Journal of Advances in Science and Technology

Vol. 10, Issue No. 21, February-2016, ISSN 2230-9659

AN ANALYSIS UPON LARGE-SCALE PREDICTION AND APPROACH OF DRUG-PROTEIN INTERACTION INFORMATION FOR COMPUTATIONAL DRUG REPURPOSING

www.ignited.in

Journal of Advances in Science and Technology
Vol. 10, Issue No. 21, February-2016, ISSN 2230-9659

An Analysis upon Large-Scale Prediction and Approach of Drug-Protein Interaction Information for Computational Drug Repurposing

Rajeev Kumar[1]* Dr. Alok Mishra[2]

[1]Research Scholar, SSSUTMS, Sehore

[2]UTD, SSSUTMS, Sehore

Abstract – Despite increased investment in pharmaceutical research and development, fewer and fewer new drugs are entering the marketplace. This has prompted studies in repurposing existing drugs for use against diseases with unmet medical needs. A popular approach is to develop a classification model based on drugs with and without a desired therapeutic effect. For this approach to be statistically sound, it requires a large number of drugs in both classes. However, given few or no approved drugs for the diseases of highest medical urgency and interest, different strategies need to be investigated.

We developed a computational method termed "drug-protein interaction-based repurposing" (DPIR) that is potentially applicable to diseases with very few approved drugs. The method, based on genome-wide drug-protein interaction information and Bayesian statistics, first identifies drug-protein interactions associated with a desired therapeutic effect. Then, it uses key drug-protein interactions to score other drugs for their potential to have the same therapeutic effect.

Detailed cross-validation studies using United States Food and Drug Administration-approved drugs for hypertension, human immunodeficiency virus, and malaria indicated that DPIR provides robust predictions. It achieves high levels of enrichment of drugs approved for a disease even with models developed based on a single drug known to treat the disease. Analysis of our model predictions also indicated that the method is potentially useful for understanding molecular mechanisms of drug action and for identifying protein targets that may potentiate the desired therapeutic effects of other drugs (combination therapies).

The emergence of large-scale genomic, chemical and pharmacological data provides new opportunities for drug discovery and repositioning. Systematic integration of these heterogeneous data not only serves as a promising tool for identifying new drug-target interactions (DTIs), which is an important step in drug development, but also provides a more complete understanding of the molecular mechanisms of drug action. In this work, we integrate diverse drug-related information, including drugs, proteins, diseases and side-e_ects., together with their interactions, associations or similarities, to construct a heterogeneous network with 12,015 nodes and 1,895,445 edges. We then develop a new computational pipeline, called DTINet, to predict novel drug-target interactions from the constructed heterogeneous network.

---------------------------♦-----------------------------

INTRODUCTION

Computational prediction of drug-target interactions (DTIs) has become an important step in the drug discovery or repositioning process, aiming to identify putative new drugs or novel targets for existing drugs. Compared to in vivo or biochemical experimental methods for identifying new DTIs, which can be extremely costly and time-consuming, in silico or computational approaches can efficiently identify potential DTI candidates for guiding in vivo validation, and thus significantly reduce the time and cost required for drug discovery or repositioning. Traditional computational methods mainly depend on two strategies, including the molecular docking-based approaches and the ligand-based approaches. However, the performance of molecular docking is limited when the 3D structures of target proteins are not available, while the ligand-based approaches often lead to poor prediction results when a target has only a small number of known binding ligands.

In the past decade, much effort has been devoted to developing the machine learning based approaches for computational DTI prediction. A key idea behind these methods is the \guilt-by-association" assumption, that is, similar drugs may share similar targets and vice versa. Based on this intuition, the DTI prediction problem is often formulated as a binary classification task, which aims to predict whether a drug-target interaction is present or not. A straightforward classification based approach is to consider known DTIs as labels and incorporate chemical structures of drugs and primary sequences of targets as input features (or kernels). Most existing prediction methods mainly focus on exploiting information from homogeneous networks. For example, Bleakley and Yamanishi applied a support vector machine (SVM) framework to predict DTIs based on a bipartite local model (BLM). Mei et al. extended this framework by combining BLM with a neighbor-based interaction-profile inferring (NII) procedure (called BLMNII), which is able to learn the DTI features from neighbors and predict interactions for new drug or target candidates. Xia et al. proposed a semi-supervised learning method for DTI prediction, called NetLapRLS, which applies Laplacian regularized least square and incorporates both similarity and interaction kernels into the prediction framework. van Laarhoven et al. introduced a Gaussian interaction profile (GIP) kernel based approach coupled with regularized least square (RLS) for DTI prediction. Rather than regarding a drug-target interaction as a binary indicator,Wang and Zeng proposed a restricted Boltzmann machine (RBM) model to predict different types of DTIs (e.g., activation and inhibition) on a multidimensional network.

