Drug Target Identification by Maximizing Information Flow

Bingjie Zhou

under the direction of

Dr. Shao Li

Bioinformatics Division and Center for Synthetic and Systems Biology

Tsinghua University

Research Science Institute-Tsinghua

August 16, 2015

Abstract

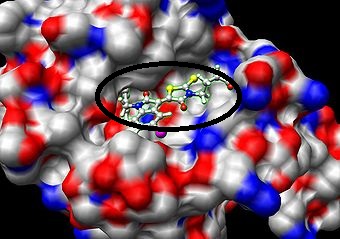
Identifying drug targets is a critical step in pharmacology.[4]However,up to now, targets that found from proteins engage only very small part of total amount of targets.Although several probable models have been proposed to predict drug target like Cipher relating pharmacological and genomic spaces, they still largely base on assumption of result model and are greatly influenced by extreme values.

We proposed combinatorial approach for drug identifying.We set up a drug protein network with known drug drug similarity, drug target interaction and protein protein interaction as capacity of each connection.Also we set thresholds to eliminate small values which have weak biological meaning in pharmacology.Then we used a MAXIF method to calculate the largest association between a pair of drug and protein based on network flow.We had leave-one-out cross validation in which we cover part of known drug target interaction and then use the MAXIF to predict and rank them.From the validation result, we could say MAXIF approach is pretty much fixed to drug target process.

Our findings demonstrated that MAXIF approach is fit and accuracy to drug identification.And this new efficient method could not only speed up our development of drugs and targets, but also improve our theory of side effects and toxicity.

1. Introduction

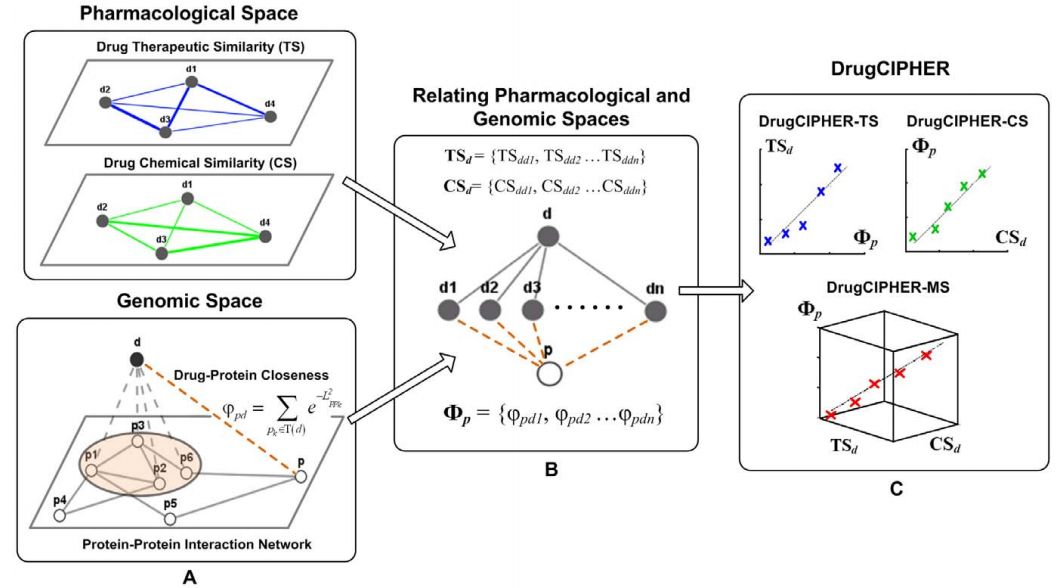
An ideal drug target is a native protein in the body whose activity is modified in a specific and controllable manner resulting a specific effect[1]. Over eighty percent of drug targets are protein and almost fifty percent are belong to GPCRs, serine, threonine and protein tyrosine kinase, MMPs, serine protease, hormone nuclear receptor and PDEs. In theory, proteins that act as drug targets must be capable of combining small molecules associated to diseases with appropriate chemical characters and affinity and could express in pathology cells and tissues resulting in therapeutic effect [2]. For example, Alzheimer’s disease has complex etiology and it attacks many molecular segments, which provides a multitude of potential drug targets. Moreover, our research aims to reduce side effects and toxicities of drugs, caused by unwanted adverse effect of targets.



