SEHS
science, engineering
and health studies

Research Article

Science, Engineering and Health Studies
https://li01.tci-thaijo.org/index.php/sehs
ISSN (Online): 2630-0087

Functionality-based similarities for uncovering relationships between drugs and diseases

Thitipong Kawichai[1, 2], Apichat Suratanee[3] and Kitiporn Plaimas[1, 4]*

[1] Advanced Virtual and Intelligent Computing (AVIC) Center, Department of Mathematics and Computer Science, Faculty of Science, Chulalongkorn University, Bangkok 10330, Thailand
[2] Department of Mathematics and Computer Science, Academic Division, Chulachomklao Royal Military Academy, Nakhon Nayok 26001, Thailand
[3] Department of Mathematics, Faculty of Applied Science, King Mongkut's University of Technology North Bangkok, Bangkok 10800, Thailand
[4] Omics Science and Bioinformatics Center, Faculty of Science, Chulalongkorn University, Bangkok 10330, Thailand

***Corresponding author**:
Kitiporn Plaimas
kitiporn.p@chula.ac.th

***Received**: 22 January 2021*
***Revised**: 27 January 2021*
***Accepted**: 11 February 2021*
***Published**: 14 December 2021*

***Citation**:*
Kawichai, T., Suratanee, A., and Plaimas, K., (2021). Functionality-based similarities for uncovering relationships between drugs and diseases. Science, Engineering and Health Studies, 15, 21040005.

ABSTRACT

Drug repositioning is a process of discovering new indication for existing drugs. The similarities based on drug- and disease-associated proteins can be used to reveal the relationships between drugs and diseases, between two drugs, or between two diseases for drug repositioning. Due to a lack of complete data about drug- and disease-associated proteins, this strategy could be directly affected by the limited number of proteins under consideration. To overcome this limitation, more extensive information about drugs and diseases such as gene ontology terms, functional annotations of genes and gene products, could be used. Herein, we provided a comprehensive exploration of using functionality-based similarities to uncover the relationships among drugs and diseases. After comparing seven different similarity indices, it is found that the derived Jaccard index was the most suitable one for computing functionality-based similarity scores. The predictions of drug-disease, drug-drug, and disease-disease associations for drug repositioning were significantly improved with an accuracy of 89%, 67%, and 83%, respectively, by utilizing functionality-based similarities. The case studies showed that our approach can identify the drug-disease associations that have been under investigation such as those between tolcapone and attention deficit-hyperactivity disorder and between nicorandil and type 2 diabetes mellitus.

Keywords: drug repositioning; functionality-based similarity; gene ontology; similarity index

1. INTRODUCTION

Developing new drugs to markets is expensive and time-consuming. For only one new drug achieved, it takes 12 to 16 years and almost US $2 billion on average (Nelson et al., 2018). In addition, many drug-like compounds have failed and could not enter the stage of clinical trials due to their inadequate safety and efficacy (Yella et al., 2018). With the availability of drug efficacy and safety information for approved drugs, the discovery of their new therapeutic indications, also known as drug repositioning, can significantly reduce the time, costs, and failure rate of drug discovery and development. To support drug repositioning, computational approaches are the most promising tools to efficiently propose plenty of potential drug-disease associations for further validation and development on wet lab experiments.

Because genes and proteins play crucial roles in drug actions and disease processes at the molecular level, similarities based on drug target proteins and disease-associated genes can be leveraged to infer new drug-disease associations. In the comparative toxicogenomics database (CTD), associations between chemicals and diseases were inferred based on the genes shared between the manually curated chemical-gene and disease-gene relationships (Davis et al., 2008). Since a gene-based similarity between two diseases could indicate shared causes or even treatments for those diseases, several studies focused on the genetic basis of disease-disease similarity. For example, Lewis and colleagues computed similarities between diseases based on the Jaccard similarity index and employed genome-wide association studies to identify disease associations (Lewis et al., 2011). Based on the assumption that similar drugs would show similar indications, drug-drug similarity scores computed based on the Tanimoto coefficients and drug-interacting proteins were used to predict drug-disease associations (Huang et al., 2015). Although the gene- or protein-based similarities could point to some potential drug-disease associations, some limitations of the methods such as using drug target proteins without considering genes or proteins that are affected downstream, could disguise true drug-disease associations or limit the predicted results to only the drugs and diseases obviously involved with each other.

Under the current situation, it is costly to identify complete sets of genes and proteins affected by drugs and diseases, utilizing more extensive information about drugs and diseases, as provided by gene ontology (GO) annotations, is a promising strategy to overcome the limitations of traditional methods. GO terms are controlled vocabularies used to describe biological functions of genes and gene products, such as RNAs and proteins (Hill et al., 2008). Davis and colleagues compared the numbers of genes and GO functions shared between the old and new diseases treated by three repositioned drugs, including raloxifene, thalidomide, and sildenafil. They found that in only one case did the new and old disease are associated with the same genes, whereas all drugs showed similar GO functions between their old and new diseases (Davis et al., 2016). This suggests that similarities based on GO functions, termed as functionality-based similarities or GO-based similarities, may improve the identification of drug-disease associations. However, there is no study providing a comprehensive analysis that uses all three functionality-based similarities, including drug-disease, drug-drug, and disease-disease similarity, for drug repositioning.

In this study, we comprehensively explored how to exploit drug-disease, drug-drug, and disease-disease similarity based on GO functions to uncover the relationships between drugs and diseases. Two main objectives of this study were to find the most appropriate similarity index for computing functionality-based similarities and to assess the utilization of functionality-based similarity scores for the classifications of drug-disease associations, drug-drug associations (in terms of being able to treat similar diseases), and disease-disease associations (in terms of being treated by similar drugs). The overview of this study is shown in Figure 1. Initially, drug-GO and disease-GO associations were constructed based on drug-protein associations, disease-protein associations, and the functional annotation of human proteins. Based on known drug-disease associations, we could generate all drug-drug pairs labeled with "shared" or "not shared" some common associated diseases. Similarly, we also had all disease-disease pairs labeled with "shared" or "not shared" some common associated drugs. Seven different similarity measures, including the Jaccard, Braun-Blanquet, Simpson, Cosine, Sorgenfrei, McConnaughey, and derived Jaccard index, were used to compute the functionality-based similarity scores of all drug-disease, drug-drug, and disease-disease pairs. Then, we compared the predicting powers of the similarity scores based on those seven similarity indices to select the best-performing similarity index for computing functionality-based similarities. After that, we compared the performance of functionality-based similarities in the classifications of the drug-disease, drug-drug, and disease-disease associations with that of protein-based similarities to evaluate our method. Lastly, we demonstrated the practicality of using functionality-based similarities to classify drug-disease, drug-drug, and disease-disease associations. Three case studies selected from the inferred associations of each association type were validated by searching for supporting evidence from published literature and public databases.

