MINING THE BIOMEDICAL LITERATURE TO PREDICT SHARED DRUG TARGETS IN DRUGBANK

Horacio Caniza*, Diego Galeano*, and Alberto Paccanaro

Abstract— The current drug development pipelines are characterised by long processes with high attrition rates and elevated costs. More than 80% of new compounds fail in the later stages of testing due to severe side-effects caused by unknown biomolecular targets of the compounds. In this work, we present a measure that can predict shared targets for drugs in DrugBank through large scale analysis of the biomedical literature. We show that using MeSH ontology terms can accurately describe the drugs and that appropriate use of the MeSH ontological structure can determine pairwise drug similarity. **Index Terms**— MeSH terms, drug descriptors, drug targets, drugbank

1 INTRODUCTION

DESPITE breakthroughs in genomics [1], elucidating the molecular modes of action of pharmaceutical drugs has proven to be a difficult problem. Current drug discovery follows a pipeline characterised by a drawn-out process with low success rates and high costs [2].

The drug discovery process begins with the identification of a specific protein target whose behaviour the drug aims to change [3]. However, drugs tend to interact with proteins other than the intended target. These off-target effects disrupt the function of proteins that are usually unrelated to the condition the drug aims to treat. Most severe off-target effects are discovered during the final clinical trials in humans, an important cause of attrition in the pipeline [4]. Computational prediction of on/off target effects of drugs can help stem the high attrition rates before clinical trials, reduce patient risk and drug development costs [5].

Starting in the 2017 XML release (version 5.0.5) DrugBank [6] incorporates MeSH (Medical Subject Headings) terms into its entries [7]. MeSH is a controlled vocabulary organised into 16 interconnected, hierarchically organised ontologies describing different areas of knowledge (e.g. Anatomy). MeSH was designed to index the biomedical literature in Medline/PubMed. The structure and content of these large-scale hand-curated ontologies have also been shown to be effective in quantifying pairwise disease similarity, determining the distance between disease modules on the Interactome, and building disease-symptoms networks [8,9, 10, 11].

In this work, we present a MeSH based method to quantify pairwise drug similarities. We show that the content of the MeSH ontologies can be used to accurately describe FDA approved drugs in DrugBank. We also show that the structure of the MeSH ontology, when used appropriately, outperforms the Tanimoto chemical similarity index, in terms of predicting share targets. Furthermore, we show that although the Tanimoto chemical similarity is good predictors of share targets for highly similar drugs (Tanimoto score > 60%) [12], in general, Tanimoto scores are not good predictors of on/off targets for chemically dissimilar drugs [12].

2 MATERIALS AND METHODS

The approach we present here characterises FDA approved drugs through sets of MeSH terms. We construct the set of annotations for a drug by combining the MeSH terms available in DrugBank and the set of MeSH terms annotating the publications referenced in the drug's DrugBank entry. The similarity between two drugs is then quantified by the semantic similarity of their annotations. An outline of our method is presented in Fig. 1.

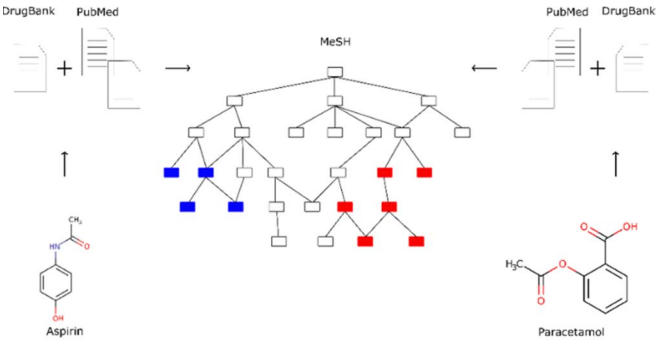


Fig.1. Outline of the Mesh ontology based drug similarity. For each pair of FDA-approved drug in Drugbank, we obtained the MeSH terms associated to the drug entry and retrieved also the MeSH terms associated to the publications for each drug as listed in the database. The set of MeSH terms associated to each drug, e.g., Aspirin (blue terms) and Paracetamol (red terms) were annotated into the MeSH ontology. Several ontology-based similarity measures were computed between these terms and obtained a single number that

- *H. Caniza is with Facultad de Ciencias de la Ingeniería. Universidad Paraguayo Alemana. Asunción, Paraguay. E-mail horacio.caniza@upa.edu.py.*
- *D. Galeano is with Department of Computer Science. Royal Holloway, University of London. Egham TW200EX UK. E-mail Diego.Galeano.2014@rhul.ac.uk*
- *A. Paccanaro is with D. Galeano is with Department of Computer Science. Royal Holloway, University of London. Egham TW200EX UK. E-mail Alberto.Paccanaro@rhul.ac.uk*

** These authors contributed equally to this work.*

978-1-5386-3057-0/17/$31.00 © 2017 IEEE European Union

indicates the similarity between the two drugs.

