Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

Subject Section

Combining Bottom-up and top-down approaches for drug-target-interaction prediction

Tilman Hinnerichs [1,*] and Robert Hoehndorf [2]

[1]Department, Institution, City, Post Code, Country and
[2]Department, Institution, City, Post Code, Country.

*To whom correspondence should be addressed.
Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text.
Results: Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text
Availability: Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text
Contact: tilman.hinnerichs@kaust.edu.sa
Supplementary information:10264703 Supplementary data are available at *Bioinformatics* online.

1 Introduction

Wang and Kurgan [2018]
In history, traditional remedies, that were known for their medicinal properties lead to drugs by extraction of the functional ingredients. Alternatively, characteristics and features of potential drugs were detected by accident like in the case of penicillin. More recently, biological drug targets can be found *in silico* through discovery of suitable computational predictors.
The challenge of accurately predicting drug-target-interactions (DTI) has shown its importance in the fields of drug repurposing and repositioning, and in the exploration of novel drugs and their interaction partners. Knowledge about those links between compounds and their target proteins help in an array of medical and pharmaceutical studies. Additionally, those associations can be utilized to identify disease specific targets, leading to desirable therapeutic effects.
With the rapidly growing field of machine learning approaches and their application to bioscientifical problems in the realm of bioinformatics, different kinds of data, such as long DNA sequences could be utilized for feature generation, while rapid advances were made. Almost all state of the art models for drug-target-interaction prediction were based on the usage of neural networks with increasing size.
Only recently, the technique of graph learning was introduced by Kipf and Welling [2016] through graph convolution algorithms, and improved and altered under usage of different kernels(Defferrard et al. [2016], Bianchi et al. [2019]) attention mechanisms (Veličković et al. [2017]), random walks (Klicpera et al. [2018]) and mixtures of both (Hamilton et al. [2017]). While based on diverse systems, they can be relevant for testing distinct hypothesis for given graphs. While convolutional filters are suitable for finding patterns among the the given graph, attention mechanisms are more relevant for discovery of important regions within. Lately, graph learning approaches found application for computing compound representations for DTI prediction.
Approaches on this rather sophisticated problem can divided into top-down or network approaches (**CITATION**), and bottom-up or molecular approaches (**CITATION**). Top-down approaches take advantage of other data such as diseases (CITATION), side effects, knowledge graphs or ontologies, in order to learn representations for both compound and protein. On the other hand, bottom-up aspirations attempt to learn from chemical properties. For drugs, molecular structure (CITATION GraphDTA), molecular fingerprints, similarity to other drugs (See Bioinf Survey), and other molecular features. On the protein side, secondary structure prediction (CITATION), contact prediction (CITATION), or simply convolution over the amino acid sequences served to obtain a feature for the given proteins. However, both kinds contain and share some problems, that are not solvable within themselves. Thus, bottom-up approaches share the lack of ability to generalize, which we will show in later sections, and usually focus on engineering sophisticated features for the drugs, while neglecting to formulate meaningful features on the protein side. Top-down approaches lack the ability to spot small differences to cope with small differences

within the drug structure and rely heavily on given data for the considered drug-target pair. The latter is not suitable for predictions on novel or unseen compounds, as e.g., data on side effects or its impact on diseases is seldom given for novel drugs.

In order to design such a feature for proteins and drugs, respectively, we make use of the interaction networks for both proteins and compounds. Drug-drug interaction networks were introduced and standardized by Ayvaz et al. [2015] and have been used for clinical decision support (Scheife et al. [2015]). Drug-drug interaction networks may give a hint on common targeted pathways. As an additional compound feature we will use semantic side effect similarity, which we will discuss later on.

Protein-protein interaction networks have shown great results in . . . (Vazquez et al. [2003], Ackerman et al. [2019]) in granting context for molecular system biology. However, these contexts were never applied to the problem of drug-target-interaction prediction. Thus we formalized our hypotheses over these interaction graphs and will test them in the following chapters.

