Methods xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Prediction of drug gene associations via ontological profile similarity with application to drug repositioning

Maria Kissa, George Tsatsaronis, Michael Schroeder *

Biotechnology Center, Technische Universität Dresden, Germany

ARTICLE INFO

Article history:
Received 26 February 2014
Accepted 25 November 2014
Available online xxxx

Keywords:
Drug gene association prediction
Drug repositioning
Ontological profiles
MEDLINE
Unsupervised
Semantic relatedness

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 a fully corpus-based and unsupervised method which utilizes the *MEDLINE* indexed titles and abstracts to infer drug gene associations and assist drug repositioning. The method measures the *Pointwise Mutual Information* (*PMI*) between biomedical terms derived from the *Gene Ontology* and the *Medical Subject Headings*. 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 method can successfully identify direct drug gene associations with high precision and prioritize them. Validation shows that the statistically derived profiles from literature perform as good as manually curated profiles. In addition, we examine the potential application of our approach towards drug repositioning. For all *FDA* approved drugs repositioned over the last 5 years, we generate profiles from publications before 2009 and show that new indications rank high in the profiles. In summary, literature mined profiles can accurately predict drug gene associations and provide insights onto potential repositioning cases.

1. Introduction

Drug repositioning is the task of finding new targets for old drugs and has been in the spotlight for the past few years. The average cost for launching a new drug into the market is estimated to 1.8 billion dollars [1]. Apart from that, the drugs that make it to the market are very few. Notably, from 1999 to 2008, only 50 compounds were *FDA* approved in the U.S., out of which 17 were identified as arising from target-based discovery methods [2]. This stresses the importance of drug repositioning in the process of drug development, since it accelerates the process, minimizes the associated costs, and, in parallel, contributes to the prevention of noxious adverse events and toxicological liabilities.

The use of terminologies for the prediction of drug–target interactions has been exploited extensively in the past. Campillos et al. showed that drugs causing the same adverse events may share the same off-targets [3]. Through this study new targets for already marketed drugs were experimentally confirmed. Lamb et al. built a Connectivity Map and used similar gene expression signatures to connect drugs, genes and diseases [4]. Lounkine et al. enriched a drug-adverse event-target network of 73 new off-targets and 656 marketed drugs with chemical features and sequential information [5]. This method led to the experimental confirmation of 125 novel drug–target interactions. Other approaches also used similarity based on pharmacological effects for generating *in silico* predictions of drug–target interactions [6].

Generally, there are few unsupervised methods towards the prediction of drug gene associations. Chen et al. build a network of so-called semantically linked entities to drugs based on publicly available repositories which comprise drug-related information [7]. By traversing the paths in that network they identify successfully drug gene associations. Wu et al. 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 [8]. Mestres et al. investigated the topological characteristics of drug–target networks in general [9]; the authors constructed 4 network of more than 10.000 interactions assembling data from various databases and *in silico* predictions and analyzed how much network topology depends on the data sources, drug properties and target families. The observations made over the 4 networks converge to the fact that small hydrophobic drugs has a very high

* Corresponding author.
E-mail addresses: maria.kissa@biotec.tu-dresden.de (M. Kissa), george.tsatsaronis@biotec.tu-dresden.de (G. Tsatsaronis), ms@biotec.tu-dresden.de (M. Schroeder).

http://dx.doi.org/10.1016/j.ymeth.2014.11.017
1046-2023/

promiscuity. On the other hand, the supervised techniques for the prediction of drug gene associations are numerous [10–13]. Bleakley et al. represent drug–target interactions as bipartite graphs and predict targets for drugs via learned local models [14]. Cheng et al. infer drug gene associations via a supervised network based approach and show that it performs better compared to drug-based and target-based drug gene association prediction [15]. Emig et al. also use the notion of an integrated network and learn global and local features towards target identification and drug-repurposing [16]. Vogt and Mestres give an overview on several drug–target network applications and point to information completeness as the main obstacle towards the efficient identification of drug–target interactions [17].

In this study, we focus on the prediction of drug gene associations. The suggested methodology utilizes the bibliography to measure corpus-based semantic relatedness between ontological terms. To the best of our knowledge, it is the first unsupervised method that predicts new drug gene associations solely by analysing systematically the co-occurrence of biomedical terms in all the scientific publications indexed by *MEDLINE*. Intermediate ontological concepts are used to form the links between drug and genes. In the past, the problem of establishing indirect links between two concepts A and C via a set of intermediate concepts B has been addressed by Srinivasan [18]. Herein, we also perform hypothesis generation from biomedical texts, and suggest putative drug gene associations on a large scale. The presented method identifies the co-occurrences of *GO* and *Medical Subject Headings* (*MeSH*) *Disease* concepts with drugs and genes in *MEDLINE* titles and abstracts. The co-occurrence information is used to rank the most related *GO* and *MeSH Disease* biomedical concepts to the drug and the gene respectively. These concepts form an individual profile for each drug and gene, which is in turn, used to assess associations between them by quantifying the degree of the relatedness between their profiles. In addition, the generated profiles can provide an insight into biomedical properties for drugs and genes and infer associations between them that might not have been included in a database nor reported in the literature. To this end, we experimentally evaluate the approach in prioritizing drug gene associations. The results show that the co-occurence based profiles perform as good as the manually curated profiles. We also validate that the proposed measure for the estimation of the semantic relatedness between the profiles outperforms traditional measures of semantic similarity. via empirical evaluation that is presented in Section 3 (*ROC* Curves), we demonstrate that the proposed method successfully recovers true interactions when applied on a dataset for which the respective drug and gene names are not co-occurring in any scientific publication. Finally, we apply our approach towards the identification of candidate drug repositioning cases. We collect all the known drug repositioning cases which were reported by the *FDA* within the last 5 years (since 2009). For all the drugs included in these cases, we generate the *MeSH Disease* profiles based on the literature data before the year of repositioning. We demonstrate that the new therapeutic indication, i.e., *Disease*, is always included and ranks high in the profiles. This suggests that the proposed method offers a meaningful insight for the task of drug repositioning. All the steps of our method are summarized in Fig. 1, and are explained in detail in Section 2.

2. Materials and methods

Our approach towards the prediction of drug gene associations is to identify latent relations between drugs and genes by creating their profiles and measuring their relatedness. The drugs are taken from *DrugBank* and the genes from *UniProtKB*. A profile consists of *GO* and/or *MeSH Disease* concepts that co-occur with the gene or the drug in the literature, i.e., in the titles and abstracts of *MEDLINE* indexed articles. *MEDLINE* abstracts and titles constitute a vast high-throughput annotation source and has been explored in the past for the identification of relationships between biomedical entities, in conjunction with the usage of co-occurrence data [19]. Only 10% of the *MEDLINE* indexed articles are *Open Access* and thus, their full text is available. On the other hand, the abstracts and titles for all *MEDLINE* indexed articles are freely accessible. It has been demonstrated that text mining tools perform better on abstracts than on article bodies [20]. Due to their condensed information and clear statements of the research findings, the co-occurrence of biological entities in scientific abstracts has been shown to reflect meaningful relationships between them [21]. As far as the *GO* terms are concerned, all concepts under biological processes (*GO:*0008371), molecular functions (*GO:*0008369) and cellular components (*GO:*0008370) may participate in a profile, while from *MeSH* only the concepts under the *Disease* tree are considered.