The Drugbank database [6] is a chemoinformatics resource that combines information about known drug compounds, e.g. drug chemical structure, their Anatomical Therapeutic and Chemical (ATC) category, their known biomolecular targets and references to the relevant publications. The current release contains 8,261 compounds and 4,338 non-redundant drug targets. In this work, we analyse the set of 1,416 FDA approved drugs with known targets and SMILES (Simplified Molecular-Input Line-Entry System) representation of the drug chemical structure.

A drug is annotated with the set of MeSH terms available in DrugBank. Terms in each MeSH ontology are organised in increasing specificity as a Directed Acyclic Graph (DAG) and a term can belong to multiple ontologies. For example, Anatomy [A] -> Body Regions [A01] -> Extremities [A01.378] -> Upper Extremity [A.01.378.000] -> Hand [A.01.378.000.667] -> Wrist [A.01.378.000.667.715]. We enrich this set of MeSH available in DrugBank terms by mining the MeSH terms from the literature referenced in the drug's entry. The procedure for annotation of drug MeSH terms and ontology combination follows the procedure presented in [10].

The similarity of a pair of drugs is given by the semantic similarity of the MeSH terms in the ontology. We evaluated a set of representative semantic similarity measures. The summary of the semantic similarity measures is presented in Table 1.

The individual MeSH ontologies are combined into a single ontology [10]. We have terms annotated in 15 MeSH ontologies. The chosen ontologies are the ones that annotate most terms, namely Anatomy (A), Diseases (C) and Phenomena and Processes (G), and remove those that may introduce bias in the performance evaluation. Importantly, the Chemical and Drugs [D] ontology in MeSH contains biomolecular information [7]. For this reason, the entire Chemicals and Drugs ontology is removed and all its terms ignored in the semantic similarity calculations. In addition, we combine the ontologies by introducing a root node linked to the root nodes of the chosen ontologies [10].

To evaluate our method, we define a binary classification problem in which the pairwise MeSH based similarity scores are used to predict share on/off targets for two drugs. The performance of the classification is given by the area under the ROC curve (AUC). The classification performance of the MeSH based similarity measures are compared to that of the Tanimoto chemical similarity [12]. For a pair of drugs, the Tanimoto similarity is determined by comparing the SMILES binary fingerprint of the drug compounds.

Formally the Tanimoto chemical similarity of two drugs Da and Db is given by:

$$T(D_a, D_b) = \frac{\sum_{bit,i} F_{a,i} \cap F_{b,i}}{\sum_{bit,i} F_{a,i} \cup F_{b,i}} \qquad (1)$$

where F_a and F_b are the hash fingerprints for drugs Da and Db, respectively.

TABLE 1
SUMMARY OF THE SEMANTIC SIMILARITY MEASURES

Method	Definition
Resnik [13]	$Resnik(a,b) = \max_{t_a \in terms_a, t_b \in terms_b} \left(-log\left(P\left(LCA\left(t_a,t_b\right)\right)\right)\right)$
Lin [14]	$Lin(a,b) = \max_{t_a \in terms_a, t_b \in terms_b} \left(\frac{2 * \log(P(LCA(t_a, t_b))}{\log(P(t_a)) + \log(P(t_b))}\right)$
Jiang [15]	$Jiang(a,b) = \max_{t_a \in terms_a, t_b \in terms_b} \left(1 - \frac{2 * \log(P(LCA(t_a,t_b)) - \log(P(t_a)) - \log(P(t_b)}{\max(Jiang \forall (t_a,t_b))}\right.$
simUI [16]	$simUI(a,b) = \frac{\lvert terms(a) \cap terms(b) \rvert}{\lvert terms(a) \cup terms(b) \rvert}$
simGIC [17]	$simGIC(a,b) = \frac{\sum_{t \in terms(a) \cap terms(b)} IC(t)}{\sum_{t \in terms(a) \cup terms(b)} IC(t)}$
Jaccard [18]	$Jaccard(a,b) = \frac{\lvert annot(a) \cap annot(b) \rvert}{\lvert annot(a) \cup annot(b) \rvert}$
Number of common terms	$NumCommon(a,b) = \lvert annot(a) \cap annot(b) \rvert$

LCA denotes the lowest common ancestor between a pair of terms in the ontology and P(a) denotes the probability of a term in the ontology, given by the fraction of terms annotated by the term and the total number of annotations.