**Figure 1: drug target combining** *small molecule drugs usually combine with fit sites on targets using hydrogen bonds*

In the most basic model, we assume that one drug acts selectively to one specific target. However, recent research has revealed a far more complex figure of drug reaction in the past decades. An elegant new theory by Yildirim illustrates that not only are there multiple drugs relating to one target but that one target could associate to many drugs [3]. Furthermore, this theory suggests that we view drug actions as networks, which improves our discovery of targets by drug-drug and protein-protein associations. Although many effective analyses such as phenotypic effect and chemical-based approaches have previously existed, there are apparent shortcomings. For the phenotypic approach, off-target responses were activated due to similar pathways or biological processes. For the chemical approach, researchers often do not have enough prior information for major chemicals to know downstream effects. Consequently, there is urgent need to combine phenotypic and chemical indexes together, to achieve an improved method to predict drug-target association in on a larger scale [4].

Taking these observations into account, the Cipher model was proposed and validated as the most accurate way to predict drug targets, given both phenotypic and chemical indexes. .More specifically, the Cipher model assumes that similarities in pharmacological space (termed as Therapeutic Similarities(TS)) and drug chemical similarities(CS) are correlated with the relatedness of targets based on protein-protein interaction(PPI) network in genomic space. Based on this assumption, a computational framework drugCipher was built to relate pharmacological and genomic spaces with multi-dimensional information and used three linear regression models to calculate concordance scores which used as standard of correlation between drug and protein [4].



**Figure 2:principle of drugCipher** *Drugs are solid nodes and presented by ‘****d****’; proteins are hollow nodes and presented by ‘****p****’. A). Drug Therapeutic Similarity (****TS****) (blue solid edges) and Drug Chemical Similarity (****CS****) (green solid edges) comprise the pharmacological space. The protein protein interaction (****PPI****) (gray solid edges) network represents the information in the genomic space. Together with drug-target interactions (gray dashed edges), the closeness (brown dashed edges) is defined to associate a drug with any arbitrary protein. B). For drug d and protein p, two similarity vectors for d in pharmacological space (****TSd*** *and* ***CSd****) and one closeness vector for* ***p*** *(****Wp****) are constructed. C. The concordance scores between drug* ***d*** *and protein* ***p*** *are computed based on three linear regression models, which assume linear correlations exist between TSd and Wp,Wp and CSd, Wp and the combination of TSd and CSd.[4]*

Although Cipher has a high sensitivity and precision score among all approaches, the linear regression model is not clear under the situation that the protein-protein interactions and drug-drug similarities are low, showing on the left, top and right, bottom part of graph. Likewise, the approaches above largely depended on PPI networks to estimate similarity. Since these approaches consider the shortest path as the optimal path between proteins and may overlook others paths, the reliability of the optimal path and the final result may be adversely affected [5].

Motivated by these observations, we propose a new combinatorial approach to calculate score of association between a pair of drug and protein in this paper. In our approach, we first introduce a threshold to select out high similarities which have biological meanings. Then we set up a drug-protein network by integrating given drug-target interactions, PPI network information and drug-drug similarity from DrugBank database. The magnitudes of associations are recorded under the assumption that drugs with high similarities have stronger associations as the capacity of each edge for final calculation. Then we will judge the strength of the relation based on the amount of information flow between drug and protein and we will develop MAXIF(Maximizing information flow) algorithm in the drug-protein network system to figure out the path containing the maxima information and compute the max information flow as the concordance score of this pair of drug-protein. Proteins with high score are considered as potential targets. We show the accuracy and competitive of our method by series designed experiments and comparisons with other approaches and we our method to do couple of predictions as successful applications.