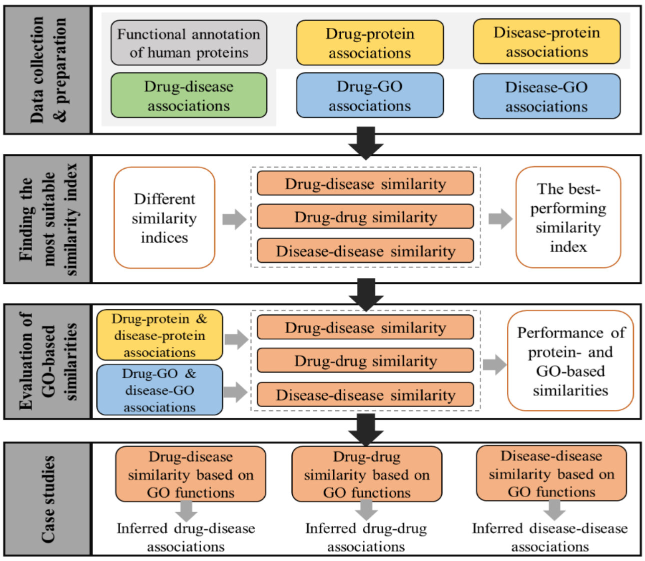

Data collection & preparation
Functional annotation of human proteins
Drug-protein associations
Disease-protein associations
Drug-disease associations
Drug-GO associations
Disease-GO associations
Finding the most suitable similarity index
Different similarity indices
Drug-disease similarity
Drug-drug similarity
Disease-disease similarity
The best-performing similarity index
Evaluation of GO-based similarities
Drug-protein & disease-protein associations
Drug-GO & disease-GO associations
Drug-disease similarity
Drug-drug similarity
Disease-disease similarity
Performance of protein- and GO-based similarities
Case studies
Drug-disease similarity based on GO functions
Drug-drug similarity based on GO functions
Disease-disease similarity based on GO functions
Inferred drug-disease associations
Inferred drug-drug associations
Inferred disease-disease associations

Figure 1. Schematic diagram describing an overview of this study

2. MATERIALS AND METHODS

2.1 Data collection

Four data sets were required for this study, including the functional annotation of human proteins, drug-disease associations, drug-protein associations, and disease-protein associations (Figure 1). The GO annotation data of human proteins were retrieved from the gene ontology annotation (GOA) database version 191 (Huntley et al., 2015). Our drug-disease associations were generated by combining two data sets provided in the study of the PREDICT method (Gottlieb et al., 2011) and the comparative toxicogenomics database (CTD), released in August 2019 (Davis et al., 2019). The former is the manually curated gold-standard data set whose drug-disease associations were assembled from different sources, and only the associations that overlapped with more than one source were kept in this data set (Gottlieb et al., 2011). In the CTD, only therapeutic drug-disease relations supported by the literature were selected. All approved drugs and their target proteins were collected from DrugBank version 5.1.3 (Wishart et al., 2018). All diseases and their associated genes were downloaded from DisGeNET version 6.0 (Piñero et al., 2016). All disease-associated genes were mapped to their corresponding protein identifiers to obtain the disease-protein associations.

2.2 Construction of drug-GO and disease-GO associations

GO can be classified into three non-overlapping classes (also known as GO aspects), including molecular function (MF), biological process (BP), and cellular component (CC). "MF" refers to a cellular activity that a gene product performs, such as "alcohol dehydrogenase activity" and "retinol dehydrogenase activity." "BP" is a molecular process comprising one or more biological activities such as "neurotransmitter secretion" and "limb development." CC is a cellular location where a gene product may function such as "plasma membrane." GO functions are expressed as a hierarchical structure, where high-level GO terms provide broader information than low-level GO terms. For example, a BP GO term "serotonin secretion" is a child of the parent BP term "neurotransmitter secretion." For each gene or protein, a particular set of relevant GO terms is annotated, and all parent GO terms of an annotated term are also associated with that gene or protein.

To create drug-GO and disease-GO associations, the drug-protein, disease-protein, and functional annotation data of human proteins were used. All aspects of GO terms, including BP, MF, and CC, were utilized to collect as much as functional information about drugs and diseases. Based on the GO annotation data, GO functions annotated for all target proteins of a drug were mapped to that drug. Similarly, GO functions annotated for all proteins associated with a disease were directly linked to that disease. Because similar GO terms from different levels could be connected to the same drug or disease, the drug-GO and disease-GO associations with the GO terms that were not the most detailed annotation terms (leaf terms) were removed. Then, we performed one-sided Fisher's exact tests to examine whether a drug (or a disease) and a GO function of a particular pair are specifically associated with each other or not. To reduce the false discovery rate (FDR) in the multiple testing, we transformed all *p*-values obtained from the Fisher's exact tests into *q*-values by using the Benjamini-Hochberg method, as shown in Equation (1), where *m* is the total number of tests, and *i* is the rank of a *p*-value when we sort all *p*-values in an ascending order. Only the drug-GO and disease-GO associations, which had the *q*-value less than 0.05 were preserved into the final list of our drug-GO and disease-GO associations.

$$q\text{-value} = \min\left\{\left(p\text{-value}\right)\left(\frac{m}{i}\right), 1\right\} \tag{1}$$

2.3 Three functionality-based relationships of drugs and diseases under investigation

Since drug- and disease-associated GO functions can indicate biological functions, which drugs and diseases are involved, it is more likely that the associations between drugs and diseases, between two drugs, and between two diseases can be detected by the drug-disease, drug-drug, and disease-disease similarities based on GO functions. To perform a comprehensive study of the functionality-based similarities, three functionality-based relationships of drugs and diseases were investigated, including the drug-disease, drug-drug, and disease-disease relationship (Figure 2). To be capable of making an inference about drug-disease associations, the drug-drug and disease-disease associations were formulated based on how they are mapped to some common diseases and drugs, respectively.

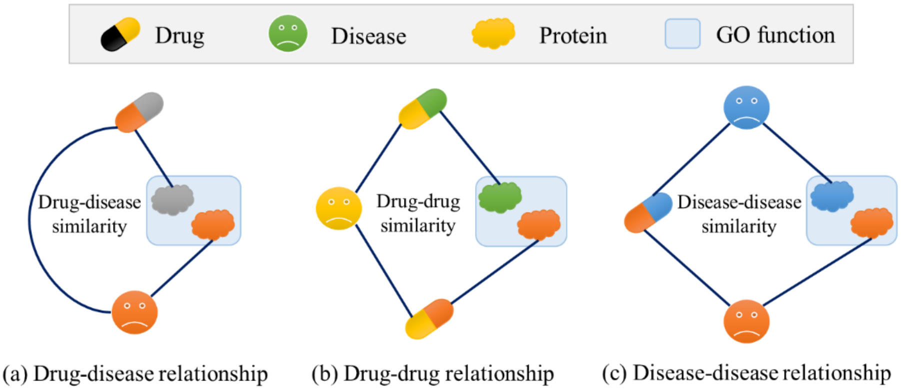

Drug
Disease
Protein
GO function
Drug-disease similarity
Drug-drug similarity
Disease-disease similarity
(a) Drug-disease relationship
(b) Drug-drug relationship
(c) Disease-disease relationship

