ISSN: 2321-7782 (Online)
Impact Factor: 6.047

Volume 4, Issue 9, September 2016

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

Similarity Measures for Drug Elucidation

P. N. Varalakshmi K[1]
Asst Professor, Department of CSE
Tirumala College of Engineering
Jonnalagadda,Narasaraopeta
Guntur(Dt), Andhra Pradesh – India

D. Komali[2]
Asst Professor, Department of CSE
Tirumala College of Engineering
Jonnalagadda,Narasaraopeta
Guntur(Dt), Andhra Pradesh – India

Abstract: ***Understanding drugs and their modes of action is a fundamental challenge in systems medicine. Key to addressing this challenge is the elucidation of drug targets, an important step in the search for new drugs or novel targets for existing drugs. Incorporating multiple biological information sources is of essence for improving the accuracy of drug target prediction. In this article, we introduce a novel framework—Similarity-based Inference of drug-TARgets (SITAR)—for incorporating multiple drug-drug and gene-gene similarity measures for drug target prediction. The framework consists of a new scoring scheme for drug-gene associations based on a given pair of drug-drug and gene-gene similarity measures, combined with a logistic regression component that integrates the scores of multiple measures to yield the final association score. We apply our framework to predict targets for hundreds of drugs using both commonly used and novel drug-drug and gene-gene similarity measures and compare our results to existing state of the art methods, markedly outperforming them.***

Keywords: ***computational molecular biology, gene expression, gene networks, genetic variation, machine learning, sequence analysis.***

I. INTRODUCTION

Deciphering drug targets is a primary task in the development of new drugs, in finding new ways to utilize existing drugs, and in pinpointing their side effects. Experimental identification of drug-target associations remains a laborious and costly task (Haggarty et al., 2003), calling for faster computational prediction methods. Such methods can be used to augment the limited available information on drug targets, which is in sharp contrast to the vast number of compounds existing in chemical databases. Early attempts in computational prediction included docking simulations (Cheng et al., 2007) and text mining (Zhu et al., 2005). The former, however, can be applied only to targets with known three dimensional (3D) structure. The latter searches for co-occurrences of drugs and genes in texts, and is limited to current knowledge and prone to detection problems due to multiple gene and compound names. Additional attempts were based on reverse engineering of gene regulatory networks, inferring possible targets from cellular responses to administered drugs (di Bernardo et al., 2005; Gardner et al., 2003; Mani et al., 2008). These methods suffer from the complex and noisy nature of molecular networks. Recently, several algorithms have been proposed to predict drug-target associations by combining drug-drug and gene-gene similarity measures (Bleakley and Yamanishi, 2009; Campillos et al., 2008; Keiser et al., 2009; Yamanishi et al., 2008). The key assumption underlying these algorithms is that similar drugs tend to share similar targets (Mitchell, 2001). This has been observed with respect to chemical similarity (Martin et al., 2002; Schuffenhauer et al., 2003), side effect similarity (Campillos et al., 2008), and more. Several authors had previously predicted drug-target interactions by combining chemical drug-drug similarity and sequence-based gene-gene similarity (Bleakley and Yamanishi, 2009; Yamanishi et al., 2008). Keiser et al. (2009) compared the chemical structure of drugs to a compendium of ligands, known to modulate the function of protein receptors, providing indirect connections between drugs and targets via these ligands. Several approaches, concentrating mainly on indirect drug-gene associations, employed additional similarity measures to gain insights on drugs. Specifically, protein-protein interaction (PPI) network similarity was

used in Hansen et al. (2009) to predict drug-gene genetic interactions (termed ‘‘pharmacogenes’’), and gene expression data was combined with drug-response data in Kutalik et al. (2008) to infer co-modules of genes and drugs. Last, a recent approach used information on compound induced fitness defects of yeast deletion strains to predict drug-targets in S.cerevisiae (Hillenmeyer et al., 2010). To overcome these limitations, we have designed a new prediction scheme—Similarity-based Inference of drug-TARgets (SITAR)—that integrates multiple measures to facilitate the prediction task. Our contribution is twofold: (i) We introduce novel drug-drug similarity measures and combine them into the prediction process; and (ii) we propose a way of integrating the drug-drug and gene-gene similarities to create classification features. The result is a new drug-target prediction algorithm, which markedly outperforms previous methods and can cope with new drugs with no known targets.

