, Atul Butte, Rong Chen, and Mark Musen. Ontology-driven indexing of public datasets for translational bioinformatics. *BMC Bioinformatics*, 10(Suppl 2):S1, 2009a.
- Thomas C. Rindfleisch and Marcelo Fiszman. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypnymic propositions in biomedical text. pages 462–477, 2003.
- Andreas Doms and Michael Schroeder. GoPubMed: exploring PubMed with the gene ontology. *Nucleic Acids Research*, 33(Web Server issue):W783–W786, July 2005.
- Ward Blonde, Vladimir Mironov, Aravind Venkatesan, Erick Antezana, Bernard De Baets, and Martin Kuiper. Reasoning with bio-ontologies: using relational closure rules to enable practical querying. *Bioinformatics*, 27(11):1562–1568, jan 2011.
- Franz Baader and Boontawee Suntisrivaraporn. Debugging SNOMED CT using axiom pinpointing in the description logic EL+. In *Proceedings of the Third International Conference on Knowledge Representation in Medicine, Phoenix, Arizona, USA, May 31st - June 2nd, 2008*, 2008.
- Despoina Magka, Markus Krotzsch, and Ian Horrocks. A rule-based ontological framework for the classification of molecules. *Journal of Biomedical Semantics*, 5(1):17, 2014.
- Warren Cheung, BF Ouellette, and Wyeth Wasserman. Inferring novel gene-disease associations using medical subject heading over-representation profiles. *Genome Medicine*, 4(9):75, 2012.
- Warren Cheung, BF Francis Ouellette, and Wyeth Wasserman. Compensating for literature annotation bias when predicting novel drug-disease relationships through medical subject heading over-representation profile (meshop) similarity. *BMC Medical Genomics*, 6(Suppl 2):S3, 2013.

- Lauri Eronen and Hannu Toivonen. Biomine: predicting links between biological entities using network models of heterogeneous databases. *BMC Bioinformatics*, 13:119, June 2012.
- Andreas Schlicker, Thomas Lengauer, and Mario Albrecht. Improving disease gene prioritization using the semantic similarity of gene ontology terms. *Bioinformatics*, 26(18):i561–i567, September 2010.
- Conrad Plake. *Gene Annotation by Automated Literature Analysis with an Application to Drug-Target Interaction Prediction*. PhD thesis, Technischen Universitaet Dresden, Germany, 2010.
- Anika Oellrich, Julius Jacobsen, Irene Papatheodorou, Sanger Mouse Genetics Project, and Damian Smedley. Using association rule mining to determine promising secondary phenotyping hypotheses. *Bioinformatics*, 30(12):i52–i59, June 2014.
- Damian Smedley, Anika Oellrich, Sebastian Köhler, Barbara Ruef, Sanger Mouse Genetics Project, Monte Westerfield, Peter Robinson, Suzanna Lewis, and Christopher Mungall. Phenodigm: analyzing curated annotations to associate animal models with human diseases. *Database : the journal of biological databases and curation*, 2013, January 2013.
- Nicole L. Washington, Melissa A. Haendel, Christopher J. Mungall, Michael Ashburner, Monte Westerfield, and Suzanna E. Lewis. Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS Biol*, 7(11):e1000247, November 2009.
- Robert Hoehndorf, Tanya Hiebert, Nigel W. Hardy, Paul N. Schofield, Georgios V. Gkoutos, and Michel Dumontier. Mouse model phenotypes provide information about human drug targets. *Bioinformatics*, pages btt613+, oct 2013.
- Cynthia Smith, Carroll-Ann Goldsmith, and Janan Eppig. The mammalian phenotype ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biology*, 6(1), 2004.
- Peter N. Robinson, Sebastian Köhler, Sebastian Bauer, Dominik Seelow, Denise Horn, and Stefan Mundlos. The human phenotype ontology: A tool for annotating and analyzing human hereditary disease. *The American Journal of Human Genetics*, 83(5):610 – 615, 2008.