The degree of the co-occurrences between a drug or a gene and the candidate concepts is quantified by measuring *Pointwise Mutual Information* (*PMI*) scores. *PMI* scores can then be used to rank the related concepts of a drug or a gene that participate in their profiles. Considering *MEDLINE* article titles and abstracts is especially important for generating the drug profiles, since, to the best of our knowledge, there do not exist drug databases with comprehensive *MeSH* and *GO* annotations. In turn, the relatedness between a drug and a gene profile is measured using *PMI* scores between the concepts of their profiles, again based on their co-occurrence in *MEDLINE* titles and abstracts. Finally, the *PMI* scores between the drug and the gene profile concepts are combined and an overall relatedness score between a drug and a gene is calculated. The overall scores between drug and gene profiles can then be used to prioritize the drug gene associations. All the steps of our method are summarized in Fig. 1, and are explained in detail in the following subsections.

2.1. Recognition of terms in text

In the first step, protein coding genes and drugs are recognized in *MEDLINE* indexed abstracts and titles. Approximately 22 million indexed articles were considered.

Regarding the gene annotation process, all genes from *UniProtKB* were considered. For their recognition in text, we applied the gene annotation system *GNAT* [22]. GNAT is a publicly available system which handles inter-species gene mention normalization. Unlike to 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 into 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. Altogether, around 58,000 genes from 31 species were identified in *MEDLINE* indexed abstracts and titles (see Table S3).

Regarding the drug annotation process, a dictionary approach was followed. The reference of a drug in literature varies and poses a significant challenge towards the correct identification of drug names in biomedical text; drugs can be referred to by their

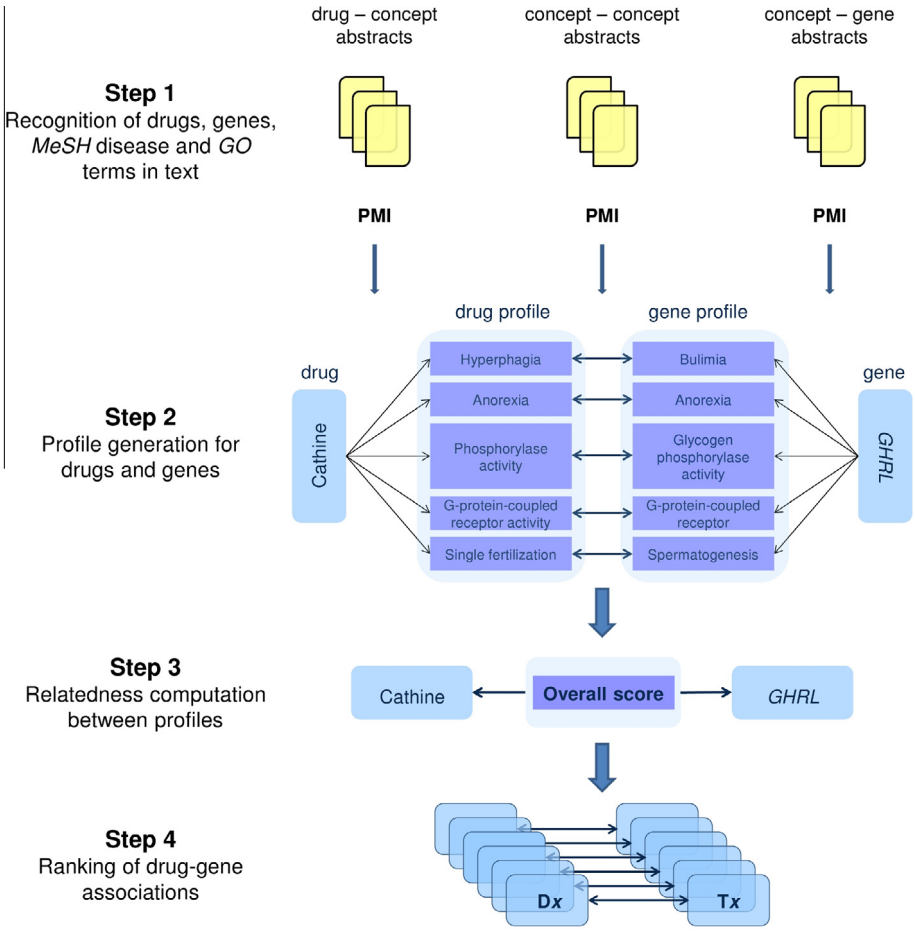

Fig. 1. Overview of our approach towards drug gene association prediction. The profile of the psychotropic drug *Cathine* is shown, along with the profile of the gene *GHRL*. *GHRL*, which encodes the hunger-stimulating peptide *Ghrelin*, is predicted by our method as a gene associated to *Cathine*. The figure illustrates the major steps of the suggested methodology. 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.

commercially assigned name (e.g., *Amiodarone*), their *IUPAC* name (e.g., (2-4-[(2-butyl-1-benzofuran-3-yl) carbonyl]-2,6-diiodophenoxyethyl) diethylamine) or even by their molecular formula (e.g., $C_{25}H_{29}I_2NO_3$). Lately, there have been approaches towards the unification of different drug related repositories to one single dictionary of drug names presenting promising performance results, e.g., the *Joint Chemical Dictionary* (*JoChem*) [23]. However, for the task at hand, an *in-house* drug dictionary based on the *DrugBank* drug names and synonyms was utilized [24]. 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* name. All regular expressions are compiled to a single labeled deterministic finite state automaton (*LDFA*). Each end state in the automaton stores the corresponding 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. Drug names fall into two categories of name ambiguity. The first pertains to the false positives that result from ambiguous abbreviations (e.g., *ACC* for *Acetylcysteine or Adenoid Cystic Carcinoma*). In such case, the abbreviations are mapped to their long forms with the use of the algorithm introduced by [25]. For a random set of 60 *PubMed* references out of the set of 22 million references utilized by the suggested methodology, we manually annotated the corresponding titles and abstracts. 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* [26], a knowledge-based search engine that organizes *PubMed* references with *MeSH* and *GO* annotations (see Table S3). With respect to *MeSH*, only *Disease* terms were considered. *MeSH* is general; it consists of more than 27,000 terms (descriptors) organized in categories that, apart from *Diseases*, include as well *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. *GO* is also large, but entirely focused on the biological processes and functions and hence, it has been fully used. The *MeSH* headers that are used for indexing *MEDLINE* abstracts and titles were also ignored. The goal of this study is to introduce a robust methodology that automatically infers drug gene associations from text and remains independent from any sort of information introduced by manual annotation, as it is in the case of *MeSH* headers. Focusing on the

terms that are literally reported in text also opts for the general applicability of the suggested methodology on other types of text, such as patent documents for example.

2.2. Profile generation for genes and drugs

For the creation of the drug and gene profiles the *MeSH Disease* and *GO* terms that co-occur in the literature with each drug and gene are considered. There is a significantly large number of documents in which drugs or genes and ontological terms co-occur and this motivates the usage of statistical measures to quantify their relatedness, such as *PMI* (see Table S4).

Considering every co-occurring term as part of a drug's or a gene's profile could result in many associations. The aim of this step is to quantify the strength of the associations. For this purpose we compute the strength of each association based on the *Pointwise Mutual Information* (*PMI*) between the drug or gene and the respective ontological term. *PMI* is a probabilistic measure used to assess the strength of word collocations in a text corpus [27]. Herein, we extent its application to co-occurring pairs of drugs/genes and ontological terms in titles and abstracts of *PubMed* documents. It has been shown that a high percent of biomedical entities co-occurring in abstracts, also co-occur in sentences [28]. For that reason, we focus on abstract level co-occurrence and aim in maximizing the recall of latent associations between drugs/genes and ontological terms. Given that the negative associations are underepresented in literature [29] and therefore would result to a low *PMI* score, we decided to apply no filtering.