In addition to chemical and genomic data, previous works have incorporated pharmacological or phenotypic information, such as side-effect, transcriptional response data, drug-disease associations, public gene expression data and functional data for DTI prediction. Heterogeneous data sources provide diverse information and a multi-view perspective for predicting novel DTIs. For instance, the therapeutic effects of drugs on diseases can generally reect their binding activities to the targets (proteins) that are related to these diseases and thus can also contribute to DTI prediction. Therefore, incorporating heterogeneous data sources, e.g., drug-disease associations, can potentially boost the accuracy of DTI prediction and provide new insights into drug repositioning. Despite the current availability of heterogeneous data, most existing methods for DTI prediction are limited to only homogeneous networks or a bipartite DTI models, and cannot be directly extended to take into account heterogeneous node or topological information and complex relations among different data sources.

Recently, several computational strategies have been introduced to integrate heterogeneous data sources to predict DTIs. A network-based approach for this purpose is to fuse heterogeneous information through a network diffusion process, and directly use the obtained diffusion distributions to derive the prediction scores of DTIs. A meta-path based approach has also been proposed to extract the semantic features of DTIs from heterogeneous networks. A collaborative matrix factorization has been developed to project the heterogeneous networks into a common feature space, which enables one to use the aforementioned homogeneous network based methods to predict new DTIs from the resulting single integrated network.

However, these approaches generally fail to provide satisfactory integration paradigms. First, directly using the diffusion states as the features or prediction scores may easily suffer from the bias induced by the noise and high-dimensionality of biological data and thus possibly lead to inaccurate DTI predictions. In addition, the hand-engineered features, such as meta-paths, often require expert knowledge and intensive effort in feature engineering, and hence prevent the prediction methods from being scaled to large-scale datasets. Moreover, collapsing multiple individual networks into a single network may cause substantial loss of network-specific information, since edges from multiple data sources are mixed without distinction in such an integrated network.

The increasing amount of publicly available chemical data creates opportunities for the analysis and integration of resources of molecular information at the interface between biology and chemistry. While large-scale data sets have long been publicly available in molecular biology, this spirit of openness began only recently to spread in chemistry. Funding bodies such as the National Institutes of Health (NIH) are fostering the creation of public databases, for example, PubChem as part of the NIH_s Molecular Libraries Roadmap Initiative. In addition, more research areas are being considered pre-competitive by the pharmaceutical industry. Consequently, we are witnessing an increasing number of public databases that store information about compounds along with properties and context.

The combined knowledge on individual drugs and targets can be advantageously integrated with new high-throughput data sets and concepts for systems-wide analysis of their relations, thus opening a new road to predict drug–target relationships and the effects of drugs on human biology. Until exhaustive screens have been performed that study the effect of all human drugs on all human proteins under various conditions, computational and systems biology approaches will be invaluable in extending our knowledge on drug–target relations systematically.

BACKGROUND

By conservative estimates, we know the molecular basis of more than 4,000 human diseases, whereas treatments are available for only about 250 of them. The modern drug discovery paradigm, i.e., starting with a disease target and looking for a highly selective small molecule that interacts strongly only with the intended target, is struggling to meet our medical and social requirements.

Current estimates indicate that it takes an average of 14 years at a cost of close to $2 billion to bring a new, safe, and efficacious drug to market. In the process, more than 90% of the drug candidates fail due to safety concerns, inadequate bioavailability, or lack of efficacy.