- Catia Pesquita, Daniel Faria, André O. Falcão, Phillip Lord, and Francisco M. Couto. Semantic similarity in biomedical ontologies. *PLoS Computational Biology*, 5(7), July 2009.
- Zhibiao Wu and Martha Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, ACL '94, pages 133–138, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics.
- Claudia Leacock, George A. Miller, and Martin Chodorow. Using corpus statistics and WordNet relations for sense identification. *Comput. Linguist.*, 24(1): 147–165, March 1998.
- Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. *arXiv:cmp-lg/9511007*, November 1995. Proceedings of the 14th International Joint Conference on Artificial Intelligence.
- Jay J. Jiang and David W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv:cmp-lg/9709008*, September 1997. In the Proceedings of ROCLING X, Taiwan, 1997.
- Dekang Lin. An information-theoretic definition of similarity. In *In Proceedings of the 15th International Conference on Machine Learning*, pages 296–304. Morgan Kaufmann, 1998.
- Francisco M. Couto, Mário J. Silva, and Pedro M. Coutinho. Semantic similarity over the gene ontology: Family correlation and selecting disjunctive ancestors. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, CIKM '05, pages 343–344, New York, NY, USA, 2005. ACM. ISBN 1-59593-140-6.
- Catia Pesquita, Daniel Faria, Hugo Bastos, António EN Ferreira, André O Falcão, and Francisco M Couto. Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics*, 9(Suppl 5):S4, April 2008.
- James Z. Wang, Zhidian Du, Rapeeporn Payattakool, Philip S. Yu, and Chin-Fu Chen. A new method to measure the semantic similarity of GO terms. *Bioinformatics*, 23(10):1274–1281, may 2007.
- Shobhit Jain and Gary D. Bader. An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology. *BMC Bioinformatics*, 11(1):562, nov 2010.

- Haixuan Yang, Tamás Nepusz, and Alberto Paccanaro. Improving GO semantic similarity measures by exploring the ontology beneath the terms and modelling uncertainty. *Bioinformatics*, 28(10):1383–1389, may 2012.
- Adrien Barton, Arnaud Rosier, Anita Burgun, and Jean-François Ethier. *Frontiers in Artificial Intelligence and Applications*, volume 267, chapter The Cardiovascular Disease Ontology, pages 409–414. 2014.
- Jie Zheng, Elisabetta Manduchi, and Christian J. Stoeckert Jr. Development of an application ontology for beta cell genomics based on the ontology for biomedical investigations. In *Proceedings of the 4th International Conference on Biomedical Ontology, ICBO 2013, Montreal, Canada, July 7-12, 2013.*, pages 62–67, 2013.
- Erik Segerdell, Virgilio Ponferrada, Christina James-Zorn, Kevin Burns, Joshua Fortriede, Wasila Dahdul, Peter Vize, and Aaron Zorn. Enhanced XAO: the ontology of xenopus anatomy and development underpins more accurate annotation of gene expression and queries on xenbase. *Journal of Biomedical Semantics*, 4(1):31, 2013.
- Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J. Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J. Mungall, Neocles Leontis, Philippe Rocca-Serra, Alan Ruttenberg, Susanna-Assunta Sansone, Richard H. Scheuermann, Nigam Shah, Patricia L. Whetzel, and Suzanna Lewis. The obo foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25(11):1251–1255, November 2007.
- Robert Hoehndorf, Michel Dumontier, and Georgios V. Gkoutos. Evaluation of research in biomedical ontologies. *Briefings in Bioinformatics*, page bbs053, sep 2012.
- Lawrence Hunter and K. Bretonnel Cohen. Biomedical language processing: What’s beyond pubmed? *Molecular Cell*, 21(5):589 – 594, 2006.
- Pierre Zweigenbaum, Dina Demner-Fushman, Hong Yu, and Kevin B. Cohen. Frontiers of biomedical text mining: current progress. *Briefings in Bioinformatics*, 8(5):358–375, 2007.
- Veronika Vincze, Gyorgy Szarvas, Richard Farkas, Gyorgy Mora, and Janos Csirik. The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(Suppl 11):S9, 2008.