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 Eq. (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.

$$pmi(E,C) = \log\frac{N \times n_{E,C}}{n_E \times n_C} \tag{1}$$

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 *normalized PMI* [30] (*nPMI*) that takes values between $[-1,+1]$. Eq. (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 value of 0 shows independence between E and C, and a value of 1 shows complete co-occurrence between E and C.

$$npmi(E,C) = \frac{pmi(E,C)}{-\log\frac{n_{E,C}}{N}} \tag{2}$$

Let A represent the set of ontological terms C that co-occur with an entity E. We retain for the profile of E only the terms C for which: $npmi(E,C) \geqslant mean_A(npmi(E,C))$ and $npmi(E,C)$ has a *p-value*< 0.05 among all the $npmi(E,C)$ scores between any entity E and any concept C.

2.3. Relatedness computation between profiles and ranking of drug gene associations

Following the generation of the profiles for drugs and genes, in this step the relatedness between all drug gene pairs is computed based on their profiles. The number of documents wherein co-occurrences between ontological terms appear is significant and this enables the utilization of *PMI* to quantify their relatedness (see Table S5).

More specifically, the computation of the relatedness between a drug and a gene is based on the *nPMI* score values between all possible pairs of the ontological terms comprising the entities' 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 based on Eq. (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 Eq. (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 as described in the work by [31] to produce the overall score between the drug and the gene. 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 Eq. (3) as $S_1(d,g)$. Similarly, $S_2(g,d)$ is computed for all $C_g \in P_g$, the way it is shown in Eq. (4). Finally, the two scores $S_1(d,g)$ and $S_2(g,d)$ are combined as shown in Eq. (5) to produce the overall score between a drug d and a gene g. Eventually, all the calculated scores between every drug and every gene are used to order the drug gene pairs in the collection, with the higher values suggesting drug gene pairs that are more likely to comprise meaningful predicted drug gene associations.

$$S_1(d,g) = \frac{1}{|P_d|}\sum_{C_d \in P_d} \max_{C_g \in P_g} npmi(C_d, C_g) \tag{3}$$

$$S_2(g,d) = \frac{1}{|P_g|}\sum_{C_g \in P_g} \max_{C_d \in P_d} npmi(C_g, C_d) \tag{4}$$

$$Score(d,g) = \max(S_1(d,g), S_2(g,d)) \tag{5}$$

2.4. Evaluation datasets

For the evaluation of the proposed method towards predicting drug gene associations we used altogether three main sets of drug gene interaction data. For each of the sets there exist both true and false examples of drug gene interactions. The first set comprises drug gene pairs derived from *DrugBank*. The positive pairs (true interactions) are the drug gene interactions listed by *DrugBank*. The negative examples (false interactions) are generated from the rest of the combinations, i.e., all drug gene combinations which are not listed in *DrugBank*. Since *DrugBank* is not complete, this set may contain true interactions, but so few that we can neglect for the purpose of evaluation. The second set is the one introduced in the work of [32], which contains both positive and negative examples. The third set comprises drug gene pairs derived from the *Comparative Toxicogenomics Database* (*CTD*). The positive pairs are again the drug gene relations listed in *CTD* and the negative pairs are generated from the rest of the combinations, as in the case of *DrugBank*. Given the fact that not all drugs/genes have profiles generated, we exclude from the datasets any drug gene interactions wherein either the drug or the gene has an empty profile. Pairs of empty profiles result in zero scored drug gene associations that when included in the evaluation falsely enhance the discriminative power of our method. By excluding these pairs we avoid any bias towards the performance of our method. Table 1 gives an overview of the three resulting datasets and reports the number of drugs, genes and interactions (true and false) per dataset.

Towards demonstrating the application of the method in identifying candidate drug repositioning cases, we needed a set of all known drugs that have been repositioned, along with their old

Table 1
Evaluation dataset statistics. The table reports the number of drugs, genes and associations included in the evaluation datasets. In the case of *DrugBank*, we discriminate true associations based on the type of the *drug*, i.e., *Approved* or *Experimental*. In the case of *CTD*, we discriminate true associations based on the type of association, i.e., *Binding* (i.e., physically interacting) or *Related*. For the validation of the method when co-occurrence based profiles are used, we use *DrugBank* and *Yamanishi*. For the validation of the method when manually curated profiles are used, we use *CTD*.

	DrugBank		Yamanishi et al.	CTD	
	Approved	Experimental		Binding	Related
Drugs	751	790	302	955	4854
Genes	659	1732	657	633	5983
True pairs	1836	2843	2493	2260	203,949
False pairs	7,660,034		486,904	29,195,261	

and new indications. To our knowledge there is no such set publicly available. For that reason we manually mined the literature and compiled the set based on *U.S. FDA*, *Wikipedia* and other web resources.[1]

Only drugs which have a *DrugBank* identifier were considered. Old and new indications are reported along with the year of approval of each drug's new indication. For our analysis, we focus on drugs that were *U.S. FDA* approved within the last 5 years with their new indications and for which we have profiles. Thus, for the application of the suggested method in this task we excluded all *MEDLINE* data from 2009 and on. The application of the method in identifying candidate drug repositioning cases is conducted as follows: the drug's profile is generated, and we examine the *MeSH Disease* terms that participate in the profile. 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. Results and discussion

3.1. Prediction of drug gene associations

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*, we associate drugs to genes by assessing the semantic relatedness of their profiles. For details of the method's definition, see Section 2. Using the datasets discussed in Section 2.4, the results of our evaluation, meaning the *Receiver Operating Characteristic* (*ROC*) curves for the datasets are shown in Figs. 2 and 3. The notion of the performance evaluation we follow for the suggested method using *ROC* curves and the respective *Area Under the Curve* (*AUC*) measurement is similar to the notion of the evaluation used in the field of Information Retrieval by utilizing *ROC* curves [33]. More precisely, the usage of *ROC* curves illustrates the ability of the method to prioritize the positive examples, having as an input only a ranked list of results. In our case the input is the ordered list of the drug gene associations.

In order to conduct our evaluation we make use of three main drug target interaction datasets (see Table 1). The first dataset comprises drug target interactions from *DrugBank* database [34]. *DrugBank* is a publicly available resource which comprises comprehensive drug target information regarding both *FDA* approved and experimental drugs. We also considered the dataset provided by Yamanishi et al. [32]. This dataset has been used in the past for the performance evaluation of machine learning methods towards the prediction of drug target interactions. The third dataset comprises drug target interactions and drug gene relationships from the *Comparative Toxicogenomics Database* (*CTD*) [35]. *CTD* is also publicly available and comprises curated data that describes relationships between drugs, genes and diseases.

We consider two aspects of evaluation. The first is to assess whether the suggested methodology is able to successfully prioritize true drug gene associations, solely based on their literature derived profiles. The second is to assess whether the suggested methodology is able to prioritize true drug gene associations in the case of manually curated profiles. For the first aspect, we consider *DrugBank* and the dataset from Yamanishi et al. For the second aspect, we utilize the manually curated *MeSH Disease* profiles for drugs and genes provided by *CTD* and based on their statistical semantic relatedness we rank the respective drug gene associations.