Figure 2. Three functionality-based relationships investigated in this study

For a drug-disease association (Figure 2a), we presumed that a drug and a disease can be associated with different proteins, but these proteins may work together in the same biological functions or be associated with similar GO functions. We measured the drug-disease similarity based on drug- and disease-associated GO functions to represent its association score. For two distinct drugs those share a common disease (Figure 2b), they could interact with different proteins, which affected similar downstream biological functions to treat the same disease. To predict drug-drug associations, we measured the similarity of any drug-drug pairs based on drug-associated GO functions. For two diseases those can be treated by a common drug (Figure 2c), one disease may be relevant to another disease by being associated with different proteins involved in some common GO functions. The functionality-based similarity of any disease-disease pairs was measured as the scores of disease-disease associations.

Initially, all drug-disease, drug-drug, and disease-disease pairs with their labels were required for further measurements of their similarities. Based on our known drug-disease associations, all pairs of any pair type can be divided into two classes, which are known (positive-labeled) and unknown (negative-labeled) associations. The methods to generate and label the drug-disease, drug-drug, and disease-disease pairs are described as follows:

Drug-disease pairs: all possible drug-disease pairs were generated by combining all drugs and all diseases that we had. The drug-disease pairs in the list of our collected drug-disease associations were labeled as positive whereas the remaining drug-disease pairs were negative.

Drug-drug pairs: all possible drug-drug pairs were constructed by pairing two different drugs together based on the list of all drugs that we had. The drug-drug pairs that share at least one common disease were labeled as positive and the remaining drug-drug pairs were labeled as negative.

Disease-disease pairs: the list of all diseases was used to generate all possible disease-disease pairs by pairing two distinct diseases together. The disease-disease pairs sharing at least one common drug were positive samples whereas the remaining pairs were negative samples.

2.4 Measurement of protein- and functionality-based similarities

To predict drug-disease, drug-drug, and disease-disease associations, the functionality-based similarities were measured between drugs and diseases, between two drugs, and between two diseases, respectively. The drug-GO and disease-GO associations were used to measure all of the functionality-based similarities. Based on the drug-protein and disease-protein associations, we also measured the protein-based similarities, and used them as the baseline to finally compare with the functionality-based similarities. To compute drug-disease, drug-drug, and disease-disease similarity scores, seven similarity indices were used in this study. Because different similarity indices can be variously computed and are suitable for different tasks, the best-performing one in the classification of the associations is considered as the most suitable similarity index for computing the functionality-based similarities.

We posited that *x* and *y* represent a drug or a disease, and $S_{SimilarityIndex}(x, y)$ is a function for computing a similarity score between *x* and *y* based on a particular similarity index. *X* and *Y* are the sets of the proteins or GO functions associated with *x* and *y*, respectively. |·| is the number of all elements in a set, and "\" is the set difference of any two sets. The formulas of those seven similarity indices are shown in Equations (2)-(8).

$$S_{Jaccard}(x,y)=\frac{|X\cap Y|}{|X\cup Y|} \tag{2}$$

$$S_{BraunBlanquet}(x,y)=\frac{|X\cap Y|}{\max(|X|,|Y|)} \tag{3}$$

$$S_{Simpson}(x,y)=\frac{|X\cap Y|}{\min(|X|,|Y|)} \tag{4}$$

$$S_{Cosine}(x,y)=\frac{|X\cap Y|}{\sqrt{|X|\cdot|Y|}} \tag{5}$$

$$S_{Sorgenfrei}(x,y)=\frac{(|X\cap Y|)^2}{|X|\cdot|Y|} \tag{6}$$

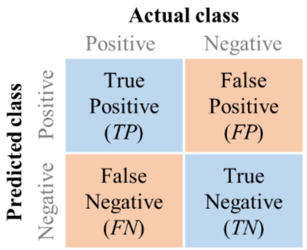
$$S_{McConnaughey}(x,y)=\frac{(|X\cap Y|)^2-(|X\setminus Y|\cdot|Y\setminus X|)}{|X|\cdot|Y|} \tag{7}$$

$$S_{DerivedJaccard}(x,y)=\frac{\log(1+|X\cap Y|)}{\log(1+|X\cup Y|)} \tag{8}$$

3. RESULTS

3.1 Performance evaluation

Drug-disease, drug-drug, and disease-disease associations are classified as either positive or negative directly based on their drug-disease, drug-drug, and disease-disease similarity scores, respectively. At a specific threshold score, the drug-disease, drug-drug, and disease-disease associations can be categorized according to their actual and predicted classes and summarized in a table called a confusion matrix (Figure 3). It is Noted that true positives, false positives, false negatives, and true negatives are *TP*, *FP*, *FN*, and *TN,* respectively.



Predicted class	Actual class: Positive	Actual class: Negative
Positive	True Positive (*TP*)	False Positive (*FP*)
Negative	False Negative (*FN*)	True Negative (*TN*)

Figure 3. Confusion matrix

To assess the performance of each similarity index and demonstrate the superiority of the functionality-based similarities, we employed the receiver operating characteristic (ROC) curves and precision-recall (PR) curves. An ROC curve is a plot showing the performance of a binary classification model at every threshold score. This plot is commonly used to compare the performance of several binary classifiers. To create an ROC curve, the true positive rates (TPRs) and false positive rates (FPRs) are computed at every changed threshold score. With an imbalanced dataset where the negatives outnumber the positives, the ROC curve may be deceptive due to a flattening of FPRs. Under this situation, the PR curve is recommended as an additional measure to the ROC curve (Saito and Rehmsmeier, 2015). A PR curve is a plot between *precision* and *recall,* which can be computed following Equation (9). To quantify the performance measures of the ROC and PR curves, the area under the ROC curve (AUROC) and the area under the PR curve (AUPRC) were estimated from the plots. The higher the AUROC and AUPRC values, the better the model. In addition to those values, we also computed *accuracy* and F_1 following Equations (10) and (11). To give a binary class for an association based on its similarity scores, we specified an optimal threshold score based on the Youden's index, a point where awards the maximum value of the difference between the FPRs and the TPRs in an ROC curve (Youden, 1950).

$$Precision = \frac{TP}{(TP+FP)}, Recall = \frac{TP}{(TP+FN)} \tag{9}$$

$$Accuracy = \frac{TP+TN}{(TP+FP+FN+TN)} \tag{10}$$

$$F_1 = \frac{2 \times Precision \times Recall}{(Precision + Recall)} \tag{11}$$

3.2 Preliminary analysis of the data

In our data set, there were a total of 904 drugs and 524 diseases. The 6,782 unique proteins interacted with those drugs or diseases. The 8,301 GO functions of any aspects were associated with the drugs or diseases. Within these GO functions, there are 901 CC terms (10.9%), 2,407 MF terms (29.0%), and 4,993 BP terms (60.1%). We considered drug-GO and disease-GO associations of all GO aspects because GO functions of any aspect can contribute functional information about drugs and diseases from different viewpoints. These would be of great advantages in discovering relationships among the drugs and diseases. The total numbers and some statistics of the drug-protein, disease-protein, drug-GO, and disease-GO associations were summarized, as shown in Table 1. Since the number of all GO functions was greater than that of all proteins, the total number of the drug-GO associations (52,038) was approximately six times greater than the total number of the interactions between drugs and proteins (9,427). Similarly, the total number of the disease-GO associations (91,998) was about three times larger than the total number of the disease-protein associations (32,659). Since a protein can be associated with more than one GO function and one GO aspect, the numbers of the relations based on GO functions were larger than those of proteins.