- K Bretonnel Cohen, Helen L Johnson, Karin Verspoor, Christophe Roeder, and Lawrence E Hunter. The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC Bioinformatics*, 11:492, September 2010b.
- José A. Gijón-Correas, Miguel A. Andrade-Navarro, and Jean F. Fontaine. Alkemio: association of chemicals with biomedical topics by text and data mining. *Nucleic Acids Research*, 42(W1):W422–W429, 2014.
- Chih-Hsuan Wei, Hung-Yu Kao, and Zhiyong Lu. Pubtator: a web-based text mining tool for assisting biocuration. *Nucleic Acids Research*, 41(W1):W518–W522, 2013.
- Dean Cheng, Craig Knox, Nelson Young, Paul Stothard, Sambasivarao Damaraju, and David S. Wishart. Polysearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic Acids Research*, 36(suppl 2):W399–W405, 2008.
- Aurélié Névéol, W. John Wilbur, and Zhiyong Lu. Improving links between literature and biological data with text mining: a case study with geo, pdb and medline. *Database*, 2012, 2012.
- Yael Garten and Russ Altman. Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text. *BMC Bioinformatics*, 10(Suppl 2):S6, 2009.
- Jorg Hakenberg, Robert Leaman, Nguyen Ha Vo, Siddhartha Jonnalagadda, Ryan Sullivan, Christopher Miller, Luis Tari, Chitta Baral, and Graciela Gonzalez. Efficient extraction of protein-protein interactions from full-text articles. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7(3):481–494, 2010.
- Philippe Thomas, Johannes Starlinger, Alexander Vowinkel, Sebastian Arzt, and Ulf Leser. Geneview: a comprehensive semantic search engine for pubmed. *Nucleic Acids Research*, 40(W1):W585–W591, 2012.
- David Campos, Sérgio Matos, and José Luís Oliveira. Gimli: open source and high-performance biomedical name recognition. *BMC Bioinformatics*, 14(1):54, feb 2013.
- Jean-Fred Fontaine, Florian Priller, Adriano Barbosa-Silva, and Miguel A. Andrade-Navarro. Génie: literature-based gene prioritization at multi genomic scale. *Nucleic Acids Research*, 39(suppl 2):W455–W461, 2011.

- Manabu Torii, Zhangzhi Hu, Cathy H Wu, and Hongfang Liu. Biotagger-gm: A gene/protein name recognition system. *Journal of the American Medical Informatics Association*, 16(2):247–255, 2009.
- Jörg Hakenberg, Conrad Plake, Robert Leaman, Michael Schroeder, and Graciela Gonzalez. Inter-species normalization of gene mentions with GNAT. *Bioinformatics*, 24(16):i126–i132, August 2008a.
- Burr Settles. Abner: an open source tool for automatically tagging genes, proteins and other entity names in text. 21(14):3191–3192, 2005.
- Martin Gerner, Goran Nenadic, and Casey Bergman. Linnaeus: A species name identification system for biomedical literature. *BMC Bioinformatics*, 11(1): 85, 2010.
- John D. Burger, Emily Doughty, Ritu Khare, Chih-Hsuan Wei, Rajashree Mishra, John Aberdeen, David Tresner-Kirsch, Ben Wellner, Maricel G. Kann, Zhiyong Lu, and Lynette Hirschman. Hybrid curation of gene-mutation relations combining automated extraction and crowdsourcing. *Database*, 2014, 2014.
- Rainer Winnenburg, Conrad Plake, and Michael Schroeder. Improved mutation tagging with gene identifiers applied to membrane protein stability prediction. *BMC Bioinformatics*, 10(Suppl 8):S3, 2009.
- J. Gregory Caporaso, William A. Baumgartner, David A. Randolph, K. Bretonnel Cohen, and Lawrence Hunter. Mutationfinder: a high-performance system for extracting point mutation mentions from text. *Bioinformatics*, 23(14):1862–1865, 2007.
- Robert Leaman, Rezarta Islamaj Doäyan, and Zhiyong Lu. Dnorm: disease name normalization with pairwise learning to rank. *Bioinformatics*, 29(22): 2909–2917, 2013.