II. Results and Discussion

2.1. SITAR: an algorithm for predicting drug targets

We designed a drug-target prediction algorithm with three main components

i. drug-drug and gene-gene similarity computations;
ii. combining the drug and gene similarity measures into classification features;
iii. feature selection and prediction using logistic regression. In the following, we describe these components in detail.

2.2. Similarity measures:

In order to overcome the limitations engulfed in using similarity measures of a single type, we set out to incorporate a multitude of similarity measures, including both novel and already published ones. Overall, we considered five drug-drug similarities and three gene-gene similarities from different biological and chemical sources. The drug-drug similarity measures were computed using chemical, registered and predicted drug side effects (Kuhn et al., 2010) of the drug, drug response gene expression profiles, and the Anatomical, Therapeutic and Chemical (ATC) classification system. The gene-gene similarity measures used are based on sequence, closeness in a protein-protein interaction network, and semantic Gene Expressions.

2.3. Feature construction and classification

At the heart of our algorithm lies the process of exhaustive construction of classification features that span the entire pair wise space of drug-target measures’ combinations. That is, each feature is constructed based on one drug-drug similarity measure and one gene-gene similarity measure. It is calculated by combining the drug-drug similarities between the query drug and other drugs and the gene-gene similarities between the query gene and other target genes across all true drug-target associations. The features are automatically combined using a logistic regression classifier that is coupled with a wrapper feature selection procedure and yield the final classification scores.

2.4. Feature selection and performance evaluation

We performed feature selection using both forward selection and backward elimination, converging to a selected set of ten features, constructed from pairs of drug-drug and gene-gene similarity measures. The area under the precision-recall curve (AUPR) scores before and after the feature selection phase, as well as the AUPR achieved when using each of the ten selected features are listed in Table 1. Similar results were obtained when using an SVM classifier (see Methods below, as well as Table S1 in the Supplementary Material, available at www.liebertonline.com/cmb). Examining the individual contribution of each of the

2.5. Comparison to other drug-target prediction methods

We compared our method to two state-of-the-art methods:

(i) The kernel regression-based method (KRM) of Yamanishi et al. (2008) embeds drugs and targets into a unified Euclidean space termed the ‘‘pharmacological space,’’ using a regression model. Predicted interacting drug-gene pairs are those that are closer to each other below a certain threshold in the pharmacological space.

(ii) The bipartite local models (BLMs) method of (Bleakley and Yamanishi, 2009) constructs local models to learn drug-target associations based on additional targets of the query drug and additional drugs targeting the query target. We note that the SEA tool of Keiser et al. (2007) provides receptors code names that cannot be mapped to our list of targets, precluding a direct comparison to their method. Figure 2 displays the precision-recall curves of the three methods, and Table 3 summarizes the AUPR and AUC scores between the different methods, overall demonstrating the marked improvement obtained by our new method (AUPR of 0.908, exceeding the KRM and BLM methods by 0.07 and 0.15 AUPR difference, respectively).

Table 1. Comparison of AUPR and AUC Scores Calculated Using Different Methods

Measure type	AUPR	AUC
This work	0.908	0.935
KRM (Yamanishi et al., 2008) 0.884	0.838	0.884
BLM (Bleakley and Yamanishi, 2009)	0.754	0.814

III. METHODS

3.1. Similarity measures

We defined and computed five drug-drug similarity measures and three gene-gene similarity measures. All similarity measures were normalized to be in the range [0, 1].