Table 1. Total numbers and statistics of drug-protein, disease-protein, drug-GO, and disease-GO relations

Statistical information	Types of drug relations		Types of disease relations	
	Drug-protein	Drug-GO	Disease-protein	Disease-GO
Total number (relations)	9,427	52,038	32,659	91,998
Mean (proteins or GO functions)	10.4	57.6	62.3	175.6
Standard deviation (proteins or GO functions)	13.1	51.4	162.7	217.6
Minimum (proteins or GO functions)	1.0	2.0	1.0	1.0
Maximum (proteins or GO functions)	188.0	545.0	1,086.0	944.0

We also investigated the number of proteins that interact with a drug or a disease and the number of GO functions associated with a drug or a disease (Table 1). The number of proteins and GO functions associated with a drug range from 1 to 188 proteins and from 2 to 545 GO functions, respectively. Also, the wide ranges of the number of proteins (1 to 1,086) and GO functions (1 to 994) were found for both proteins and GO functions associated with a disease. On average, a drug was normally associated with a mean of 10.4 proteins, with a standard deviation (SD) of 13.1 proteins, whereas a drug was associated with a larger number of GO functions, a mean of 57.6 GO functions, with an SD of 51.4 GO functions. Similarly, a disease was associated with a higher average number of GO functions (175.6 ± 217.6) than proteins (62.3 ± 162.7). With this more extensive information of GO functions relative to proteins, we suggested that higher numbers of drug-disease, drug-drug, and disease-disease associations can be detected using the functionality-based similarities.

Based on drug-disease association data, we can classify all drug-disease, drug-drug, and disease-disease pairs into the group of positive (known) and negative (unknown) samples, as shown in Table 2. Relative to the negative drug-disease pairs (467,552 pairs), we had a few positive drug-disease associations (6,144 pairs). This suggested that there was still room for discovering potential drug-disease associations. Out of 408,156 drug-drug pairs, 47,094 pairs (11.5%) shared some common diseases and were labeled as positive, whereas 361,062 pairs (88.5%) did not have any common diseases and were labeled as negative. Relative to all disease-disease pairs, the positive (sharing some common drugs) and negative (no shared drugs) paired number 17,129 (12.5%) and 119,897 (87.5%), respectively.

Table 2. The numbers of drug-disease, drug-drug, and disease-disease pairs categorized into positive and negative classes

Types of pairs	Numbers of pairs in each class (%)		Total numbers of pairs
	Positive	Negative	
Drug-disease pairs	6,144 (1.3%)	467,552 (98.7%)	473,696
Drug-drug pairs	47,094 (11.5%)	361,062 (88.5%)	408,156
Disease-disease pairs	17,129 (12.5%)	119,897 (87.5%)	137,026

3.3 Selection of the most appropriate similarity index

To select the most suitable similarity index for this study, seven similarity indices were used to compute both protein- and functionality-based similarity scores for all drug-disease, drug-drug, and disease-disease pairs. Then, we directly classified drug-disease, drug-drug, and disease-disease associations based on those similarity scores. The AUROC values of each similarity index based on both proteins and GO functions are shown in Figure 4.

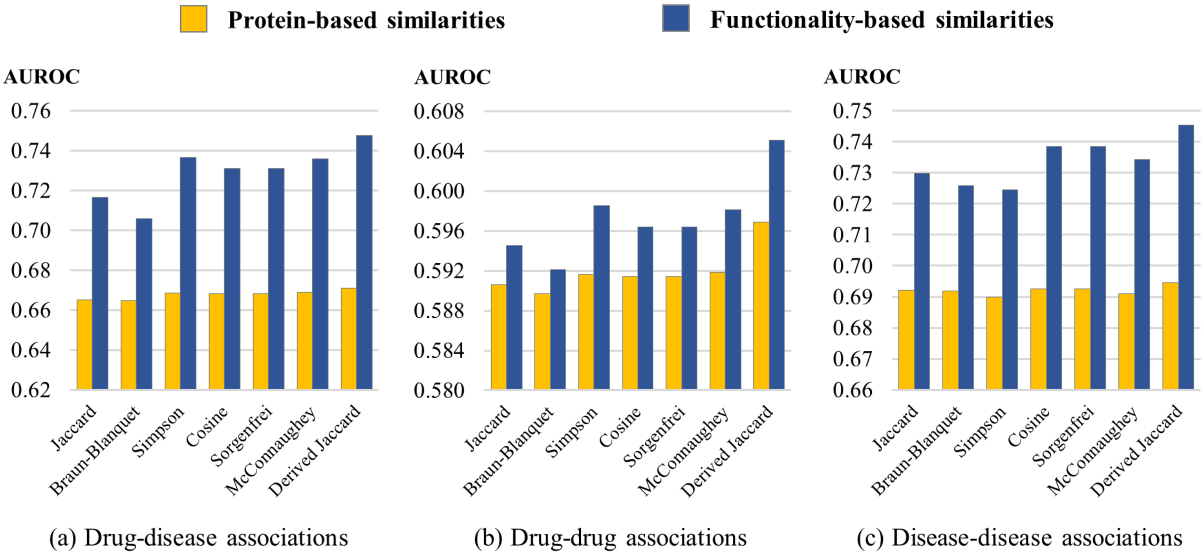


Figure 4. Area under the ROC curves (AUROC) of all similarity indices based on proteins and GO functions