With regards to the first aspect of evaluation, given that we want to rank drug gene associations solely based on their literature derived profiles, we removed datasets all the interactions between any drug and any gene that occur together at least once in a *PubMed* citation. This way the suggested method computes relatedness between drugs and genes solely based on the co-occurrence of their profile's terms. The method remains unbiased by the inclusion of drug gene association pairs that can be easily picked up from the literature. As Fig. 2 shows, the suggested method obtains an *AUC* of 0.84 for the *DrugBank* drug gene associations in which *approved* drugs participate and 0.58 when *experimental* drugs participate. An *AUC* of 0.74 is reported for the *Yamanishi* dataset. 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, we examined the number of literature references for both types of drugs. 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 underepresentation 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 S1). Consequently, latent associations between experimental drugs and genes are not as strongly established as in the case of the approved drugs, which explains the difference between the two *AUC*s. Accordingly, if we consider drug gene associations wherein *human* genes participate, the *AUC* values increase for both approved and experimental drugs to 0.88 and 0.77 respectively, as shown in Fig. 3. This is because, *human* genes are discussed more in literature than the genes that belong to other species. Altogether, we identified genes from 31 species 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 (61). This explains the improvement in the performance of our method when applied on the respective subset.

Moreover, considering the reported results for the *Yamanishi* dataset, and taking into account the small overlap with the *DrugBank* dataset (the *Yamanishi* true drug gene associations constitute only 8% of the true drug gene associations included in *DrugBank*), the value of the *AUC* (0.74) suggests that the method is robust.

In addition, we examined the impact in performance of different ontology terms that participate in drug and gene profiles. We

[1] <http://www.cancer.gov/><http://www.centerwatch.com/><http://www.drugs.com><http://www.medicalnewstoday.com/><http://www.medscape.com/><http://www.webmd.com/>.

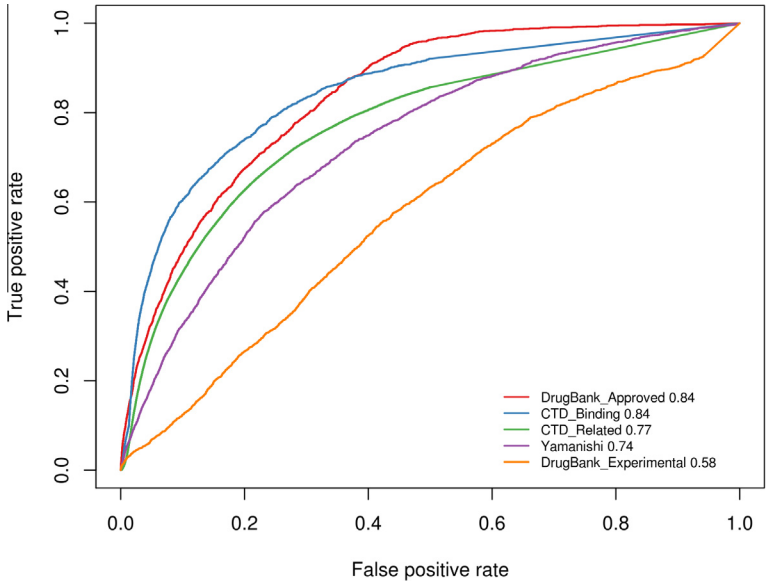

Fig. 2. *ROC* curves for all evaluation datasets. *ROC* curves plotting the true positive rate against the false positive rate for the used datasets. The input for the plotting of the *ROC* curves is the ranked list of drug gene associations produced by the suggested method. The *AUC* values quantify the ability of the suggested method to rank the positive associations higher. The curves show that the prediction method works well if there is sufficient underlying data (*Approved* and *Yamanishi*). For experimental drugs little is published and hence the method performs worse. The suggested method produces comparable results for the datasets *Binding* and *Related*, wherein manually curated profiles were utilized. The prediction method works better in the case of direct (i.e., physically interacting) drug gene associations, as shown by the *AUC* values that were achieved on the *Binding* and *Related* dataset.

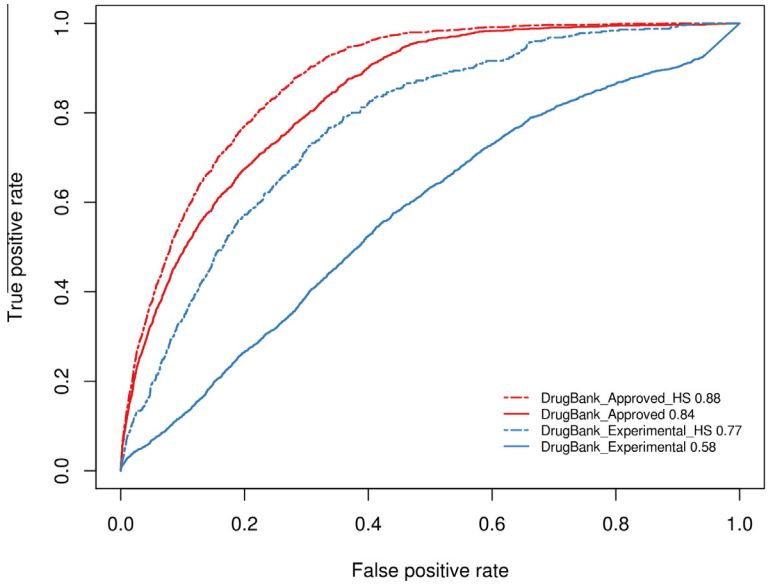

Fig. 3. *ROC* curves for *Human* and non-*Human* genes. The curves show that the suggested prediction method performs better for associations between drug and *Human* genes. For *Human* genes there is a lot published and that has a beneficial impact on the method's performance.

repeated the evaluation process considering only one type of ontology terms in each experiment; either *MeSH Disease* or *GO* terms. We observed that *GO* terms have a significantly larger impact in the prediction performance compared to the *MeSH Disease* terms. More precisely, considering only *MeSH Disease* terms, the *AUC*s are 0.77, 0.39 and 0.66 respectively for the *DrugBank* approved, *Drugbank* experimental, and the *Yamanishi* dataset. When only *GO* terms are considered, the respective *AUC*s become 0.84, 0.68 and 0.78. The contribution of their combination is exactly what is reported in Fig. 2. Hence, the contribution of *GO* terms in the drug–target prediction is significantly higher than the contribution of the *MeSH Disease* terms.

With regards to the second aspect of evaluation, we want to assess whether the semantic relatedness of manually curated profiles is able to successfully prioritize known drug gene associations. For that reason, we utilize *CTD*. *CTD* comprises both curated and inferred drug gene, gene-disease and drug-disease associations. We focus only on the curated data, and with the help of the disease profiles provided by *CTD* for both drugs and genes we calculate their statistical semantic similarity and rank the drug gene associations. We divide the set of true drug gene associations based on the type of the association, i.e., the *binding* associations and the *related* associations. The *binding* drug gene associations correspond to drug gene pairs wherein the drug is known to bind the protein