We used the following drug-drug similarity measures:

(1) Chemical-based: Canonical simplified molecular input line entry specification (SMILES) of the drug molecules were downloaded from DrugBank (Wishart et al., 2008). Hashed fingerprints were computed using the Chemical Development Kit (CDK) with default parameters (Steinbeck et al., 2006). The similarity score between two drugs is computed on their fingerprints according to the two-dimensional Tanimoto score (Tanimoto, 1957), which is equivalent to the Jaccard score (Jaccard, 1908) of their fingerprints, i.e., the size of the intersection over the union when viewing each fingerprint as specifying a set of elements.

(2) Ligand-based: The Similarity Ensemble Approach (SEA) (Keiser et al., 2007) relates protein receptors based on the chemical 2D similarity of the ligand-sets modulating their function. Given a drug’s canonical SMILES, the SEA search tool compares it against a compendium of ligand-sets and computes E-values for those ligand sets. To compute a drug-drug similarity, we queried drugs using their canonical SMILES on the SEA tool. To obtain robust results, we queried the drug against the two ligand databases provided in the tool (MDL Drug data report and WOMBAT [Olah et al., 2005]) and used two different methods to compute the drug fingerprint (Scitegic ECFP4 and Daylight), resulting in four lists of similar ligand sets. Unifying the four lists and filtering drug-ligand set pairs with E-values >10_5, we obtained a list of relevant protein-receptor families for each drug. Finally, the similarity between a pair of drugs was computed as the Jaccard score between the corresponding sets of receptor families. We note that, due to a partial mapping of the receptor families to proteins, we could not use the drug-receptor associations directly as classification features.

(3) Expression-based: Gene expression responses to drugs were retrieved from the Connectivity Map project (Lamb et al., 2006). We experimented with three different methods to calculate drug similarity from Connectivity Map ranked gene expression profiles:

(i) Spearman rank correlation coefficient

(ii) calculating a Jaccard score between the 500 most differentially expressed genes (250 Most up-regulated and 250 most down-regulated genes); and (iii) using the method proposed by (Iorio et al., 2009), employing Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005) as a similarity measure. We dealt with multiple experiments per drug as follows: In the Spearman calculation, we averaged over the d1xd2 different correlation coefficients obtained between the d1 experiments of one drug against the d2 repeated experiments of the second drug. In the Jaccard case, we used differentially expressed genes that appeared in at least 50% of the gene expression responses to the same drug. The method proposed by Iorio et al. (2009) handles repeated experiments of the same drug through iterative merging. (4) Side-effect based: Drug side effects were obtained from SIDER (Kuhn et al., 2010), an online database containing drug side effects associations extracted from package inserts using text mining methods. Recently, we developed an algorithmic framework to predict side effects for drugs by combining side effect information on known drugs with their chemical properties (Atias and Sharan, 2010). Following this work, we defined the similarity between drugs according to the Jaccard score between their top ten predicted side effects.

(5) Annotation-based: We used the World Health Organization (WHO) ATC classification system (Skrbo et al., 2004). This hierarchical classification system categorizes drugs according to the organ or system on which they act, their therapeutic and their chemical characteristics. ATC codes were obtained from Drug Bank. To define a similarity between ATC terms, we used the semantic similarity algorithm of Resnik (1999). This algorithm associates probabilities p(x) with all the nodes (i.e., terms) x in the hierarchy and calculates the similarity of two drugs as the maximum over all their common ancestors c of –log (p(c)).

The gene-gene similarity measures we used include:

(1) Sequence similarity: based on a Smith-Waterman sequence alignment score (Smith et al., 1985). Following the normalization suggested in Bleakley and Yamanishi (2009), we divide the Smith-Waterman score between two protein sequences by the geometric mean of the scores obtained from aligning each sequence against itself.

(2) Closeness in a protein-protein interaction (PPI) network: Human protein-protein interactions were compiled from the literature (Breitkreutz et al., 2008; Ewing et al., 2007; Rual et al., 2005; Stelzl et al., 2005; Xenarios et al., 2002). The distances between each pair of genes were calculated on their corresponding proteins using an all-pairs shortest paths algorithm.