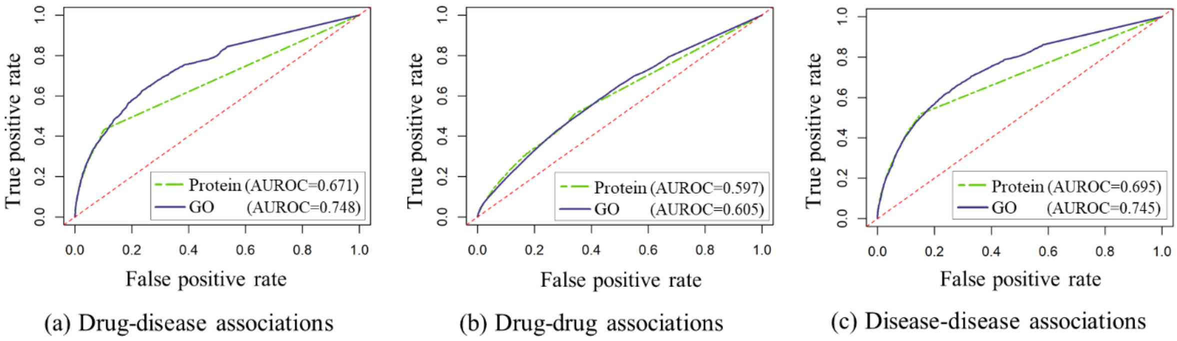
According to Figure 4, the classifications of drug-disease, drug-drug, and disease-disease associations based on functionality-based similarity scores can produce higher AUROC values than those of the classifications based on protein-based similarity scores. Despite a variety of similarity indices utilized, the AUROC values of protein-based similarities were slightly improved in all cases, especially in the cases of drug-disease and disease-disease associations. Moreover, the Cosine and Sorgenfrei similarity indices always give us the same AUROC values in all cases. For example, the AUROC values of both similarity indices computed based on GO functions were equally 0.731, 0.596, and 0.738 for drug-disease, drug-drug, and disease-disease associations, respectively. This is because they were correlated with each other, as can be seen in Equations (5) and (6). For both the protein- and functionality-based similarities, we can achieve the highest AUROC values in all association types by using the derived Jaccard similarity index. By applying a logarithmic transformation, as shown in Equation (8), the derived Jaccard similarity index was less correlated with its original one (Consonni and Todeschini, 2012). Consequently, this may suggest why the derived Jaccard similarity index was more appropriate for measuring the protein- and functionality-based similarities.

3.4 Comparison of protein- and functionality-based similarities

To assess the predicting power of the functionality-based similarities, we compared the classifications of drug-disease, drug-drug, and disease-disease associations using functionality-based similarities with those using protein-based similarities. In this experiment, the protein- and functionality-based similarity scores of all drug-disease, drug-drug, and disease-disease pairs were computed based on the derived Jaccard similarity index because it performed best among all similarity indices. Initially, the overall performance of protein- and functionality-based similarity scores was evaluated with respect to several threshold scores by using the ROC and the precision-recall (PR) curves. Based on the optimal threshold scores specified by the Youden's index, we additionally created the confusion matrices and computed the values of some standard performance measures, including *precision*, *recall*, *accuracy*, and F_1, to demonstrate that functionality-based similarities provide improved classifications of the different associations.

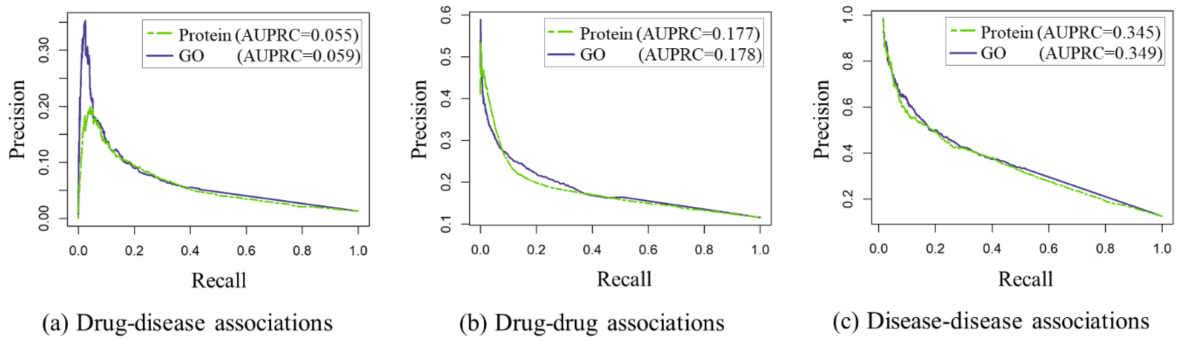
3.4.1 ROC and PR curves

The ROC curves and their corresponding AUROC values are shown in Figure 5. The classifications based on both protein- and functionality-based similarity scores can improve the performance of the completely random classifier, shown as the red-dashed straight lines. In all cases, utilizing the functionality-based similarity scores to classify drug-disease associations (AUROC = 0.748), drug-drug associations (AUROC = 0.605), and disease-disease associations (AUROC = 0.745) can produce higher AUROC values when compared to the classifications using the protein-based similarities. In addition to the ROC curves, the PR curves and computed the area under the PR curves (AUPRC) are also plotted, as shown in Figure 6. By comparing the AUPRC values, we found that the classifications of drug-disease associations (AUPRC = 0.059), drug-drug associations (AUPRC = 0.178), and disease-disease associations (AUPRC = 0.349) using functionality-based similarity scores achieved better AUPRC values than those of the classifications using the protein-based similarity scores.



(a) Drug-disease associations (b) Drug-drug associations (c) Disease-disease associations

Figure 5. ROC plots of the derived Jaccard similarity index based on proteins and GO functions



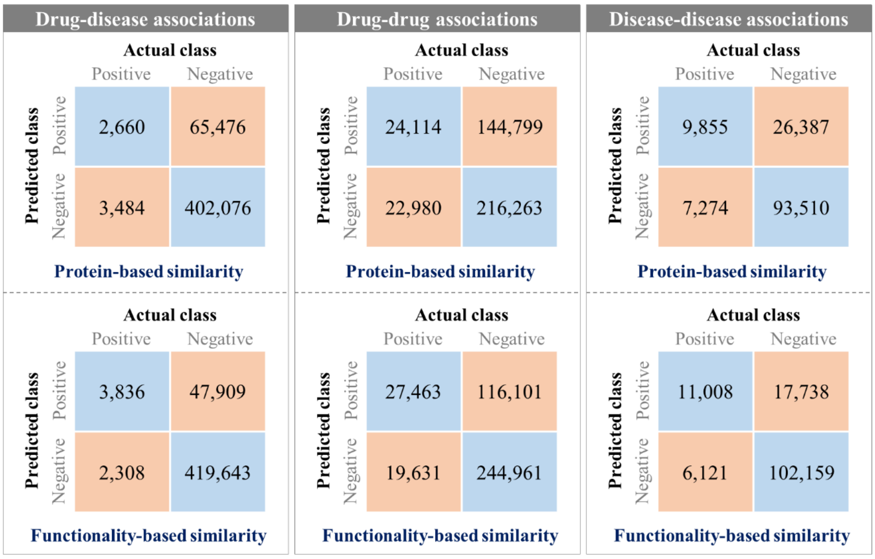
(a) Drug-disease associations (b) Drug-drug associations (c) Disease-disease associations

Figure 6. Precision-recall curves of the derived Jaccard similarity index based on proteins and GO functions