In reality, highly selective compounds are rare. The large number of safety and bioavailability issues facing candidate drugs, as well as reported side effects of marketed drugs, is a reflection of many undocumented and deleterious interactions between drugs and human targets. In addition, many underlying disease causes are multifactorial and can be due to dysfunctional processes involving multiple biomolecules. Even though significant efforts are devoted to understanding the molecular details of drug action, the fact is that the mechanisms of action of many efficacious drugs are poorly understood and, in many cases, remain largely unknown.

Drug interactions with unintended targets may lead to devastating side effects. However, these interactions may also signal the possibility for a drug to have therapeutic potential for diseases other than those for which it was approved. Indeed, there are many examples of a drug developed for a specific disease that was later approved for treating an unrelated disease. One of the most dramatic examples is thalidomide, a drug first marketed as an anti-nausea and sedative agent prescribed to treat morning sickness in pregnant women. Thalidomide was found to cause severe birth defects and was withdrawn from the market, but it later proved to have other therapeutic effects and was approved by the United States Food and Drug Administration (FDA) for skin lesions caused by leprosy and multiple myeloma.

In addition, thalidomide has shown promise in treating cutaneous lupus and Behcet's disease, human immunodeficiency virus (HIV)-related mouth and throat ulcers, and blood and bone marrow cancers. Although the recorded effects of thalidomide are multifaceted, with multiple underlying mechanisms possible, clarification of a mechanism that distinguishes between the teratogenic and anticancer therapeutic effects of thalidomide was only recently identified. The success of drug repurposing, i.e., finding new uses for existing drugs, via serendipitous discoveries inspired the development of many computational approaches for the discovery of the yet unknown therapeutic potentials of existing drugs.

A common approach used in drug repurposing is to build a binary classifier based on a training set consisting of drugs with and without a desired therapeutic effect as the positive and negative classes, respectively. One of the requirements for developing a statistically sound classifier is the availability of a relatively large number of drugs in the training set. Furthermore, it is desirable to have an equal number of drugs with and without the desired therapeutic effect in the training set so as not to bias classifier training. However, when there are many drugs approved for treating a disease, the need for discovering more drugs for the same disease is less than that for a disease for which there is a very limited number of drugs or no drugs at all. Thus, there is a need to develop computational drug repurposing methods that can be applied to diseases for which there are very few known pharmacological options.

CONCEPTS FOR LARGE-SCALE DRUG–TARGET PREDICTIONS

Predicting relations based on molecular features of chemicals and proteins -

Exploiting similarities between chemical structures is a common way to infer the activity of compounds. The most prevalent approach for comparing compounds is to convert the two-dimensional representation of each compound into a fingerprint either by using a defined list of substructures or by encoding (hashing) all the encountered substructures up to a certain size. This results in fixed-length bit vectors for which the Tanimoto (or Jacquard) similarity measure is computed by dividing the size of intersection of the set bits by the size of the union. Alternatively, chemical similarity can be determined by aligning three-dimensional models of the compounds. To illustrate these similarity measures, we show two- and three-dimensional structure comparisons of the monoamine oxidase inhibitor pargyline with five other compounds.

Initial optimistic results on the relationship between chemical similarity and activity were put into perspective by the analysis of more unbiased chemical libraries. For these, there is only a 30% chance of binding the same compound at the similarity level previously thought to warrant >80% chance. To overcome the limited predictive power of pairwise chemical structure comparison, Keiser and co-workers developed a statistical model to detect remote, yet significant similarities between groups of drugs and used it to predict novel drug–target relations. Other groups used Bayesian classifiers to correlate the presence or absence of chemical substructures with protein binding properties and reported high success rates for known interactions. More specialized chemical similarity methods have also been developed that take, for example, the similarity of target proteins into account.

Homology relations between proteins can be exploited to predict binding of drugs to proteins that are related to known drug targets. A study on crystal structures of alpha-helical proteins in the PDB showed that the chemical similarity between ligands is higher for proteins with similar sequences. Here, we generalize this to all proteins for which ligand binding constants are available from the PDSP Ki database. Using Ki = 10 lM as the threshold for what is considered ''binding'', we quantify the probability that two proteins bind the same ligand as a function of their sequence similarity separately for four classes of target proteins.

METHODOLOGY

Source of large-scale chemical-protein interaction Information -

To create large-scale chemical-protein interaction profiles for FDA-approved drugs and drug development candidates, we exploited the Search Tool for Interactions of Chemicals (STITCH) database. The October 2013 release of the database (STITCH 3.1) contains chemical-protein interaction information, derived from a broad range of sources, between 300,000 small molecules and 2.6 million proteins from 1,133 organisms.

