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