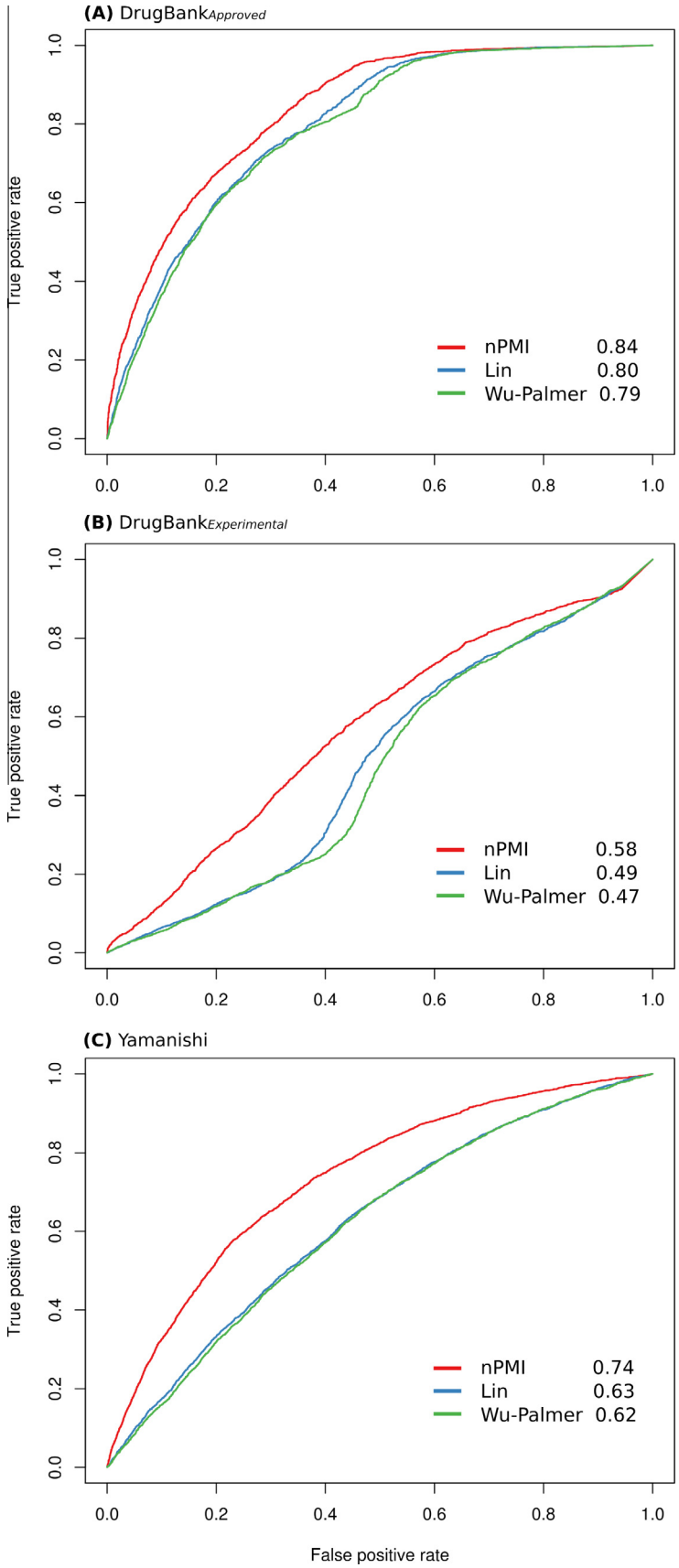

Fig. 4. *ROC* curves for different measures of semantic similarity. *ROC* curves plotting the true positive rate against the false positive rate of each semantic similarity measure on the datasets (A) *Approved*, (B) *Experimental*, and (C) *Yamanishi*. The curves show that the suggested prediction method outperforms both other measures of semantic similarity, i.e., *Wu–Palmer* and *Lin*, towards the prioritization of drug gene associations. *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.

Please cite this article in press as: M. Kissa et al., Methods (2014), http://dx.doi.org/10.1016/j.ymeth.2014.11.017

Table 2
Examples of *nPMI* computations. 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.

	Pair A		Pair B	
	Coronary disease	Myocardial ischemia	Kearns–Sayre syndrome	Ophthalmoparesis
$n_{C_{d/g}}$	164,596	55,757	514	846
n_{C_d,C_g}	18,136		134	
Distance	1		2	
nPMI	0.46		0.71	
Wu–Palmer	0.89		0.83	

that the respective gene codes, while the *related* associations correspond to drug gene pairs wherein the drug is known to have an impact on the gene's products. As shown in Fig. 2, the suggested method manages an *AUC* 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, the method achieves an *AUC* of 0.77. This value increases to 0.79 when taking into account only drug–*human* gene pairs. Clearly, the proposed method demonstrates a better performance on the *binding* dataset than the *related* one, thus justifying the method's potential to successfully prioritize the drug–target interactions over drug gene associations.

The above results clearly demonstrate that the use of co-occurrence based ontological profiles leads to a performance comparable to the one achieved when using manually curated profiles. For all the datasets used in our evaluation, the respective *Precision–Recall* Curves are also provided in the Supplementary Material and Fig. S1. As shown, the efficacy of the proposed methodology in the *Precision–Recall* space is concordant with its performance in the *ROC* space, both when co-occurrence based and manually curated profiles are utilized.

3.2. Evaluation against traditional measures of semantic similarity

To calculate the semantic relatedness of the profiles, we utilize the statistical semantic similarity *nPMI*. To evaluate the efficacy of the suggested measure towards the prioritization of drug gene associations, we compared *nPMI* against two traditional metrics of semantic similarity; the *Wu–Palmer* [36] and *Lin* [37] metrics. Unlike *nPMI*, these metrics compute the semantic similarity between two terms by taking into account the ancestor information. According to *Wu–Palmer* the shorter the path that connects two terms in the ontology via their *Lowest Common Ancestor* (*LCA*), the higher the similarity between two terms. According to *Lin*, the semantic similarity between two terms is computed by considering their *Information Content* (*IC*) and the *Information Content* of their *LCA*, as well. *Lin* takes into account both corpus but also ancestor information.

For the respective evaluation we used the *DrugBank Approved* and *Experimental*, and the *Yamanishi* datasets. The results are analytically represented in Fig. 4. The statistical semantic similarity measure *nPMI* obtains the highest *AUC* value in all three datasets. The curves in the *Precision–Recall* space are concordant; *nPMI* achieves higher *Precision* for all *Recall* values compared to metrics *Wu–Palmer* and *Lin* (see Fig. S2 in Supplementary Material). The *Lin* metric that combines both ontology and corpus information achieves the second best performance. A statistical *t-test* was performed to assess the discriminative power of each measure of semantic relatedness. It is shown (Table S2) that the *nPMI* measure can discriminate with statistical significance true from false drug gene associations. The proposed measure has the highest values of distance between true and false *ECDF* distributions of drug gene associations, according to the *Kolmogorov–Smirnoff* test (see Table S2).

The arising question is why *nPMI* performs better from the traditional measures of semantic similarity. Apart from being consistent with the profile generation for drugs and genes, the statistical metric *nPMI* has two basic characteristics that differentiate it from the remaining measures of semantic similarity and account for its good performance.

First, *nPMI* is a probabilistic and fully-corpus based measure 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 (see Eq. (1) and (2)). But why does this characteristic of *nPMI* result in a positive impact of the methods performance? Why taking into account the ontology (as in the case of *Wu–Palmer* and *Lin*) proves insufficient?

The answer is as follows. If profiles are connected via the ontology this results in precise relations, however the absolute number of possible relations decreases because ontology based connections constitute only a part of the relations that *nPMI* suggests overall.

Table 3
Examples of drug repositioning potential.

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

Please cite this article in press as: M. Kissa et al., Methods (2014), http://dx.doi.org/10.1016/j.ymeth.2014.11.017

Indeed, to quantify the degree of *nPMI*'s coverage, we check how many of the nPMI suggested profile links can be connected via the ontology (meaning they have an *LCA*). For the set of drug gene associations between *Approved* drugs and genes in DrugBank 73% of the *MeSH Disease* profile links and 67% of the *GO* profile links can be established via the ontology. This means that *nPMI* provides 27% and 33% more relations in each case. When considering drug gene associations wherein *Experimental* drugs participate, *nPMI* provides 38% more *MeSH Disease* and 37% more *GO* connections. Similarly, in the *Yamanishi* dataset an additional percentage of relations of 27.4% and 31.2% respectively.