3 Results

We analyse 1,416 FDA approved drugs with known

SMILES 2D chemical structure. We evaluate the performance of the measures via binary classification problem where the pairwise drug similarities are used to predict whether two drugs share an on/off target. The performance of the measures is evaluated by computing the area under the ROC curve (AUC). Fig. 2 shows the ROC curves for the representative set of semantic similarity measures evaluated (Resnik [13], Lin [14], Jiang and Conrah [15], SIMui [16] and simGIC [17]), similarities measure based on set intersection that do not consider the structure of the ontology (Overlap, Jaccard and Shared terms) and our baseline Tanimoto chemical similarity measure. The best performing method is SIMui, a semantic based measure.

To understand the performance of the prediction, the probability density function for Tanimoto chemical similarity and the MeSH based method SIMui is shown in Figure 3. The top figure shows the distribution of similarities for all the pairs, with the blue bars representing the Tanimoto similarity values and the orange bars the SIMui values. The mean similarity for the Tanimoto similarity values is 0.37 and for SIMui 0.16.

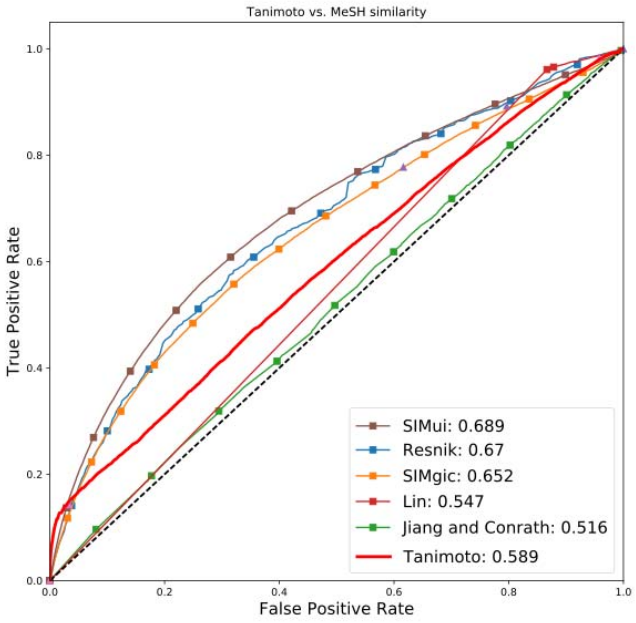


Fig. 2. Performance of the semantic-based similarity measures. Each ROC curve shows the performance of the corresponding semantic similarity method. The AUC for each method is shown in the legend. The baseline method is the Tanimoto chemical similarity (red line). SIMui is the semantic-based similarity that obtains the highest AUC in the prediction.

Fig. 3 bottom compares, for both measures, the distribution of similarity scores for diseases with and without shared pairs. In both cases the difference in means is significant (P value 0.0 within machine precision). Comparing the difference in means between pairs with shared targets and not shared targets for SIMui is 0.1 and for Tanimoto 0.04. The difference is significant with a P value of 0.0 (within machine precission).