- Tim Rocktäschel, Michael Weidlich, and Ulf Leser. Chemspot: a hybrid system for chemical named entity recognition. *Bioinformatics*, 28(12):1633–1640, 2012.
- David Jessop, Sam Adams, Egon Willighagen, Lezan Hawizy, and Peter Murray-Rust. Oscar4: a flexible architecture for chemical text-mining. *Journal of Cheminformatics*, 3(1):41, 2011.

- Alan R Aronson and François-Michel Lang. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 2010.
- Nigam Shah, Nipun Bhatia, Clement Jonquet, Daniel Rubin, Annie Chiang, and Mark Musen. Comparison of concept recognizers for building the open biomedical annotator. *BMC Bioinformatics*, 10(Suppl 9):S14, 2009b.
- Minlie Huang, Jingchen Liu, and Xiaoyan Zhu. Genetukit: a software for document-level gene normalization. *Bioinformatics*, 27(7):1032–1033, 2011.
- Joachim Wermter, Katrin Tomanek, and Udo Hahn. High-performance gene name normalization with geno. *Bioinformatics*, 25(6):815–821, 2009.
- Ning Kang, Bharat Singh, Zubair Afzal, Erik M van Mulligen, and Jan A Kors. Using rule-based natural language processing to improve disease normalization in biomedical text. *Journal of the American Medical Informatics Association*, 2012.
- Christopher Funk, William Baumgartner, Benjamin Garcia, Christophe Roeder, Michael Bada, K Cohen, Lawrence Hunter, and Karin Verspoor. Large-scale biomedical concept recognition: an evaluation of current automatic annotators and their parameters. *BMC Bioinformatics*, 15(1):59, 2014.
- R. Jelier, G. Jenster, L. C. J. Dorssers, C. C. van der Eijk, E. M. van Mulligen, B. Mons, and J. A. Kors. Co-occurrence based meta-analysis of scientific texts: retrieving biological relationships between genes. *Bioinformatics*, 21(9):2049–2058, May 2005.
- A. M. Cohen, W. R. Hersh, C. Dubay, and K. Spackman. Using co-occurrence network structure to extract synonymous gene and protein names from MEDLINE abstracts. *BMC Bioinformatics*, 6(1):103, April 2005.
- Emily Doughty, Attila Kertesz-Farkas, Olivier Bodenreider, Gary Thompson, Asa Adadey, Thomas Peterson, and Maricel G. Kann. Toward an automatic method for extracting cancer- and other disease-related point mutations from the biomedical literature. *Bioinformatics*, 27(3):408–415, 2011.
- Hyojung Paik, Hyoung-Sam Heo, Hyo-jeong Ban, and Seong B. Cho. Unraveling human protein interaction networks underlying co-occurrences of diseases and pathological conditions. *Journal of Translational Medicine*, 12(1):99, April 2014.

- Rong Xu, Li Li, and QuanQiu Wang. Towards building a disease-phenotype knowledge base: extracting disease-manifestation relationship from literature. *Bioinformatics*, 29(17):2186–2194, 2013.
- Darby Tien-Hao Chang, Chao-Hsuan Ke, Jung-Hsin Lin, and Jung-Hsien Chiang. Autobind: Automatic extraction of protein-ligand binding affinity data from biological literature. *Bioinformatics*, 2012.
- Tor-Kristian Jenssen, Astrid Lægrend, Jan Komorowski, and Eivind Hovig. A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, 28(1):21–28, May 2001.
- Yael Garten, Nicholas P. Tatonetti, and Russ B. Altman. *Improving the prediction of pharmacogenes using text-derived drug-gene relationships.*, chapter 33, pages 305–314. 2010.
- Adriano Barbosa-Silva, Jean-Fred Fontaine, Elisa R. Donnard, Fernanda Stussi, J. M. Ortega, and Miguel A. Andrade-Navarro. PESCADOR, a web-based tool to assist text-mining of biointeractions extracted from PubMed queries. *BMC Bioinformatics*, 12(1):435, November 2011.