Once the approach is validated to be accurate, we could use it to predict potential targets which could excavate novel applications of existing drugs and narrow the range of test experiments, to pinpoint drugs with specific curative effect. If we generate enough data for such targets, we could minimize problems of side effects and raise efficiency of drug-based therapies. Thus, identifying drug targets in a more efficient way will act as an important stepping-stone in pharmacology.

2. Materials and Methods

2.1 Construction of network

2.1.1 Data preparing

To construct the drug-protein network, we need to prepare data including drug-drug similarity, protein-protein interaction(PPI) and known drug targets interaction first.

First, we downloaded instructions of 6810 drugs from DrugBank[6] website version 4.3 and selected out chemical structure of every drug. OpenBabel[7] is a chemical toolbox which could convert one chemical structure format to a zero-one series and FP3[8] is one of methods in OpenBabel which defines 55 kinds of substructures and records as either one when the substructure is in the chemical structure of drug or zero when it is not in. By turning chemical structures into FP3 format and comparing 0-1 series, we could calculate out corresponding similarity score. Additionally,we assumed that drugs with more similar structures have more connections. So drugs with high similarity score have stronger association. We finally got 23177836 pairs of drug similarity scores.

For protein-protein interaction, we also downloaded protein protein comparisons from String[10] website in July, 2015. Then we selected out pairs with confidence score higher than 700 according to STRING document[9] and ultimately obtained 650466 pairs of proteins.

We also downloaded drug details from DrugBank website and selected known targets and experimental targets of each drug out from the file and finally got 15305 pairs of drug-target links. The drug-target links are significantly important because they connect protein-protein network and drug-drug network to form the final interactions between drugs and proteins.

After all these data have been ready, we use each drug-drug interaction, protein-protein interaction and drug target interaction as one edge, each drug and protein as one node and interaction score of every interaction as weight of every edge to form the final network.

2.1.2 Threshold and construction of network

Since most small similarity scores for drug drug interactions are disturbing and only high similarity scores have true biological meanings, we set up a threshold α for drug-drug similarity scores which means we only use drug pairs with similarity scores higher than or equal to the threshold. [5] And because drug target interactions have been demonstrated already, we must provide the influence of flows in drug target connections. So we set a very large coefficient β which could reach millions for drug target interaction score.

Then we regard every drug and protein as a node in the network and record every edge linking nodes.Every edge has attributes including the name of beginning drug or protein, next nodes connecting to it, capacity of the edge which is interaction score,current flow in this arc,and reverse arc of the edge.With these definitions, the drug-protein network is denoted as indirected graph G=(V,E) in which V means vertexes and E means edges.Each edge E (u,v) (from u to v) has a positive capacity c(u,v) which record their maximum flow in the edge.

2.2 Maximizing information flow in network to obtain interaction score

Our purpose is to identify drug targets with other known drug-drug ,protein-protein and drug-protein information.The process is just like a substance flows from source point to the sink point by consuming energy in a network. By coincidence, model of network flow has the same principal. And we wonder the strongest association, so we use a method called Maximizing information flow(MAXIF) which is a special computation in network flow to calculate out largest association score between arbitrary pair of drug and protein using the drug protein network.

Each directed edge of the flow network could be considered as tubes to transport substances.Every substance has its fixed capacity which we define as the interaction score we first imported.Vertexes are connecting points of tubes.Two vertexes, source and sink, are special because source is the point where all flows come out and sink is where all flows converge finally. Except source and sink points, flows only pass these vertexes instead of accumulating in them.In other words, amount of substance flowing into the node is equal to amount flowing out.The characteristic is called flow conservation.[11]Maximizing information flow is a specific problem in network flow.It’s goal is to calculate out the maximum flow to transport a substance from source to sink.In our project, the goal is converted to calculate out the maximum interaction score between a pair of unknown drug and protein which is quite similar to MAXIF model.Here are some properties in network flow:

(1) current flow of each edge is lower than its capacity

(2) reverse flow of edge is equal to its opposite number

(3) total amount of flow from source is equal to total amount of flow entering sink