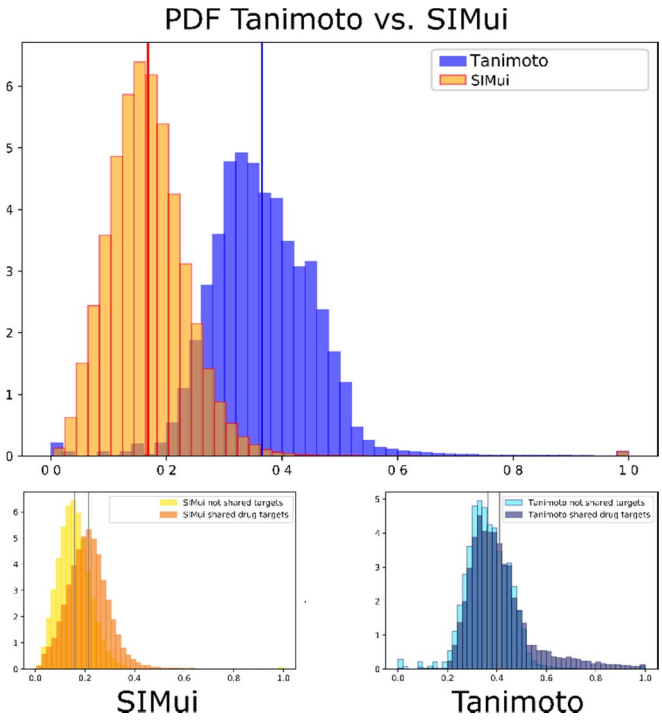


Fig.3. Distribution of similarity values. (**Top**) Probability density function of the pairwise similarity values for SIMui and Tanimoto similarities. (**Bottom**) Comparison of the distribution of pairwise similarity values between drug pairs with and without shared targets.

One important question is to understand to what extent the method's performance is dependent on the quality of the annotations, on the ontological structure, or both. To assess the effects of MeSH's ontological structure, we analysed the performance of similarity measures which disregard the ontology structure and are based on the overlap of the MeSH ontology terms annotating the drugs. For instance, in Figure 4 we show that Jaccard index performs with an AUC of 0.68, only 1% below our best performing method. While the structure of the ontology provides valuable information, and improves the performance of our method, the MeSH terms themselves are accurate descriptors of FDA approved drugs.

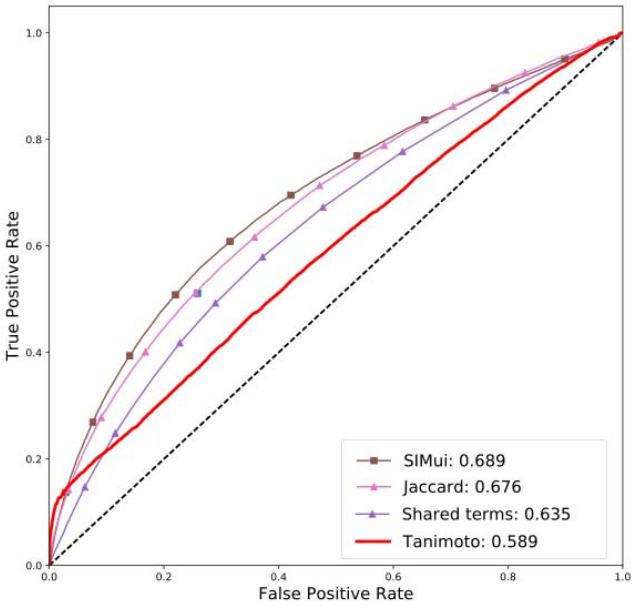


Fig.4. Performance of set intersection-based similarity measures. Each ROC curve shows the performance of the corresponding semantic similarity method. The AUC for each method is shown in the legend. The baseline method is the Tanimoto chemical similarity (red line). Jaccard index is as good predictor of shared targets as semantic-based similarity measures.

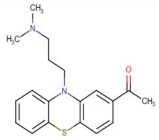
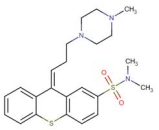
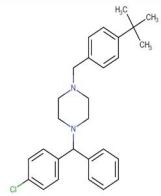
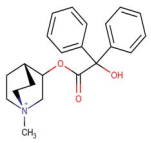
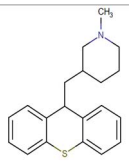
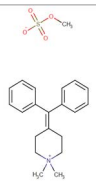
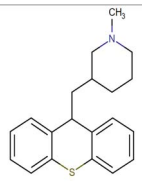
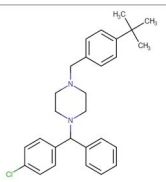
4 Discussion

In this paper, we introduced a method that uses MeSH ontology terms to accurately describe 1,416 FDA-approved drugs in DrugBank. We have shown that the semantic similarity measures can be used to predict whether two drugs share targets. Our method SIMui outperforms the Tanimoto chemical similarity by 10% AUC, as showed in Fig. 2.