The database provides a confidence measure for each chemical-protein interaction calculated by the equation score= 1 – Đi(1 – pi), with corrections that take into account the possibility of observing an interaction by chance. In the equation, pi denotes the confidence of interaction from the i-th information source. Based on STITCH, a score between 0.40 and 0.70 indicates medium confidence, between 0.70 and 0.90 indicates high confidence, and between 0.90 and 1.00 indicates the highest confidence.

To retain high-confidence chemical-protein interactions, we filtered out entries in STITCH 3.1 with confidence scores of <0.70. In addition, we removed all entries of chemical interaction with non-human proteins. The filtering reduced the total number of small molecule-protein interaction entries from >171 million to just over a half million. The categories of chemical-protein interactions with the highest occurrence in the database are binding (chemical binds to protein), inhibition (chemical inhibits protein function), and activation (chemical enhances protein function). Because the therapeutic effects of most drugs are due to chemical modulation of protein function, functional information of chemical-protein interactions, i.e., inhibition or activation, is important. However, this information is not always available. Instead, the most prevalent type of interaction information is binding. To create drug-protein interaction profiles relevant for drug repurposing, we retained interactions of only these three categories. This left 445,162 interactions between chemicals identified by 232,765 unique STITCH chemical identifiers and 6,399 unique human proteins.

Source of FDA-approved drugs and drug development Candidates -

To generate a list of FDA-approved drugs and drug development candidates, we retrieved the SMILES strings of all structurally unique small molecule compounds in Drug-Bank. Molecular structures represented by the SMILES strings were standardized, i.e., we stripped salts, standardized charge representation, removed stereochemistry labeling, removed single atom fragments, neutralized bonded zwitterions, and protonated acids/deprotonated bases. After structure standardization, we generated canonical SMILES and removed duplicates, resulting in 4,902 unique entries. They consisted of 1,163 FDA-approved drugs, 3,630 drug development candidates, 55 nutraceuticals, and 54 drugs withdrawn from market. These molecules are all referred to as “drugs” in the remainder of this article.

Computational prediction of drug-protein interactions -

Most of the compounds in DrugBank are in the biological activity screening libraries of pharmaceutical companies, government research laboratories, and academic institutions. However, not all of the Drug Bank compounds have been tested in all assays evaluating chemical-protein interactions and, hence, the data collected in the STITCH database do not cover all drug-protein interactions. Thus, to create as complete drug-protein interaction profiles as possible, we complemented the drug-protein interactions contained in the STITCH database with predicted drug-protein interactions based on chemical structural similarity. This was accomplished by re-implementing the similarity ensemble approach (SEA) and predicting additional drug-protein interactions based on the collection of chemical-protein interactions contained in STITCH 3.1. SEA predictions are based on two-dimensional molecular structure similarity as measured by Tanimoto coefficients between a drug molecule and all known ligands of a protein. When the similarity score is high, the probability that the drug interacts with the same protein is high. In this study, we retained drug-protein interaction predictions with a p-value cutoff of 0.01, and combined these predictions with the high-confidence drug-protein interactions contained in the STITCH database. The so-constructed final set of drugprotein interactions is available for non-commercial use (via download at http://www.bhsai.org/downloads/drugrepurposing/).

RESULTS AND DISCUSSION

Details of model development and quality assessment To assess performance of the drug repurposing method described above, we used three model development procedures. Type I model development represents a conventional machine learning process in which a data set is segregated into a training set and a testing set. The training set consists of a subset of samples of the positive class and a subset of samples of the negative class. The remaining samples, including both positive and negative samples, are grouped into the testing set. The model parameters are determined by the training set only. The model is then applied to the testing set to assess its ability to distinguish the positive from the negative samples. In principle, type I models are not suitable for drug repurposing applications because most drugs were developed for treating a specific disease. Accordingly, for most drugs, their ability to treat other diseases has not been systematically evaluated and, in most cases, one cannot confidently label true negative drugs (samples) in the training set.