- C. Plake, L. Royer, R. Winnenburg, J. Hakenberg, and M. Schroeder. GoGene: gene annotation in the fast lane. *Nucleic Acids Research*, 37(Web Server):W300–W304, May 2009.
- Lawrence Hunter, Zhiyong Lu, James Firby, William A. Baumgartner, Helen L. Johnson, Philip V. Ogren, and K. Bretonnel Cohen. OpenDMP: An open source, ontology-driven concept analysis engine, with applications to capturing knowledge regarding protein transport, protein interactions and cell-type-specific gene expression. *BMC Bioinformatics*, 9(1):78, January 2008.
- Katrin Fundel, Robert Küffner, and Ralf Zimmer. Relex:relation extraction using dependency parse trees. *Bioinformatics*, 23(3):365–371, 2007.
- Using text to build semantic networks for pharmacogenomics. *Journal of Biomedical Informatics*, 43(6):1009 – 1019, 2010.
- Makoto Miwa, Paul Thompson, and Sophia Ananiadou. Boosting automatic event extraction from the literature using domain adaptation and coreference resolution. *Bioinformatics*, 2012.
- Rong Xu and QuanQiu Wang. Automatic construction of a large-scale and accurate drug-side-effect association knowledge base from biomedical literature. *Journal of Biomedical Informatics*, (0), 2014.

- Razvan Bunescu, Raymond Mooney, Arun Ramani, and Edward Marcotte. Integrating co-occurrence statistics with information extraction for robust retrieval of protein interactions from medline. In *Proceedings of the Workshop on Linking Natural Language Processing and Biology: Towards Deeper Biological Literature Analysis*, BioNLP '06, pages 49–56, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.
- Sarah L. Kinnings, Nina Liu, Nancy Buchmeier, Peter J. Tonge, Lei Xie, and Philip E. Bourne. Drug discovery using chemical systems biology: Repositioning the safe medicine comtan to treat multi-drug and extensively drug resistant tuberculosis. *PLoS Comput Biol*, 5(7):e1000423, July 2009.
- Yvonne Y. Li, Jianghong An, and Steven J. M. Jones. A computational approach to finding novel targets for existing drugs. *PLoS Comput Biol*, 7(9):e1002139, September 2011.
- Twan van Laarhoven and Elena Marchiori. Predicting drug-target interactions for new drug compounds using a weighted nearest neighbor profile. *PLoS ONE*, 8(6):e66952, June 2013.
- Jian-Ping Mei, Chee-Keong Kwoh, Peng Yang, Xiao-Li Li, and Jie Zheng. Drug-target interaction prediction by learning from local information and neighbors. *Bioinformatics*, 29(2):238–245, January 2013.
- Shobeir Fakhraei, Louiqa Raschid, and Lise Getoor. Drug-target interaction prediction for drug repurposing with probabilistic similarity logic. In *Proceedings of the 12th International Workshop on Data Mining in Bioinformatics*, BioKDD '13, pages 10–17. ACM, 2013.
- Liat Perlman, Assaf Gottlieb, Nir Atias, Eytan Ruppin, and Roded Sharan. Combining drug and gene similarity measures for drug-target elucidation. *Journal of Computational Biology*, 18(2):133–145, February 2011.
- Kevin Bleakley and Yoshihiro Yamanishi. Supervised prediction of drug–target interactions using bipartite local models. *Bioinformatics*, 25(18):2397–2403, September 2009.
- Shinhyuk Kim, Daeyong Jin, and Hyunju Lee. Predicting drug-target interactions using drug-drug interactions. *PLoS ONE*, 8(11):e80129, November 2013a.
- Hua Yu, Jianxin Chen, Xue Xu, Yan Li, Huihui Zhao, Yupeng Fang, Xiuxiu Li, Wei Zhou, Wei Wang, and Yonghua Wang. A systematic prediction of multiple

- drug-target interactions from chemical, genomic, and pharmacological data. *PLoS ONE*, 7(5):e37608, May 2012.