Interestingly, biological evidence indicates that chemical similarity is not always a good indicator of shared targets [18], i.e.: i) the chemical space of the binding site of the protein target may be wide, meaning that drugs with different chemical structures can bind to the same protein; ii) binding sites in proteins can be rather different; iii) high chemical similarity does not always imply shared protein targets [12]. To illustrate this point, we present in Table 2 a set of examples that have low Tanimoto similarity and high SIMui that are known to share targets.

Our analysis shows that the MeSH terms associated to publications referenced in DrugBank are good descriptors of FDA-approved drugs, and that the MeSH ontology structure provides valuable hierarchical information for calculating distances between the sets of terms. The pairwise drug similarity has shown to be a good predictor of shared targets between drugs and can be further incorporated in advanced Machine Learning methods to enhance drug target prediction.

TABLE 2
DRUGS KNOWN TO SHARE TARGETS

Drug 1	Drug 2	Tanimoto	SIMui
 Acepromazine	 Thiothixene	30%	85%
 Buclizine	 Clidinium	36%	83%
 Metixene	 Diphemanil Methylsulfate	36%	83%
 Metixene	 Buclizine	33%	85%

Representative pair of FDA-approved drugs known to share targets. Tanimoto chemical similarity does not suggest shared targets (< 40%) whereas SIMui drug similarity is above 80%.

Acknowledgment

This work was supported, in part, by the the Consejo Nacional de Ciencia y Tecnología Paraguay (CONACYT)

Paraguay Grant INVG01-112 (14-INV-088) and PINV15-315 (14-INV-088), the Biotechnology and Biological Sciences Research Council (BBSRC), grants BB/K004131/1, BB/F00964X/1 and BB/M025047/1, and the Programa Nacional de Becas de Postgrado en el Exterior Don Carlos Antonio López (BECAL) from the Republic of Paraguay.

REFERENCES

[1] Kennedy, Donald. "Breakthrough of the year." Science 318.5858 (2007): 1833-1833.

[2] Rawlins, Michael D. "Cutting the cost of drug development?" Nature reviews Drug discovery 3.4 (2004): 360-364.

[3] Lombardino, Joseph G., and John A. Lowe. "The role of the medicinal chemist in drug discovery—then and now." Nature Reviews Drug Discovery 3.10 (2004): 853-862.

[4] Kola, Ismail, and John Landis. "Can the pharmaceutical industry reduce attrition rates?" Nature reviews Drug discovery 3.8 (2004): 711-716.

[5] Jorgensen, William L. "The many roles of computation in drug discovery." Science 303.5665 (2004): 1813-1818.

[6] Law, Vivian, et al. "DrugBank 4.0: shedding new light on drug metabolism." Nucleic acids research 42. D1 (2014): D1091-D1097. URL: https://www.drugbank.ca/

[7] Medical Subject Headings. URL: https://www.nlm.nih.gov/mesh/

[8] Van Driel, Marc A., et al. "A text-mining analysis of the human phenome." European journal of human genetics 14.5 (2006): 535-542.

[9] Robinson, Peter N., et al. "The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease." The American Journal of Human Genetics 83.5 (2008): 610-615.

[10] Caniza, Horacio, Alfonso E. Romero, and Alberto Paccanaro. "A network medicine approach to quantify distance between hereditary disease modules on the interactome." Scientific reports 5 (2015).

[11] Zhou, XueZhong, et al. "Human symptoms–disease network." Nature communications 5 (2014).

[12] Galeano, Diego, and Alberto Paccanaro. "Drug targets prediction using chemical similarity." Computing Conference (CLEI), 2016 XLII Latin American. IEEE, 2016.

[13] Resnik, Philip. "Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language." J. Artif. Intell. Res.(JAIR) 11 (1999): 95-130.

[14] Lin, Dekang. "An information-theoretic definition of similarity." ICML. Vol. 98. No. 1998. 1998.

[15] Jiang, J. J., and D. W. Conrath. "International Conference Research on Computational Linguistics (ROCLING X)." (1997): 9008-9022.

[16] Pesquita, Catia, et al. "Metrics for GO based protein semantic similarity: a systematic evaluation." BMC bioinformatics 9.5 2008): S4.

[17] Jaccard, Paul. "The distribution of the flora in the alpine zone." New phytologist 11.2 (1912): 37-50.

[18] James L. Kofron, and Linda M. Traphagen. "Do structurally similar molecules have similar biological activity?" Journal of medicinal chemistry 45.19 (2002): 4350-4358.