A more robust model development approach is represented by a type II model, which is trained with a subset of the positive drugs as the positive class and all other drugs collected in a baseline class, i.e., a large set of compounds that may or may not include drugs with a desired therapeutic effect. Because all drugs are used for model development, there is no testing set. However, for drug repurposing, one can simply score all the drugs assigned to the baseline class with the model and evaluate the degree of enrichment of the (known) positive drugs in the highest-scored samples. Type II models are more appropriate than type I models for drug repurposing, based on the premise that there exist drugs with yet unknown desirable therapeutic effects for a disease among the marketed drugs.

CONCLUSIONS

In this article, we described the development of a Bayesian statistics-based computational drug repurposing method termed DPIR and assessed its performance. We demonstrated that the method required very few known drugs to build a successful predictive model for test cases for which there are many approved drugs. We also demonstrated that for trauma-induced hemorrhage, for which only one FDA-approved drug is available, the method gave high scores to two drugs approved for unrelated indications, but with potential therapeutic effects against hemorrhage as supported by literature reports. These results indicate that DPIR is potentially applicable to diseases with as few as one approved drug, a challenging situation for methods based on a binary classifier approach. DPIR relies on largescale drug-protein interaction information. In principle, if one knows the molecular mechanisms of a disease and the details of drug-protein interactions, one can predict whether a drug will have the desired therapeutic effect for a specific disease. However, details of molecular mechanisms of drug action are not well understood and even unknown for many efficacious drugs, complicated by the fact that most drugs interact with a large number of proteins. Bayesian statistics provide a powerful and unbiased approach to identify specific drugprotein interactions critical for a desired therapeutic effect.

REFERENCES

Ashburn, T.T. and Thor, K.B. (2004). Drug repositioning: Identifying and developing new uses for existing drugs, Nat. Rev. Drug Discov., 3(8): pp. 673-683.

Ashburn, T.T., Thor, K.B. (2004). Drug Repositioning: Identifying and Developing New Uses for Existing Drugs. Nature Reviews Drug Discovery 3, pp. 645–646.

Cheng F, Liu C, Jiang J, Lu W, Li W, Liu G, Zhou W, Huang J, Tang Y (2012). Prediction of drug-target interactions and drug repositioning via network-based inference. PLoS Comput Biol, 8(5):e1002503.

Chong, C.R. and Sullivan Jr., D.J. (2007). New uses for old drugs. Nature 448, pp. 645–646.

Dudley JT, Deshpande T, Butte AJ (2011). Exploiting drug-disease relationships for computational drug repositioning. Brief Bioinform, 12 (4): pp. 303–311.

Ghofrani HA, Osterloh IH, Grimminger F: Sildenafil: from angina to erectile dysfunction to pulmonary hypertension and beyond. Nat Rev Drug Discov 2006, 5(8): pp. 689–702.

Haupt VJ, Schroeder M (2011). Old friends in new guise: repositioning of known drugs with structural bioinformatics. Brief Bioinform, 12(4): pp. 312–326.

Hurle, M.R., Yang, L., Xie, Q., Rajpal, D.K., Sanseau, P., Agarwal, P. (2013). Computational drug repositioning: from data to therapeutics. Clin. Pharmacol. Ther. 93(4), pp. 335–341.

Kotelnikova, E., Yuryev, A., Mazo, I., Daraselia, N. (2010). Computational approaches for drug repositioning and combination therapy design. J. Bioinform Comput. Biol. 8(3), pp. 593–606.

Moriaud F, Richard SB, Adcock SA, Chanas-Martin L, Surgand JS, Ben Jelloul M, Delfaud F. (2011). Identify drug repurposing candidates by mining the protein data bank. Brief Bioinform, 12(4): pp. 336–340.

Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR (2010). Schacht AL: How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nat Rev Drug Discov, 9(3): pp. 203–214.

Sardana, D., Zhu, C., Zhang, M., Gudivada, R.C., Yang, L., Jegga, A.G. (2011). Drug repositioning for orphan diseases. Brief Bioinform 12(4), pp. 346–356.

Sonner JM, Cantor RS: Molecular mechanisms of drug action: an emerging view. Annu Rev Biophys 2013, 42: pp. 143–167.

Corresponding Author

Rajeev Kumar*

Research Scholar, SSSUTMS, Sehore

E-Mail – chairman.iab@gmail.com