The statistical measure *nPMI* is able to uncover additional connections between the profiles. For example, *nPMI* is able to assign a similarity score between two *GO* terms which belong to different subontologies, e.g., a term from the *Biological Process* and a term of the *Molecular Function* subontology. These terms cannot be connected through the ontology; the similarity between a concept from the *Biological Process* subontology and a concept from the *Molecular Function* subontology always receives a zero score by the *Lin* and the *Wu–Palmer* metrics. The same happens in the case of two *MeSH Disease* terms, wherein the latter disease is a symptom of the former. Let us, for example, consider *Prader–Willi Syndrome*, a congenital disease affecting many parts of the body. According to the *MeSH* definition, symptoms of the disease include the *Hypogonadism* condition. By a quick look-up in the hierarchy of *MeSH*, we can see that *Prader–Willi Syndrome* and *Hypogonadism* cannot be connected through the ontology since they have no *Lowest Common Ancestor* (*LCA*) (their *LCA* is the root). Consequently, both the *Lin* and the *Wu–Palmer* metrics would assign a similarity of 0.0 to these concepts, while *nPMI* assigns a similarity score of 0.42.

The second characteristic of *nPMI* is the ability to discriminate the 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 [27]. More precisely, let us consider two 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 rare, then the rare pair is more informative. 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 2 shows an example of this phenomenon.

Assume $C_d = Coronary\ Disease$ and $C_g = Myocardial\ Ischemia$ two terms that participate in the profile of a drug d and a gene g respectively. Table 2 reports the number of *PubMed* documents where C_d occurs (n_{C_d}), the number of *PubMed* documents where C_g occurs (n_{C_g}), and the number of *PubMed* 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 $C_d = Kearns—SayreSyndrome$ and $C_g = Ophthalmoparesis$, 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 number 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.

3.3. Drug repositioning

We manually mine the literature and compile a set of drugs repositioning cases (see Section *Evaluation Datasets*). Table 3 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 of the 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 drugs' profiles. In almost all of the cases they appear among the top 30 associated disease terms of the respective profiles. For the new indications we calculated 2 types of z-scores. The first is the z-score the new indication achieves inside the profile of the drug. The second z-score corresponds to the overall distribution of drug-disease associations. Roughly, we see that 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 that the proposed methodology can be utilized towards identifying new indications for already existing drugs.

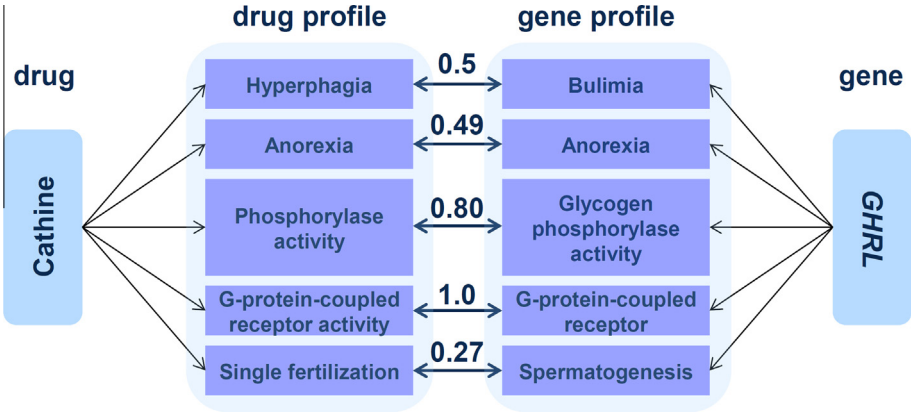

Fig. 5. *Cathine – GHRL* association representation. The figure illustrates the intermediate connections of *Cathine – GHRL*. Clearly, the most surprising connection is established via the concepts *Single Fertilization* and *Spermatogenesis* that pertain to *Reproduction*. These concepts are semantically quite distinct from the *Eating Disorders* concepts to which *Cathine* and *GHRL* have known relations.

3.4. Case studies

In the following we present a case study demonstrating the drug gene association prediction and a case study illustrating the application towards drug repositioning. In both cases, we discuss the scientific literature findings that support them, and we show the inferred associations through graphical representations.

3.4.1. Cathine-GHRL association

In Fig. 5, all the hypotheses suggesting an association between *Cathine* (*DrugBank:DB*01486), a psychotropic compound, and *GHRL* (*Entrez Gene:*51738), a gene coding for *Ghrelin*, which is a growth hormone-releasing peptide, are shown. *Cathine* was 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* was ranked among the top 0.7% of these genes (position 86) with a *z*-score of 2.83 and a *p-value*<0.05.

Table 4 shows representative textual pieces of evidence suggesting this association which was discovered by the suggested methodology. *Cathine* and *GHRL* are interconnected via the concepts *hyperphagia* (*MeSH:D*006963) and *bulimia* (*MeSH:D*002032). According to *MeSH*, *bulimia* is a form of *hyperphagia*. The association emerges from the following: *Cathine* (*d-norpseudoephedrine*) acts as a stimulant and it can be isolated from the plants *Catha Edulis* and *Ephedra Sinica*. It is a phenylpropanolamine (*PPA*) isomer, along with *norephedrine*. There exist 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:D*052018) (*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 (*PMID:*21354264). In the same paper, the beneficial effects of *Ghrelin Receptor* (*GHS-R1A*) antagonists towards the suppression of food intake, are stated. This finding accounts for *Cathine*'s appetite-suppressing effects, when considering the blockage of *Ghrelin*'s receptor as a potential mechanism of action for the drug. The same article also reports that the 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:D*000855). *Cathine*'s product information describes the drug as anorexic. It has also been stated that *Ghrelin* in hypothalamic neurons controls *anorexia* and *cachexia* (*MeSH:D*002100) (*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). 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 activities* (*GO:*0008184) increased, whilst kidney *glycogen* levels decreased. The above suggests similar responses after *Ghrelin*'s or *Cathine*'s administration. 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). Conclusively, all the described links account for the putative association between the drug *Cathine* and the functions mediated by the gene *GHRL*.

3.4.2. Milnacipran-SLC6A4 association

The following case study represents the repositioning potential of the suggested methodology. It describes the known association between the drug *Milnacipran* (*DrugBank:DB*04896) and the gene *SLC6A4* (*Entrez Gene:*6532), which codes for *Milnacipran*'s known target, *serotonin transporter* (*SERT*). *SERT* was ranked at the top (1) 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

Table 4
Literature evidence for the *Cathine-GHRL* association.