IV. Conclusion

We introduced a novel method, SITAR, for predicting drug-target interactions. Our method incorporates an extensive set of drug-drug and gene-gene similarity measures. Newly incorporated drug-drug similarities are based on predicted side effects, gene expression drug response profiles, and the ATC classification system. The classification features are constructed based on a new score integrating the drug-drug and gene-gene similarity spaces. These features are integrated via a logistic regression classifier, combined with a feature selection process. Our method is flexible and allows the incorporation of new emerging measures without altering already computed scores on other measures. Using our method, we show marked improvement of classification performance over previous drug-target prediction approaches. We provide novel predictions of drug-target interactions and validate them against public databases. Last, we predict targets for drugs which to-date have no known targets. Having shown that our method is robust with respect to different score choices, selected features, and different classification methods, it seems that the primary reason for the increased performance compared to previous methods stems from the use of multiple similarity measures. Each of the resulting features alone does not have enough predictive power, but the combination of multiple features allows the classification procedure to perform well. Accordingly, we noticed that using a low number of features (less than five) deteriorates the results. Nevertheless, our method can be enhanced in several ways. First, one could improve and expand the measures used. Of special interest is improving the gene co-expression similarity based on the Connectivity Map data, which currently exhibits the worst performance. Another extension would be to increase the number of represented drugs and genes shared between the different measures. This could be achieved either by predicting missing

similarities from existing ones (Atias and Sharan, 2010) or by incorporating imputation methods to overcome missing information in some of the measures.

References

1. Accelrys, Inc. 2009. Available at: http://accelrys.com/products/scitegic/. Accessed November 1, 2010.
2. Ashburner, M., Ball, C.A., Blake, J.A., et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25, 25–29.
3. Atias, N., and Sharan, R. 2010. An algorithmic framework for predicting side-effects of drugs. RECOMB 2010 (to appear).
4. Bleakley, K., and Yamanishi, Y. 2009. Supervised prediction of drug-target interactions using bipartite local models. Bioinformatics 25, 2397–2403.
5. Breitkreutz, B.J., Stark, C., Reguly, T., et al. 2008. The BioGRID Interaction Database: 2008 update. Nucleic Acids Res. 36, D637-D640.
6. Burns, K., Opas, M., and Michalak, M. 1997. Calreticulin inhibits glucocorticoid- but not cAMP-sensitive expression of tyrosine aminotransferase gene in cultured McA-RH7777 hepatocytes. Mol. Cell Biochem. 171, 37–43.
7. Campillos, M., Kuhn, M., Gavin, A.C., et al. 2008. Drug target identification using side-effect similarity. Science 321, 263–266.
8. Chang, C.-C., and Lin, C.-J. 2001. LIBSVM: a library for support vector machines. Available at: http://www.csie.ntu .edu.tw/*cjlin/libsvm. Accessed November 1, 2010.
9. Chen, M., Ona, V.O., Li, M., et al. 2000. Minocycline inhibits caspase-1 and caspase-3 expression and delays mortality in a transgenic mouse model of Huntington disease. Nat. Med. 6, 797–801.
10. Cheng, A.C., Coleman, R.G., Smyth, K.T., et al. 2007. Structure-based maximal affinity model predicts small-molecule drug ability. Nat. Biotechnology. 25, 71–75.
11. Davis, A.P., Murphy, C.G., Saraceni-Richards, C.A., et al. 2009. Comparative Toxicogenomics Database: a knowledge base and discovery tool for chemical-gene-disease networks. Nucleic Acids Res. 37, D786-D792.
12. Davis, J., and Goadrich, M. 2006. The relationship between Precision-Recall and ROC curves. Proc. ICML '06 233– 240. di Bernardo, D., Thompson, M.J., Gardner, T.S., et al. 2005. Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. Nat. Biotechnol. 23, 377–383.
13. Ehnholm, C., Mahley, R.W., Chappell, D.A., et al. 1984. Role of apolipoprotein E in the lipolytic conversion of betavery low density lipoproteins to low density lipoproteins in type III hyperlipoproteinemia. Proc. Natl. Acad. Sci. USA 81, 5566–5570.