3.4.2 Confusion matrices and standard evaluation metrics

Based on the optimal threshold scores determined by the Youden's index, we created the confusion matrices of the classifications of drug-disease, drug-drug, and disease-disease associations, as shown in Figure 7. From all confusion matrices, it is noticeable that the classifications of all associations by using functionality-based similarities were improved. In Figure 7a, the number of accurate predictions using drug-disease similarity scores based on GO functions (*TP* = 3,836 and *TN* = 419,643) was significantly greater than those using the protein-based similarity scores (*TP* = 2,660 and *TN* = 402,076). Similarly, using drug-drug similarity scores based on GO functions (*TP* = 27,463 and *TN* = 244,961) increased the number of accurate predictions compared with those using the protein-based similarity scores (*TP* = 24,114 and *TN* = 216,263), as can be seen in Figure 7b. Furthermore, in the classification of disease-disease associations (Figure 7c), the number of accurate predictions was noticeably improved by using the functionality-based similarity scores (*TP* = 11,008 and *TN* = 102,159) compared with those using the protein-based similarity scores (*TP* = 9,855 and *TN* = 93,510).

According to the confusion matrices, the values of the evaluation metrics (*precision*, *recall*, *accuracy*, and F_1) can be computed to measure the performance of the protein- and functionality-based similarities, as shown in Tables 3 - 5. The values of all evaluation metrics were noticeably improved when we used the functionality-based similarity scores in the association classification. Especially, the *accuracy* values of the classifications of drug-disease, drug-drug, and disease-disease associations using the protein-based similarities were improved from 0.854 to 0.894, from 0.589 to 0.668, and from 0.754 to 0.826, respectively, by using the functionality-based similarities. These results are in accordance with the results of the confusion matrices (Figure 7), which show that using the functionality-based similarities increased the number of accurate predictions in the association classifications. Similarly, the *recall* values of the classifications of drug-disease, drug-drug, and disease-disease associations were improved from 0.433 to 0.624, from 0.512 to 0.583, and from 0.575 to 0.643, respectively, by using the functionality-based similarities. The *precision* values of the association classifications were quite low in all cases due to the highly imbalanced classes between positives and negatives in the data, resulting in a much higher number of false positives detected relative to true positives. Due to the low values of *precision*, the values of F_1 were also low in all cases. However, whatever the metrics considered may be, it is noticeable that using the functionality-based similarities in the classifications of drug-disease, drug-drug, and disease-disease associations outperformed the classifications using the protein-based similarities.



Drug-disease associations

		Actual class	
		Positive	Negative
Predicted class	Positive	2,660	65,476
	Negative	3,484	402,076

Protein-based similarity

		Actual class	
		Positive	Negative
Predicted class	Positive	3,836	47,909
	Negative	2,308	419,643

Functionality-based similarity

(a) Drug-disease associations

Drug-drug associations

		Actual class	
		Positive	Negative
Predicted class	Positive	24,114	144,799
	Negative	22,980	216,263

Protein-based similarity

		Actual class	
		Positive	Negative
Predicted class	Positive	27,463	116,101
	Negative	19,631	244,961

Functionality-based similarity

(b) Drug-drug associations

Disease-disease associations

		Actual class	
		Positive	Negative
Predicted class	Positive	9,855	26,387
	Negative	7,274	93,510

Protein-based similarity

		Actual class	
		Positive	Negative
Predicted class	Positive	11,008	17,738
	Negative	6,121	102,159

Functionality-based similarity

(c) Disease-disease associations

Figure 7. Confusion matrices of the association classification using the protein- and functionality-based similarities

Table 3. Values of evaluation metrics for the classification of drug-disease associations

Evaluation metrics	Drug-disease associations	
	Protein-based similarity	Functionality-based similarity
Precision	0.039	0.074
Recall	0.433	0.624
Accuracy	0.854	0.894
F_1	0.072	0.133

Table 4. Values of evaluation metrics for the classification of drug-drug associations

Evaluation metrics	Drug-drug associations	
	Protein-based similarity	Functionality-based similarity
Precision	0.143	0.191
Recall	0.512	0.583
Accuracy	0.589	0.668
F_1	0.223	0.288

Table 5. Values of evaluation metrics for the classification of disease-disease associations

Evaluation metrics	Disease-disease associations	
	Protein-based similarity	Functionality-based similarity
Precision	0.272	0.383
Recall	0.575	0.643
Accuracy	0.754	0.826
F_1	0.369	0.480

3.5 Case studies of the inferred associations

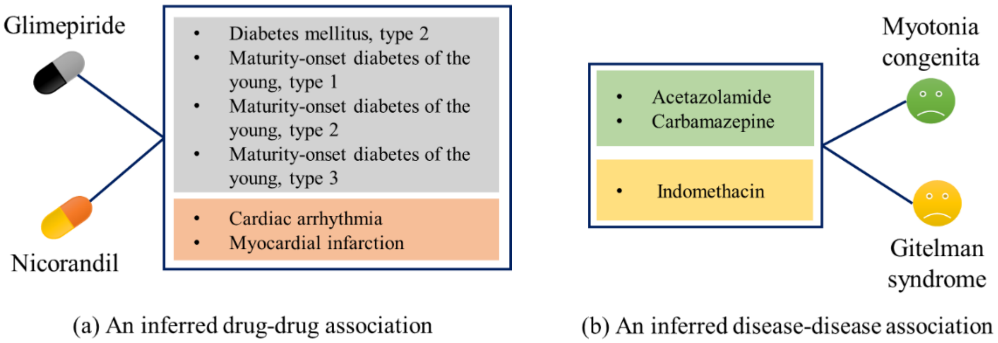
In this section, the practicality of using the functionality-based similarities for predicting drug-disease, drug-drug, and disease-disease associations were demonstrated. The optimal threshold similarity scores specified by the Youden's index were 0.264, 0.369, and 0.297 for drug-disease, drug-drug, and disease-disease associations, respectively. Each of three case studies was selected from the negative-labeled drug-disease pairs, drug-drug pairs, and disease-disease pairs having the functionality-based similarity scores greater than the given threshold scores. After that, these three inferred associations were validated by finding supporting evidence from the published literature and a database of clinical studies (ClinicalTrails.gov).

3.5.1 Tolcapone and attention deficit-hyperactivity disorder (ADHD)

Tolcapone (DB00323) is a drug used to treat Parkinson's disease (PD) by inhibiting the enzyme catechol-O-methyl transferase (COMT) (Antonini et al., 2008). ADHD (OMIM: 143465) is a mental health disorder that is characterized by several abnormal behaviors such as inattention, hyperactivity, and high impulsivity (Kuntsi et al., 2014). Recently, we found a clinical study in ClinicalTrials.gov (NCT03904498) that had the aim to assess the therapeutic effects of tolcapone in participants who have both ADHD and alcohol use disorder. The functionality-based similarity score between tolcapone and ADHD was 0.484. There were 9 GO functions shared between tolcapone and ADHD, mainly related to neural processes such as dopamine catabolic process (GO: 0042420), catechol-O-methyltransferase activity (GO: 00162606), and short-term memory (GO: 0007614). Until now, the mechanism of tolcapone is still unclear, but it is believed that the ability to inhibit COMT can sustain the dopaminergic system, resulting in relieving PD (Bonifácio et al., 2007). Moreover, the COMT gene's product can degrade dopamine mainly within the prefrontal cortex which mediates several executive behaviors related to ADHD (Sun et al., 2014).