The input of our MAXIF program includes total number of proteins and drugs which are considered as nodes in the network n and total number of pairs which are considered as edges m.And the drug-protein network G = ( V , E ) including begin node u[i] ,end node v[i] and capacity of edge c ( u , v ) flowing from u to v. First, we incorporate the source vertex s and define each drug a directed edge of infinite capacity pointing from source vertex s to drug vertexes d which means there is no connection between the nodes. Similar, we incorporate a sink vertex t and define each protein a directed edge of infinite capacity pointing from each protein vertex p to sink vertex t. So we obtain a directed graph G = ( V , E ) , where V = V ∪ {s,t} and E = E ∪ { ( s , d ) ∪ ( p , t ) }.We also define capacity function c ( u , v ) = c ( v , u ) and c ( s , d ) = c ( p , t ) = ∞ . Then we input the network file which contains nodes and capacities and build up the drug protein network with exact capacity of every edge.

It is obvious that the network is a flow network if drug and protein networks are connected.And there is must at least one route of link between each drug and each protein.We regard G as a information flow network and interpret information flow f as a scheme of distribution of total amount of information from source s over all the edges in the network such that total amount of flow from source is equal to total amount of flow accepted by sink.

The total amount of flow from source is determined by value of information flow in network.And the flow with max value in the scheme is the route we desire because the route allowing the maximum information flow from the source has the maximum relation between the pair of drug and protein. According to literature[],a maximum flow could be calculated out by ISAP algorithm. And we multiply all capacities by a ten times number to turn the result capacities into integers.

Once the score has been calculated out, we obtained the exact association between the drug and protein and could make a deeper analyze between the drug and protein. Cause the efficient algorithm and simple construction of network, we could simply get association score from the maximum flow in the network to firstly identify whether the protein is a target or not.[5]In this way, the efficiency of matching drug target is raised to a large extent and also the range of experiments that scientists need to do to seek targets is reduced too.

2.3 Validation methods

We perform a leave-one-out cross validation experiment to examine the capability of our method in discovering targets that are known to be associated with certain drug from a set of candidates. First, in validation, we take a known drug target association in each run, assume that the association is not known and prioritize the target against 99 control drug that are selected randomly from proteins. Second, after each validation run ,we are able to gain a list of rank proteins.We then calculate rank ratios of drugs by dividing their ranks with the number of drugs in the list. Third, we are able to generate a receiver operating characteristics(ROC) curve [12] to appraisal the sensitivity and specifity of the method and further calculate the area below the curve(AUC).Obviously, larger AUC values indicate higher performance of a prioritization method.

3. Results

Replace this text with your own writing, but retain the font and formatting: This section should state what are your research findings, either in chronological order, or most often, with the most important findings described first. This section should include figures and possible tables with captions, such as the example figure below. Your minipaper should have at least one figure, one table, and one equation (see examples).

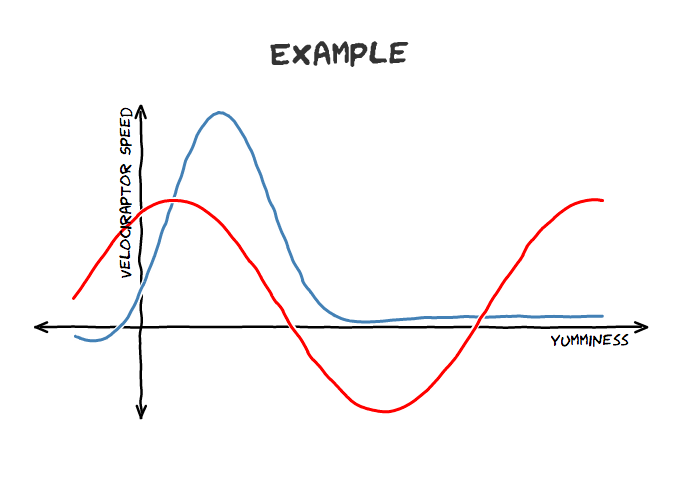


Figure 1 This example figure offers a simple graphical representation of velociraptor speed vs. yumminess. Figures (which can include graphs, charts, pipelines or any visual representation of a process or data) are typically favorable over a table.

|  |  |
| --- | --- |
| Velociraptor | Trex |
| 1 | 3 |
| 2 | 4 |

Table 1 This is an example table. Your minipaper should have at least one table. Your final paper may or may not have a table.