Cathine is also called **PPA**	**Cathine**, . . .is one of the optical isomers of phenylpropanolamine (**PPA**), *Wikipedia*
PPA suppresses **hyperphagia**	**PPA** is capable of suppressing appetite in rats made **hyperphagic** by various stimuli (PMID:3310024)
Bulimia is an **hyperphagia**	*MeSH*
Ghrelin is involved in **bulimia**	***Ghrelin*** increases food intake . . .relevance in the regulation of human feeding behavior in individuals with eating disorders such as **bulimia** nervosa (PMID:2 1453750)
Cathine is an **anorexic** drug	Product Information
Ghrelin controls **cachexia**	***Ghrelin*** in concert with hypothalamic neurons control **anorexia** and cachexia (PMID:2 2632865)
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**	*Gene Ontology*
Ghrelin affects **Glycogen phosphorylase activity**	After the injection of intraperitoneal ***Ghrelin***, kidney **glycogen phosphorylase activities** increased (PMID:1 5302267)
Cathine affects **adrenergic receptors**	Regulation of adenylyl cyclase/*cAMP* in a **G protein**-mediated fashion by **Cathine** may possibly involve **adrenergic receptors** (PMID:1 5513978)
Adrenergic receptors are **G-protein coupled receptors**	*Gene Ontology*
Ghrelin's receptor is a **G-protein coupled receptor**	Growth hormone secretagogue receptor is a **G-protein coupled receptor** that binds ***Ghrelin*** (PMID:1 6382107)
Cathine boosts **single fertilization**	**Cathine** can enhance chances of **fertilization** in vivo (PMID:1 7158213)
Single fertilization and **spermatogenesis** pertain to reproduction	*Gene Ontology*
Ghrelin modulates **spermatogenesis**	***Ghrelin*** may be considered as a modulator of **spermatogenesis** (PMID:2 2360851)

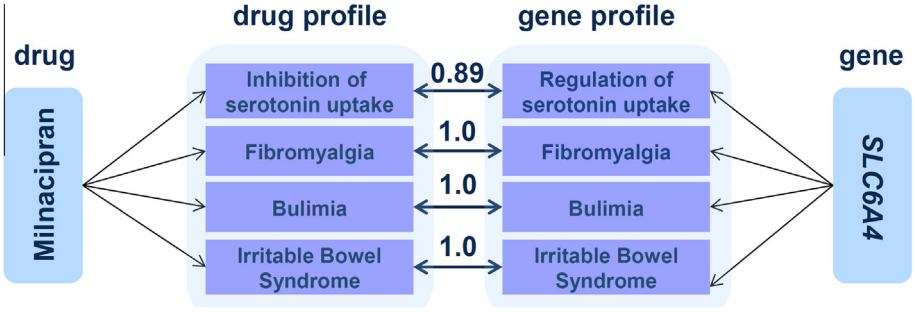

Fig. 6. *Milnacipran – SLC6A4* association representation. The figure demonstrates the indirect connections for the known association *Milnacipran* to its target-coding gene *SLC6A4*. The connections are generated from data published before 2009. As shown, *Fibromyalgia*, a condition to which *Milnacipran* was repositioned after 2009, participates in the establishment of the respective association.

Table 5
Literature evidence for the *Milnacipran-SLC6A4* association.

Milnacipran is a **serotonin–norepinephrine reuptake inhibitor**	*Wikipedia*
Inhibition of serotonin uptake is a **regulation of serotonin uptake**	*Gene Ontology*
SERT is responsible for the **regulation of serotonin uptake**	*Wikipedia*
Milnacipran cures **fibromyalgia**	In this Phase II study, **Milnacipran** led to statistically significant improvements in pain and other symptoms of **fibromyalgia** (PMID:1 6206355)
SLC6A4 polymorphism is related to **fibromyalgia**	Confirmation of an association between **fibromyalgia** and **serotonin transporter** promoter region polymorphism (PMID:1 1920428)
Milnacipran treats **bulimia nervosa**	**Milnacipran** in the treatment of **bulimia nervosa**: a report of 16 cases. (PMID:1 2650949)
SLC6A4 polymorphism is related to **bulimia nervosa**	The **serotonin transporter**, encoded by the *SLC6A4* gene, may also have an important role in eating disorders, as its availability is decreased in patients with **bulimia nervosa**. . .(PMID:1 4987118)
Milnacipran treats **Irritable Bowel Syndrome**	... **milnacipran** has potential clinical application in the treatment of visceral pain, such as in irritable bowel syndrome ... (PMID:2 1996314)
SLC6A4 is a biomarker of **Irritable Bowel Syndrome**	... suggesting that *SLC6A4* is a potential candidate gene involved in the pathogenesis of **Irritable Bowel Syndrome**. (PMID:2 2457857)

inhibitor (*SNRI*) initially approved for the treatment of *depression* (*MeSH:D*003863) (1996). In January 2009 *Milnacipran* was also approved for the treatment of *fibromyalgia* (*MeSH:D*005356). 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). Fig. 6 shows the suggested connections and Table 5 summarizes the textual pieces of evidence that support them. The connections were generated based on the *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 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*. 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 *fibromyalgia* (*PMID:*15378666; 16206355). Other reports confirm that the polymorphic region of *SLC6A4* is associated to *fibromyalgia* (*PMID:*11920428; 10555044). 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:D*043183) 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). 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.

4. Conclusion

In this paper we focus on interrelating drugs to genes via intermediate ontological concepts. We introduce a method that generates concept profiles for both drugs and genes from the literature. *Pointwise Mutual Information* is used to create the profiles by finding associated ontological terms, and also to measure the relatedness between a drug and a gene via the *nPMI* scores of their profiles' terms. We demonstrated the application of our approach towards predicting drug gene associations, and identifying candidate drug repositioning cases. In the first application, we evaluated the approach on three datasets. Results show that the suggested method achieves an *AUC* up to 0.88 in prioritizing true associations between approved drugs and *human* genes. The suggested approach has some advantages. First, it can successfully prioritize direct from indirect drug gene associations. Second, the method's performance towards the prediction of drug gene associations is equally good both when co-occurence based profiles and manually curated profiles are used. This suggests that the co-occurence based profiles that are statistically derived from literature are indeed reliable for associating drugs to genes. Other advantages of the suggested methodology is that it constitutes a fully corpus-based and unsupervised approach. The case study between *Cathine* and *GHRL* was analysed to demonstrate the potential of

our approach towards mining drug gene associations that were not discussed previously in the literature. For *Cathine*, the gene *GHRL* ranked among the top associated genes. Digging into the literature, we provided concrete pieces of evidence which suggest that this association is meaningful. With regards to the second application, we compiled a dataset comprising all the known drug repositioning cases that were *FDA* approved within the last 5 years. For these drugs we generated the profiles based on literature data before the year of approval of each repositioning case. We demonstrated that the application of the suggested approach results in respective drug profiles which always include the new indications, and that in the majority of the cases these indications are ranked high among drugs' profile terms. In addition, we analysed the case study between *Milnacipran* and *SLC6A4*, where we provided evidence that the new indication (*Fibromyalgia*) actively participated in the generated associations.

With regards to the limitations of the suggested method, our approach is bound by two factors: (1) the usage of *MEDLINE* titles and abstracts in order to extract the profiles' terms and measure the relatedness between drugs and genes, and, (2) the performance of the existing annotation methodologies used to identify drugs, genes, *MeSH Disease* and *GO* terms in text. Though the former limitation remains unremedied as long as the open access to full text articles is restricted, in our future work we will focus on improving and alleviating the impact or bias introduced by the used annotation methodologies. For example, one way to improve the annotation results would be using additional tools and considering the products of the inter-annotation agreement. We will also examine any potential improvements of the suggested method with the application of sentence level co-occurrence based statistics or the incorporation of *NLP* techniques. Scoring schemes considering the number of papers participating in each drug gene association will also be examined as a confidence value towards the further discrimination of putative drug gene associations. Finally, future work will be towards enriching the profiles of drugs and genes with physicochemical properties and additional protein structural information respectively. Work will also be conducted towards analysing the predicted drug gene associations, to identify possible repositioning of existing drugs (e.g., the use of docking analysis could be considered).

Acknowledgments

We kindly acknowledge funding by EU and BMWi (GeneCloud).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ymeth.2014.11.017.