3.5.2 Glimepiride and nicorandil

Glimepiride (DB00222) and nicorandil (DB09220) are two drugs having the functionality-based similarity score as 0.488 and sharing four GO functions, such as potassium ion import across plasma membrane (GO: 1990573), ion channel binding (GO: 0044325), and sulfonylurea receptor activity (GO: 0008281). Because of their involvement in similar GO functions, these drugs may be able to treat the same diseases. According to Figure 8a, glimepiride is a drug used for the treatment of type 2 diabetes mellitus or T2D (OMIM: 125853), maturity-onset diabetes of the young, type 1 (OMIM: 125850), type 2 (OMIM: 125851), and type 3 (OMIM: 600496). Nicorandil is used to treat cardiac arrhythmia (OMIM: 115000) and myocardial infarction (OMIM: 608446). Interestingly, we found a clinical study in ClinicalTrials.gov (NCT03775902) that was performed to investigate the effect of nicorandil in diabetic patients. Furthermore, there is a study showing that the abnormality of insulin handling in T2D patients is related to the dysfunction of the ATP-dependent potassium channel activity (Bonfanti et al., 2015). In addition, defective Ca^{2+} handling, which is mediated by several ion channel activities, can impact β-cell function in T2D (Jacobson and Shyng, 2020). Based on those supporting studies, it is suggested that the shared GO functions between these drugs may hint at the mechanisms of nicorandil in the treatment of T2D.


Glimepiride
Nicorandil
Diabetes mellitus, type 2
Maturity-onset diabetes of the young, type 1
Maturity-onset diabetes of the young, type 2
Maturity-onset diabetes of the young, type 3
Cardiac arrhythmia
Myocardial infarction
Myotonia congenita
Acetazolamide
Carbamazepine
Indomethacin
Gitelman syndrome

(a) An inferred drug-drug association

(b) An inferred disease-disease association

Figure 8. Inference of the shared diseases from the drug-drug association and the shared drugs from the disease-disease association

3.5.3 Myotonia congenita and Gitelman syndrome

Myotonia congenita (OMIM: 255700) is a disease with abnormality of skeletal muscles treated by acetazolamide (DB00819) and carbamazepine (DB00564), as shown in Figure 8b. Gitelman syndrome (OMIM: 263800) is a rare disease affecting the balance of several ions in the body such as magnesium, calcium, and potassium (Cruz et al., 2001), and can be treated by indomethacin (DB00328). Myotonia congenita and Gitelman syndrome have a functionality-based similarity score equal to 0.667 and share three GO functions between them, which are chloride transmembrane transport (GO: 1902476), voltage-gated chloride channel activity (GO: 0005247), and chloride channel complex (GO: 0034707). With the highly functional correlation between these diseases, myotonia congenita and Gitelman syndrome may share some common treatments. Although there is no clinical study in ClinicalTrials.gov that can support this theory, there are some studies suggesting the relationship between carbamazepine and Giltelman syndrome. The studies revealed that carbamazepine can affect the sodium transport in the toad *Pleurodema thaul* (Suwalsky et al., 2006), and that dysfunction of sodium chloride cotransporters can cause Giltelman syndrome (Graziani et al., 2010).

4. DISCUSSION

In this study, we comprehensively explore how to exploit functionality-based similarities to identify potential drug-disease, drug-drug, and disease-disease associations for drug repositioning. Drug-disease, drug-drug, and disease-disease similarity scores based on GO functions were used to predict drug-disease associations, drug-drug associations (sharing some common diseases), and disease-disease associations (sharing some common drugs), respectively. To assess the predicting power of functionality-based similarities, the performance of predictions based on protein-based similarities served as the baseline and were compared with the performance of functionality-based similarities. Both protein- and functionality-based similarities were computed based on seven commonly used similarity indices, which include the Jaccard, Braun-Blanquet, Simpson, Cosine, Sorgenfrei, McConnaughey, and derived Jaccard index, to find the most appropriate similarity index for protein and GO information used in this study. It was found that the derived Jaccard index performs better than the others for computing both protein- and functionality-based similarities. This result is consistent with the study of Wijaya and co-workers, who showed that the derived Jaccard index produced the highest AUROC value in classifying the efficacy matching of Indonesian and Japanese herbal medicines (Wijaya et al., 2016).

In the classifications of drug-disease, drug-drug, and disease-disease associations, functionality-based similarities based on the derived Jaccard similarity index significantly improved the classifications based on protein-based similarities. With the enlarging scope of drug and disease information using GO functions, the relationships of drugs and diseases were more efficiently detected than by using drug- and disease-associated proteins. This finding is supported by the study of Davis and co-workers, where two diseases having some common drugs significantly share their BP GO terms but rarely share their associated genes due to broader biological concepts compared (Davis et al., 2016). Furthermore, by the investigation of drug-protein-disease relationships on a protein-protein interaction (PPI) network, it has been revealed that disease-associated proteins interact with drug target proteins through more complex interaction than one-step direct interaction on the PPI network (Rutherford et al., 2018). This implied that not only drug target proteins and disease-associated proteins but also the downstream proteins affected by drugs and diseases are important for connecting drugs to diseases. Nevertheless, it is unlikely that all drug- and disease-affected proteins for computing the protein-based similarities can be identified. Therefore, the utilization of more extensive information about drugs and diseases, i.e., drug- and disease-associated GO functions, can overcome the limitation of the protein-based similarities and improve the classifications of drug-disease, drug-drug, and disease-disease associations.

Despite the advantages of functionality-based similarities were shown, there are some limitations of these similarities, which require improvement. Potential drug-drug and disease-disease associations, identified by the drug-drug and disease-disease functionality-based similarity scores, cannot directly point to the corresponding drug-disease associations for drug repositioning. Only a set of possible diseases shared between two drugs or a set of possible drugs shared between two diseases can be inferred from a drug-drug and disease-disease association, respectively. To overcome this limitation, more complex methods that can systematically integrate all these independent measures of the functional similarities are required to obtain more accurate and reliable predictions of the drug-disease associations.

5. CONCLUSION

The utilization of functionality-based similarities to identify drug-disease, drug-drug, and disease-disease associations for drug repositioning was explored. Herein, the derived Jaccard similarity index is recommended for computing functionality-based similarity scores due to its better performance than other similarity indices. By using the functionality-based similarity scores, the performance of the association classification based on the protein-based similarity scores is significantly improved. The case studies guarantee that the functionality-based similarity scores can be used to identify some potential drug-disease associations, which have been under investigation. In addition, the shared GO functions of the potential associations could provide a guide to the underlying mechanisms related to drugs and diseases.