- Yoshihiro Yamanishi, Masaaki Kotera, Minoru Kanehisa, and Susumu Goto. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics*, 26(12):i246–i254, June 2010.
- Masataka Takarabe, Masaaki Kotera, Yosuke Nishimura, Susumu Goto, and Yoshihiro Yamanishi. Drug target prediction using adverse event report systems: a pharmacogenomic approach. *Bioinformatics*, 28(18):i611–i618, September 2012.
- N. T. Hansen, S. Brunak, and R. B. Altman. Generating genome-scale candidate gene lists for pharmacogenomics. *Clinical Pharmacology & Therapeutics*, 86(2):183–189, 2009.
- Dorothea Emig, Alexander Ivliev, Olga Pustovalova, Lee Lancashire, Svetlana Bureeva, Yuri Nikolsky, and Marina Bessarabova. Drug target prediction and repositioning using an integrated network-based approach. *PLoS ONE*, 8(4), April 2013.
- Bin Chen, Ying Ding, and David J. Wild. Assessing drug target association using semantic linked data. *PLoS Comput Biol*, 8(7):e1002574, July 2012.
- Yonghui Wu, Mei Liu, W Jim Zheng, Zhongming Zhao, and Hua Xu. Ranking gene-drug relationships in biomedical literature using latent dirichlet allocation. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 422–433, 2012.
- Salvatore Alaimo, Alfredo Pulvirenti, Rosalba Giugno, and Alfredo Ferro. Drug-target interaction prediction through domain-tuned network-based inference. *Bioinformatics*, 29(16):2004–2008, August 2013.
- Hailin Chen and Zuping Zhang. A semi-supervised method for drug-target interaction prediction with consistency in networks. *PLoS ONE*, 8(5):e62975, May 2013.
- Mehmet Gönen. Predicting drug-target interactions from chemical and genomic kernels using bayesian matrix factorization. *Bioinformatics*, 28(18):2304–2310, September 2012.

- Twan van Laarhoven, Sander B. Nabuurs, and Elena Marchiori. Gaussian interaction profile kernels for predicting drug-target interaction. *Bioinformatics*, 27(21):3036–3043, January 2011.
- Feixiong Cheng, Chuang Liu, Jing Jiang, Weiqiang Lu, Weihua Li, Guixia Liu, Weixing Zhou, Jin Huang, and Yun Tang. Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS Comput Biol*, 8(5):e1002503, May 2012a.
- Zhisong He, Jian Zhang, Xiao-He Shi, Le-Le Hu, Xiangyin Kong, Yu-Dong Cai, and Kuo-Chen Chou. Predicting drug-target interaction networks based on functional groups and biological features. *PLoS ONE*, 5(3):e9603, March 2010.
- Yuhao Wang and Jianyang Zeng. Predicting drug-target interactions using restricted boltzmann machines. *Bioinformatics*, 29(13):i126–i134, January 2013.
- Yu-Fei Gao, Lei Chen, Guo-Hua Huang, Tao Zhang, Kai-Yan Feng, Hai-Peng Li, and Yang Jiang. Prediction of drugs target groups based on ChEBI ontology. *BioMed Research International*, 2013, 2013.
- Feixiong Cheng, Yadi Zhou, Jie Li, Weihua Li, Guixia Liu, and Yun Tang. Prediction of chemical-protein interactions: multitarget-QSAR versus computational chemogenomic methods. *Molecular BioSystems*, 8(9):2373–2384, July 2012b.
- Shanfeng Zhu, Yasushi Okuno, Gozoh Tsujimoto, and Hiroshi Mamitsuka. A probabilistic model for mining implicit chemical compound-gene relations from literature. *Bioinformatics*, 21(suppl 2):ii245–ii251, January 2005.
- Tapio Pahikkala, Antti Airola, Sami Pietil, Sushil Shakyawar, Agnieszka Szwarda, Jing Tang, and Tero Aittokallio. Toward more realistic drug-target interaction predictions. *Briefings in Bioinformatics*, page bbu010, April 2014.