References

[1] S.M. Paul, D.S. Mytelka, C.T. Dunwiddie, C.C. Persinger, B.H. Munos, S.R. Lindborg, A.L. Schacht, Nat. Rev. Drug Discovery 9 (3) (2010) 203–214, http://dx.doi.org/10.1038/nrd3078.
[2] M.R. Hurle, L. Yang, Q. Xie, D.K. Rajpal, P. Sanseau, P. Agarwal, Clin. Pharmacol. Ther. 93 (4) (2013) 335–341, http://dx.doi.org/10.1038/clpt.2013.1.
[3] M. Campillos, M. Kuhn, A.-C. Gavin, L.J. Jensen, P. Bork, Science 321 (5886) (2008) 263–266, http://dx.doi.org/10.1126/science.1158140.
[4] J. Lamb, E.D. Crawford, D. Peck, J.W. Modell, I.C. Blat, M.J. Wrobel, J. Lerner, J.-P. Brunet, A. Subramanian, K.N. Ross, M. Reich, H. Hieronymus, G. Wei, S.A. Armstrong, S.J. Haggarty, P.A. Clemons, R. Wei, S.A. Carr, E.S. Lander, T.R. Golub, Science 313 (5795) (2006) 1929–1935, http://dx.doi.org/10.1126/science.1132939 (PMID: 17008526).
[5] E. Lounkine, M.J. Keiser, S. Whitebread, D. Mikhailov, J. Hamon, J.L. Jenkins, P. Lavan, E. Weber, A.K. Doak, S. Ct, B.K. Shoichet, L. Urban, Nature 486 (7403) (2012) 361–367, http://dx.doi.org/10.1038/nature11159.
[6] Y. Yamanishi, M. Kotera, M. Kanehisa, S. Goto, Bioinformatics 26 (12) (2010) i246–i254, http://dx.doi.org/10.1093/bioinformatics/btq176.
[7] B. Chen, Y. Ding, D.J. Wild, PLoS Comput. Biol. 8 (7) (2012) e1002574, http://dx.doi.org/10.1371/journal.pcbi.1002574.
[8] Y. Wu, M. Liu, W.J. Zheng, Z. Zhao, H. Xu, Pac. Symp. Biocomput. (2012) 422–433.
[9] J. Mestres, E. Gregori-Puigjané, S. Valverde, R.V. Solé, Mol. BioSyst. 5 (9) (2009) 1051–1057.
[10] T. van Laarhoven, E. Marchiori, PLoS ONE 8 (6) (2013) e66952, http://dx.doi.org/10.1371/journal.pone.0066952.
[11] H. Chen, Z. Zhang, PLoS ONE 8 (5) (2013) e62975, http://dx.doi.org/10.1371/journal.pone.0062975.
[12] M. Takarabe, M. Kotera, Y. Nishimura, S. Goto, Y. Yamanishi, Bioinformatics 28 (18) (2012) i611–i618, http://dx.doi.org/10.1093/bioinformatics/bts413.
[13] S. Zhao, S. Li, PLoS ONE 5 (7) (2010) e11764, http://dx.doi.org/10.1371/journal.pone.0011764.
[14] K. Bleakley, Y. Yamanishi, Bioinformatics 25 (18) (2009) 2397–2403, http://dx.doi.org/10.1093/bioinformatics/btp433.
[15] F. Cheng, C. Liu, J. Jiang, W. Lu, W. Li, G. Liu, W. Zhou, J. Huang, Y. Tang, PLoS Comput. Biol. 8 (5) (2012) e1002503, http://dx.doi.org/10.1371/journal.pcbi.1002503.
[16] D. Emig, A. Ivliev, O. Pustovalova, L. Lancashire, S. Bureeva, Y. Nikolsky, M. Bessarabova, PLoS ONE 8 (4) (2013), http://dx.doi.org/10.1371/journal.pone.0060618.
[17] I. Vogt, J. Mestres, Mol. Inf. 29 (1–2) (2010) 10–14.
[18] P. Srinivasan, J. Am. Soc. Inform. Sci. Technol. 55 (5) (2004) 396–413.
[19] R. Jelier, G. Jenster, L.C.J. Dorssers, C.C.v.d. Eijk, E.M.v. Mulligen, B. Mons, J.A. Kors, Bioinformatics 21 (9) (2005) 2049–2058, http://dx.doi.org/10.1093/bioinformatics/bti268.
[20] K.B. Cohen, H.L. Johnson, K. Verspoor, C. Roeder, L.E. Hunter, BMC Bioinf. 11 (2010) 492, http://dx.doi.org/10.1186/1471-2105-11-492.
[21] T.-K. Jenssen, A. Lgreid, J. Komorowski, E. Hovig, Nat. Genet. 28 (1) (2001) 21–28, http://dx.doi.org/10.1038/ng0501-21.
[22] J. Hakenberg, C. Plake, R. Leaman, M. Schroeder, G. Gonzalez, Bioinformatics 24 (16) (2008) i126–i132, http://dx.doi.org/10.1093/bioinformatics/btn299.
[23] K.M. Hettne, R.H. Stierum, M.J. Schuemie, P.J.M. Hendriksen, B.J.A. Schijvenaars, E.M. v. Mulligen, J. Kleinjans, J.A. Kors, Bioinformatics 25 (22) (2009) 2983–2991, http://dx.doi.org/10.1093/bioinformatics/btp535.
[24] C. Plake, Gene annotation by automated literature analysis with an application to drug–target interaction prediction (Ph.D. thesis), Technischen Universitaet Dresden, Germany, 2010.
[25] A.S. Schwartz, M.A. Hearst, Pac. Symp. Biocomput. (2003) 451–462.
[26] A. Doms, M. Schroeder, Nucl. Acids Res. 33 (2005) W783–W786 (Web Server issue).
[27] C.D. Manning, H. Schütze, Foundations of Statistical Natural Language Processing, MIT Press, 1999 (Ch. Collocations).
[28] Y. Niu, D. Otasek, I. Jurisica, Bioinformatics 26 (1) (2010) 111–119, http://dx.doi.org/10.1093/bioinformatics/btp602.
[29] A.J. Prez, C. Perez-Iratxeta, P. Bork, G. Thode, M.A. Andrade, Bioinformatics 20 (13) (2004) 2084–2091, http://dx.doi.org/10.1093/bioinformatics/bth207.
[30] G. Bouma, Normalized (pointwise) mutual information in collocation extraction (2009) 31–40.
[31] A. Schlicker, F.S. Domingues, J. Rahnenführer, T. Lengauer, BMC Bioinf. 7 (1) (2006) 302, http://dx.doi.org/10.1186/1471-2105-7-302.
[32] Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda, M. Kanehisa, Bioinformatics 24 (13) (2008) i232–i240, http://dx.doi.org/10.1093/bioinformatics/btn162.
[33] C.D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008 (Chapter 8).
[34] C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, Y. Djoumbou, R. Eisner, A.C. Guo, D.S. Wishart, Nucl. Acids Res. 39 (suppl. 1) (2011) D1035–D1041, http://dx.doi.org/10.1093/nar/gkq1126.
[35] A.P. Davis, C.G. Murphy, R. Johnson, J.M. Lay, K. Lennon-Hopkins, C. Saraceni-Richards, D. Sciaky, B.L. King, M.C. Rosenstein, T.C. Wiegers, C.J. Mattingly, Nucl. Acids Res. 41 (D1) (2012) D1104–D1114, http://dx.doi.org/10.1093/nar/gks994.
[36] Z. Wu, M. Palmer, Verbs semantics and lexical selection, in: Stroudsburg, PA, USA, 1994, pp. 133–138. doi: 10.3115/981732.981751.
[37] D. Lin, An Information-Theoretic Definition of Similarity, Morgan Kaufmann, 1998, pp. 296–304.