ACKNOWLEDGMENT

Thitipong Kawichai was supported by the Development and Promotion of Science and Technology Talents Project (DPST). Apichat Suratanee was supported by the Office of the Higher Education Commission (OHEC) and Thailand Research Fund (TRF) (Grant No. MRG6180021). The authors would like to acknowledge the National e-Science Infrastructure Consortium (http://www.e-science.in.th) for allowing the use of their facility to produce the results of this study. All authors read and approved the final manuscript and declared that no competing interests exist.

REFERENCES

Antonini, A., Abbruzzese, G., Barone, P., Bonuccelli, U., Lopiano, L., Onofrj, M., Zappia, M., and Quattrone, A. (2008). COMT inhibition with tolcapone in the treatment algorithm of patients with Parkinson's disease (PD): Relevance for motor and non-motor features. *Neuropsychiatric Disease and Treatment*, 4(1), 1-9.

Bonfanti, D. H., Alcazar, L. P., Arakaki, P. A., Martins, L. T., Agustini, B. C., de Moraes Rego, F. G., and Frigeri, H. R. (2015). ATP-dependent potassium channels and type 2 diabetes mellitus. *Clinical Biochemistry*, 48(7-8), 476-482.

Bonifácio, M. J., Palma, P. N., Almeida, L., and Soares-da-Silva, P. (2007). Catechol-O-methyltransferase and its inhibitors in Parkinson's disease. *CNS Drug Reviews*, 13(3), 352-379.

Consonni, V., and Todeschini, R. (2012). New similarity coefficients for binary data. *MATCH Communications in Mathematical and in Computer Chemistry*, 68(2012), 581-592.

Cruz, D. N., Simon, D. B., Nelson-Williams, C., Farhi, A., Finberg, K., Burleson, L., Gill, J. R., and Lifton, R. P. (2001). Mutations in the Na-Cl cotransporter reduce blood pressure in humans. *Hypertension*, 37(6), 1458-1464.

Davis, A. P., Grondin, C. J., Johnson, R. J., Sciaky, D., McMorran, R., Wiegers, J., Wiegers, T. C., and Mattingly, C. J. (2019). The comparative toxicogenomics database: Update 2019. *Nucleic Acids Research*, 47, D948-D954.

Davis, A. P., Murphy, C. G., Saraceni-Richards, C. A., Rosenstein, M. C., Wiegers, T. C., and Mattingly, C. J. (2008). Comparative toxicogenomics database: A knowledgebase and discovery tool for chemical–gene–disease networks. *Nucleic Acids Research*, 37, D786-D792.

Davis, A. P., Wiegers, T. C., King, B. L., Wiegers, J., Grondin, C. J., Sciaky, D., Johnson, R. J., and Mattingly, C. J. (2016). Generating gene ontology-disease inferences to explore mechanisms of human disease at the comparative toxicogenomics database. *PLoS ONE*, 11(5), e0155530.

Gottlieb, A., Stein, G. Y., Ruppin, E., and Sharan, R. (2011). PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Molecular Systems Biology*, 7(1), 496.

Graziani, G., Fedeli, C., Moroni, L., Cosmai, L., Badalamenti, S., and Ponticelli, C. (2010). Gitelman syndrome: pathophysiological and clinical aspects. *QJM: An International Journal of Medicine*, 103(10), 741-748.

Hill, D. P., Smith, B., McAndrews-Hill, M. S., and Blake, J. A. (2008). Gene ontology annotations: what they mean and where they come from. *BMC Bioinformatics*, 9(Suppl 5), S2.

Huang, H., Nguyen, T., Ibrahim, S., Shantharam, S., Yue, Z., and Chen, J. Y. (2015). DMAP: a connectivity map database to enable identification of novel drug repositioning candidates. *BMC Bioinformatics*, 16(Suppl 13), S4.

Huntley, R. P., Sawford, T., Mutowo-Meullenet, P., Shypitsyna, A., Bonilla, C., Martin, M. J., and O'Donovan, C. (2015). The GOA database: Gene ontology annotation updates for 2015. *Nucleic Acids Research*, 43, D1057-D1063.

Jacobson, D. A., and Shyng, S. L. (2020). Ion channels of the islets in type 2 diabetes. *Journal of Molecular Biology*, 432(5), 1326-1346.

Kuntsi, J., Pinto, R., Price, T. S., van der Meere, J. J., Frazier-Wood, A. C., and Asherson, P. (2014). The separation of ADHD inattention and hyperactivity-impulsivity symptoms: Pathways from genetic effects to cognitive impairments and symptoms. *Journal of Abnormal Child Psychology*, 42(1), 127-136.

Lewis, S. N., Nsoesie, E., Weeks, C., Qiao, D., and Zhang, L. (2011). Prediction of disease and phenotype associations from genome-wide association studies. *PLoS ONE*, 6(11), e27175.

Nelson, B. S., Kremer, D. M., and Lyssiotis, C. A. (2018). New tricks for an old drug. *Nature Chemical Biology*, 14(11), 990-991.

Piñero, J., Bravo, À., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., García-García, J., Sanz, F., and Furlong, L. I. (2016). DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research*, 45, D833-D839.

Rutherford, K. D., Mazandu, G. K., and Mulder, N. J. (2018). A systems-level analysis of drug-target-disease associations for drug repositioning. *Briefings in Functional Genomics*, 17(1), 34-41.

Saito, T., and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, 10(3), e0118432.

Sun, H., Yuan, F., Shen, X., Xiong, G., and Wu, J. (2014). Role of COMT in ADHD: a systematic meta-analysis. *Molecular Neurobiology*, 49(1), 251-261.

Suwalsky, M., Mennickent, S., Norris, B., and Cardenas, H. (2006). The antiepileptic drug carbamazepine affects sodium transport in toad epithelium. *Toxicology in Vitro*, 20(6), 891-898.

Wijaya, S. H., Afendi, F. M., Batubara, I., Darusman, L. K., Altaf-Ul-Amin, M., and Kanaya, S. (2016). Finding an appropriate equation to measure similarity between binary vectors: Case studies on Indonesian and Japanese herbal medicines. *BMC Bioinformatics*, 17(1), 520.

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., Assempour, N., Iynkkaran, I., Liu, Y., Maciejewski, A., Gale, N., Wilson, A., Chin, L., Cummings, R., Le, D., Pon, A., Knox, C., and Wilson, M. (2018). DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Research*, 46, D1074-D1082.

Yella, K. J., Yaddanapudi, S., Wang, Y., and Jegga, G. A. (2018). Changing trends in computational drug repositioning. *Pharmaceuticals*, 11(2), 57.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, 3(1), 32-35.