- David S. Wishart, Craig Knox, An Chi Guo, Dean Cheng, Savita Shrivastava, Dan Tzur, Bijaya Gautam, and Murtaza Hassanali. DrugBank: a knowledge-base for drugs, drug actions and drug targets. *Nucleic Acids Research*, 36 (Database issue):D901–D906, January 2008.
- The Gene Ontology Consortium. The gene ontology project in 2008. *Nucleic Acids Research*, 36(suppl 1):D440–D444, 2008.
- Jorg Hakenberg, Conrad Plake, Loic Royer, Hendrik Strobelt, Ulf Leser, and Michael Schroeder. Gene mention normalization and interaction extraction

- with context models and sentence motifs. *Genome Biology*, 9(Suppl 2):S14, 2008b.
- Ariel S Schwartz and Marti A Hearst. A simple algorithm for identifying abbreviation definitions in biomedical text. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 451–462, 2003.
- Ralph Delfs, Andreas Doms, Er Kozlenkov, and Michael Schroeder. Gopubmed: ontology-based literature search applied to geneontology and pubmed. In *In Proceedings of German Bioinformatics Conference. LNBI*, pages 169–178. Springer, 2004.
- Christofer D. Manning and Hinrich Schütze. *Foundations of statistical natural language processing*, chapter Collocations. MIT Press, 1999.
- Gerlof Bouma. Normalized (pointwise) mutual information in collocation extraction. pages 31–40, 2009.
- Iraklis Varlamis, Michalis Vazirgiannis, Maria Halkidi, and Benjamin Nguyen. THESUS, a closer view on web content and management enhanced with link semantics. *IEEE Transactions on Knowledge and Data Engineering*, 16(6): 685–700, June 2004.
- Maria Halkidi, Benjamin Nguyen, Iraklis Varlamis, and Michalis Vazirgianis. THESUS: organizing web document collections based on link semantics. *VLDB J.*, 12(4):320–332, 2003.
- Adam Zagorecki and Marek J. Druzdzel. An empirical study of probability elicitation under noisy-or assumption. In *FLAIRS Conference*, pages 880–886, 2004.
- Yoshihiro Yamanishi, Michihiro Araki, Alex Gutteridge, Wataru Honda, and Minoru Kanehisa. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, 24(13):i232–i240, January 2008.
- Allan Peter Davis, Cynthia Grondin Murphy, Robin Johnson, Jean M. Lay, Kelley Lennon-Hopkins, Cynthia Saraceni-Richards, Daniela Sciaky, Benjamin L. King, Michael C. Rosenstein, Thomas C. Wieggers, and Carolyn J. Mattingly. The comparative toxicogenomics database: update 2013. *Nucleic Acids Research*, 2012.

- Andriy I. Bandos, Howard E. Rockette, and David Gur. Use of likelihood ratios for comparisons of binary diagnostic tests: Underlying roc curves. *Medical Physics*, 37(11):5821–5830, 2010.
- M H Zweig and G Campbell. Receiver-operating characteristic (roc) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, 39(4): 561–77, 1993.
- Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.
- Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*, chapter 8. Cambridge University Press, 2008.
- Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 233–240, 2006.
- Hao Ding, Ichigaku Takigawa, Hiroshi Mamitsuka, and Shanfeng Zhu. Similarity-based machine learning methods for predicting drug-target interactions: a brief review. *Briefings in Bioinformatics*, 2013.
- Chong Sun, Narasimhan Rampalli, Frank Yang, and AnHai Doan. Chimera: Large-scale classification using machine learning, rules, and crowdsourcing. *PVLDB*, 7(13):1529–1540, 2014.
- Tie-Yan Liu, Yiming Yang, Hao Wan, Hua-Jun Zeng, Zheng Chen, and Wei-Ying Ma. Support vector machines classification with a very large-scale taxonomy. *SIGKDD Explorations*, 7(1):36–43, 2005.
- Laurent Jacob and Jean-Philippe Vert. Protein-ligand interaction prediction: an improved chemogenomics approach. *Bioinformatics*, 24(19):2149–2156, 2008.