An example of an equation that you may use for your minipaper would be that describing standard deviation (go to “insert” and “equation” to find equation tools in Word):

4. Discussion

4.1 Works of previous methods

5. Conclusion

Replace this text with your own writing, but retain the font and formatting: The conclusion should be a brief summary of what you have done. You wish to leave the reader with a clear impression of your ideas and what you have accomplished. This section is typically extremely short and in the past tense.

6. Acknowledgements

First, I want to give my best thanks to my mentor Dr. Shao Li and comentor Mr.Songpeng Xu.They gave me great help no matter in the way of thinking or skills in writing codes and data analyzing. They also make me develope a very good behavior on dealing with my study files during the research.Also I want to thank my tutor Ana Lyons who gave me great help in paper writing and skills of presentation.Ana is a very serious and inspiring person.She gave me a lot of encouragements and very detail remarks of my work which may influence my writing in whole life.In addition , I want to thank RSI and CEE giving me this excellent opportunity to approach the true scientist research and make me have such a unforgettable summer.I also want to thank my friends and labmates in RSI.They are really creativity and friendly and truly gave me a lot of help in different area.Finally, thanks to Tsinghua university giving us such a wonderful study environment.Thank you.

References

[1] Rang HP, Dale MM, Ritter JM, Flower RJ, Henderson G (2012). ["Chapter 2: How drugs act: general principles"](http://www.inkling.com/read/rang-dale-pharmacology-ritter-flower-henderson-7th/chapter-2/protein-targets-for-drug-binding). *Rang and Dale's Pharmacology*. Edinburgh; New York: Elsevier/Churchill Livingstone. pp. 6–19. [ISBN](https://en.wikipedia.org/wiki/International_Standard_Book_Number" \o "International Standard Book Number) [0-7020-3471-1](https://en.wikipedia.org/wiki/Special:BookSources/0-7020-3471-1" \o "Special:BookSources/0-7020-3471-1).

[2]Jianxiong Yang,Zhihui Liu,Effect of drug targets in developing novel drugs,*Lishizhe Medicine and Materia Medica Research* ,vol. 20 no.3(2009)

[3]Andrew L. Hopkins,Network biology illustrates our understanding of drug actions, *Nature Biotechnology,*vol 25,no.10 (October 2007)

[4]Shiwen Zhao, Shao Li,Network-based relating pharmacological and genomic spaces for drug target identification.*Plos one ,*vol 5,issue 7(July 2010)

[5]Yong Chen,Tao Jiang,Shao Li,Uncovering disease genes by maximizing information flow in the phenome-interactome network,*Bioinformatics,* vol 27, page i167-i176(2011)

[6]Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, et al. (2008) DrugBank:  
a knowledge base for drugs, drug actions and drug targets. Nucleic Acids Res 36:  
D901–D906.

[7]O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. (2011). ["Open Babel: An open chemical toolbox"](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3198950). *Journal of Cheminformatics*

[8]unknown;”*openbabel instruct” ;http://emuch.net/html/201207/4695885.html*

[9]Peer Bork,Lars Juhl Jensen,Christian von Mering ,”Chapter 4 :FAQ” ,String documentation(2000-2011)

[10] Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M, Bork P, von Mering C (2009). ["STRING 8--a global view on proteins and their functional interactions in 630 organisms"](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2686466). *Nucleic Acids Res* **37** (Database issue): D412–6.

[11]Thomas H. Cormen, Charles E.Leiserson,Ronald L.Rivest, Clifford Stein [Chapter 26:network flow] *Introduction to algorithms* (Second Edition)(ISBN :0-262-03293-7)

1. Title of Appendix

This section is OPTIONAL and may be deleted. Appendices may appear after the paper proper. Appendices may hold extra information that would interrupt the flow of the paper and that is not absolutely necessary for the reader to appreciate the work. For example, a large number of related figures or a mathematical derivation could go nicely in an appendix.