2 Approach

As shown, recent work lacks the ability to combine both top-down and bottom-up approaches for their predictions, however performing quite well on their datasets. When building the train-test split over compound-protein pairs, there are the following options:

1. Build split over drugs
2. Build split over drug-target pairs
3. Build split over proteins

In general, recent works do perform their split over the drugs or drug-target pairs (Wang and Kurgan [2018], CITATION). The first is more relevant for novel drugs, as it is much more likely to test a new compound than a innovative protein. However, it lies in the very nature of the used datasets, making the prediction for new drugs much easier. Thus, drugs are often built by minor variations of existing drugs, thus leading to no deviations in the functional group of that very compound (CITATION/EXAMPLE). When distributed over both train and test split, the models do not perform inductive inference and generalize, but rather implement transductive inference by just predicting the recently seen structures. Hence, when entirely new molecules are seen, the models perform much worse.

The same applies to splits of drug-target pairs, as all drugs were already seen, and novelty cannot be coped with.

As mentioned in the introduction it is quite difficult to learn suitable features from proteins. In general, attempts search for motifs in the protein sequences under usage of convolutional neural networks and filters, which is more suitable for tasks like protein function prediction, than for for drug-target interaction prediction, and lack a more in-depth hypothesis on the protein side, while investing in refined drug features.

Thus, building splitting over proteins is the most challenging of the three options.

In this work we are testing the following hypotheses:

1. Can we build a best performing model that outperforms state of the art approaches, combining top-down and bottom-up approaches?
2. Are interaction networks sufficient to improve the performance of simple molecular predictors?

We will test the first hypothesis by building a model that takes both top-down and bottom-up features into account. For a

3 Approach

4 Methods

5 Discussion

6 Conclusion

Acknowledgements

Funding

This work has been supported by the... Text Text Text Text.

References

Emily E. Ackerman, John F. Alcorn, Takeshi Hase, and Jason E. Shoemaker. A dual controllability analysis of influenza virus-host protein-protein interaction networks for antiviral drug target discovery. *BMC Bioinformatics*, 20(1), June 2019. doi: 10.1186/s12859-019-2917-z. URL https://doi.org/10.1186/s12859-019-2917-z.

Serkan Ayvaz, John Horn, Oktie Hassanzadeh, Qian Zhu, Johann Stan, Nicholas P. Tatonetti, Santiago Vilar, Mathias Brochhausen, Matthias Samwald, Majid Rastegar-Mojarad, Michel Dumontier, and Richard D. Boyce. Toward a complete dataset of drug–drug interaction information from publicly available sources. *Journal of Biomedical Informatics*, 55:206–217, June 2015. doi: 10.1016/j.jbi.2015.04.006. URL https://doi.org/10.1016/j.jbi.2015.04.006.

Filippo Maria Bianchi, Daniele Grattarola, Cesare Alippi, and Lorenzo Livi. Graph neural networks with convolutional arma filters. 2019.

Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. 2016.

William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. 2017.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. 2016.

Johannes Klicpera, Aleksandar Bojchevski, and Stephan Günnemann. Predict then propagate: Graph neural networks meet personalized pagerank. 2018.

Richard T. Scheife, Lisa E. Hines, Richard D. Boyce, Sophie P. Chung, Jeremiah D. Momper, Christine D. Sommer, Darrell R. Abernethy, John R. Horn, Stephen J. Sklar, Samantha K. Wong, Gretchen Jones, Mary L. Brown, Amy J. Grizzle, Susan Comes, Tricia Lee Wilkins, Clarissa Borst, Michael A. Wittie, and Daniel C. Malone. Consensus recommendations for systematic evaluation of drug–drug interaction evidence for clinical decision support. *Drug Safety*, 38(2):197–206, January 2015. doi: 10.1007/s40264-014-0262-8. URL https://doi.org/10.1007/s40264-014-0262-8.

Alexei Vazquez, Alessandro Flammini, Amos Maritan, and Alessandro Vespignani. Global protein function prediction from protein-protein interaction networks. *Nature Biotechnology*, 21(6):697–700, May 2003. doi: 10.1038/nbt825. URL https://doi.org/10.1038/nbt825.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. 2017.

Chen Wang and Lukasz Kurgan. Review and comparative assessment of similarity-based methods for prediction of drug–protein interactions in the druggable human proteome. *Briefings in Bioinformatics*, 20(6):2066–2087, 08 2018. ISSN 1477-4054. doi: 10.1093/bib/bby069. URL https://doi.org/10.1093/bib/bby069.