- Qin Iris Wang, Dale Schuurmans, and Dekang Lin. Semi-supervised convex training for dependency parsing. In *ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, USA*, pages 532–540, 2008.
- Jeongkyun Kim, Seongeun So, Hee-Jin Lee, Jong C. Park, Jung-jae Kim, and Hyunju Lee. Digsee: disease gene search engine with evidence sentences (version cancer). *Nucleic Acids Research*, 41(W1):W510–W517, 2013b.

- Aurélie Névéol, W. John Wilbur, and Zhiyong Lu. Extraction of data deposition statements from the literature: a method for automatically tracking research results. *Bioinformatics*, 27(23):3306–3312, 2011.
- Paul Fontelo, Alex Gavino, and Raymond Francis Sarmiento. Comparing data accuracy between structured abstracts and full-text journal articles: implications in their use for informing clinical decisions. *Evidence Based Medicine*, 18(6):207–211, 2013.
- M. Y. Yildiz and W. Pratt. Using statistical and knowledge-based approaches for literature-based discovery. *Journal of Biomedical Informatics*, 39(6):600–611, dec 2006.
- Yun Niu, David Otasek, and Igor Jurisica. Evaluation of linguistic features useful in extraction of interactions from PubMed; application to annotating known, high-throughput and predicted interactions in I2D. *Bioinformatics*, 26(1):111–119, January 2010.
- Antonio J. Pérez, Carolina Perez-Iratxeta, Peer Bork, Guillermo Thode, and Miguel A. Andrade. Gene annotation from scientific literature using mappings between keyword systems. *Bioinformatics*, 20(13):2084–2091, September 2004.
- Marc Aubry, Annabelle Monnier, Celine Chicault, Marie de Tayrac, Marie-Dominique Galibert, Anita Burgun, and Jean Mosser. Combining evidence, biomedical literature and statistical dependence: new insights for functional annotation of gene sets. *BMC Bioinformatics*, 7(1):241, 2006.
- Kazuhiro Seki and Javed Mostafa. Discovering implicit associations between genes and hereditary diseases. In *Biocomputing 2007, Proceedings of the Pacific Symposium, Maui, Hawaii, USA, 3-7 January 2007*, pages 316–327, 2007.
- Minlie Huang, Xiaoyan Zhu, Yu Hao, Donald G. Payan, Kunbin Qu, and Ming Li. Discovering patterns to extract protein-protein interactions from full texts. *Bioinformatics*, 20(18):3604–3612, 2004.
- Patricia L. Whetzel, Natalya F. Noy, Nigam H. Shah, Paul R. Alexander, Csongor Nyulas, Tania Tudorache, and Mark A. Musen. Bioportal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. *Nucleic Acids Research*, 39(suppl 2):W541–W545, 2011.
- Stuart J. Nelson. Medical terminologies that work: The example of mesh. *Parallel Architectures, Algorithms, and Networks, International Symposium on*, 0, 2009.

- Michael F. Chiang, John C. Hwang, Alexander C. Yu, Daniel S. Casper, James J. Cimino, and Justin Starren. Reliability of SNOMED-CT coding by three physicians using two terminology browsers. In *AMIA 2006, American Medical Informatics Association Annual Symposium, Washington, DC, USA, November 11-15, 2006*, 2006.
- Robert Stevens and Uli Sattler. Post-coordination: Making things up as you go along, 2013. URL <http://ontogenesis.knowledgeblog.org/1305>.
- Roman Klinger, Corinna Kolářik, Juliane Fluck, Martin Hofmann-Apitius, and Christoph M. Friedrich. Detection of iupac and iupac-like chemical names. *Bioinformatics*, 24(13):i268–i276, 2008.
- Simone Daminelli, V. Joachim Haupt, Matthias Reimann, and Michael Schroeder. Drug repositioning through incomplete bi-cliques in an integrated drug-target-disease network. *Integr. Biol.*, 4:778–788, 2012.
- V. Joachim Haupt, Simone Daminelli, and Michael Schroeder. Drug promiscuity in pdb: Protein binding site similarity is key. *PLoS ONE*, 8(6):e65894, 06 2013.