Hybridization of Chemoinformatic and Bioinformatic Information Networks for Graph Diffusion of Drug-Target Interactions

Bruno Kaufman[1] and Ariel Chernomoretz[1,2]

[1]Fundación Instituto Leloir
[2]Physics Department, FCEN, University of Buenos Aires/IFIBA (CONICET)

Abstract

Drug development is a costly, time-consuming process. It is therefore convenient to repurpose existing drugs for novel treatments. Massive amounts of information exist on the topic of drug-protein interactions. In order to extract new insights from this wealth of data, massive computational approaches are warranted, combining bioinformatics and chemoinformatics. In our work, we explore the combination of several sources of data: drug-protein interactions and chemical similarity measures based on molecular fingerprints (Tanimoto similarity) and drug scaffold hierarchies. These come in the form of complex networks, and a hybrid network is obtained by combining them. In this contribution we present a network diffusion strategy to predict new drug-protein interactions.

Keywords — network diffusion, drug repurposing, structural similarity

1 Introduction

Drug repurposing is a problem which seeks to use prior knowledge to infer new possible uses for existing drugs. One way of going about this task is to use data on drug-protein interactions, and attempt to predict new interactions of the same variety. Proteins are, in this context, referred to as "drug targets".

2 Methods

In our work, we make use of TDR-targets, a curated drug-target interaction database (Urán Landaburu et al., 2020). Our goal is to recover excluded parts of this database by combining the remaining information with chemical similarity data between drugs and making use of network diffusion.

Our procedure uses three separate graphs, each constructed with its own measure of drug-drug similarity: (1) Drug similarity by shared targets; (2) Chemical similarity by Tanimoto score; (3) Sharing of chemical scaffolds by Bemis-Murcko criterion (Bemis and Murcko, 1996).

These three networks are combined using a weighted linear combination of their edge values. Each of these contributes to our hybrid, combined network through a separate weight parameter ($\vec{w} = (w_{targets}, w_{Tanimoto}, w_{scaffold})$), which should be learned through training. For a given hyperparameter set $\vec{w}$, we considered the unweighted Laplacian of the integrated network to perform a label propagation procedure (Chapelle et al., 2009).

3 Results

Our integrated network is composed by 5k target and 7M drug nodes linked by 100M drug-drug and 2M drug-target edges. In order to implement a performant prioritization methodology we partitioned the chemical space into Louvain drug communities and looked for cluster specific hyperparameters $\vec{w}$) during the training phase. The rationale for this methodology was that the relative importance of each knowledge layer could be differentially adjusted, better

reflecting the interplay between the propagation of information from known drugs and network topology for different parts of the network.

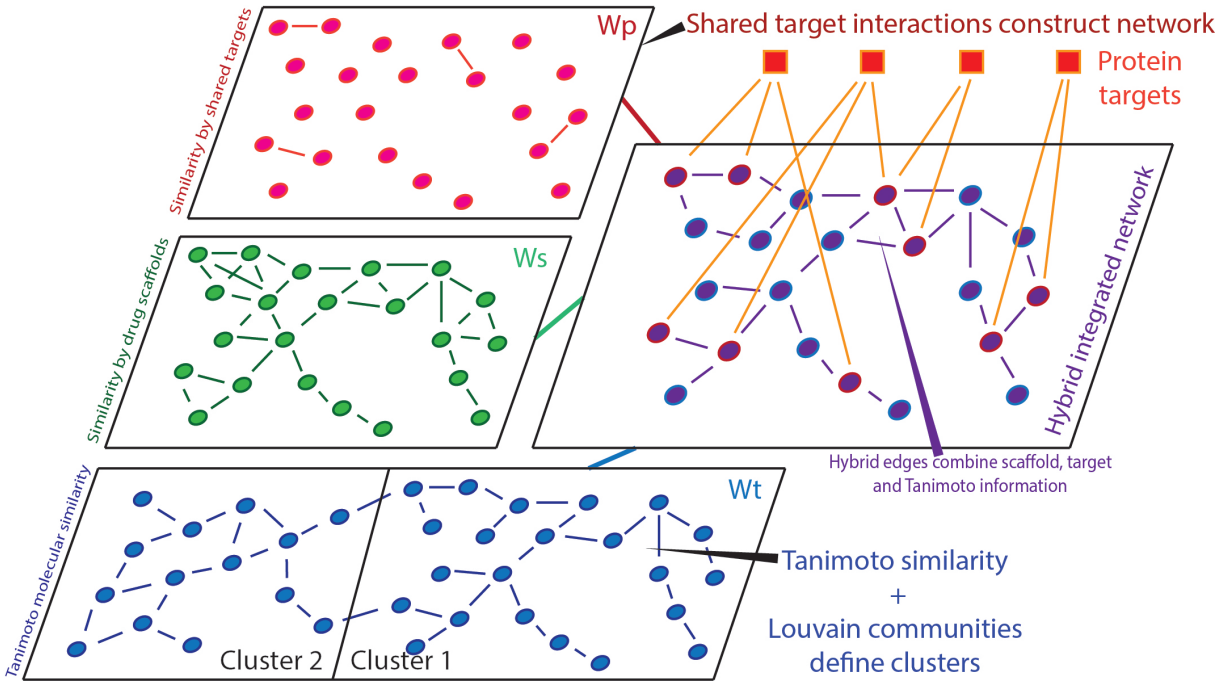


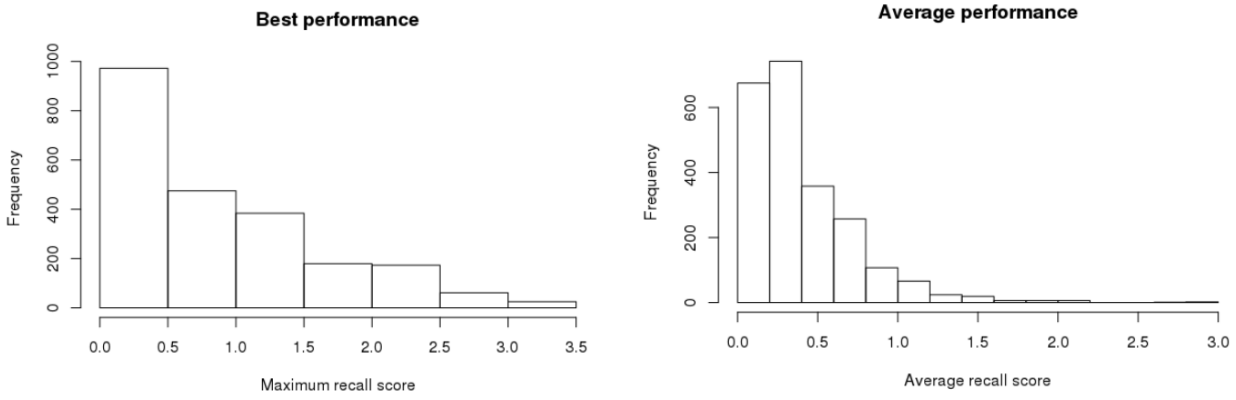
Figure 1: Architecture of information used. Drug targets are depicted as orange squares. According to how many targets a given drug shares, they may be linked by target similarity (top left, red). The same drugs may also be linked if they are comprised of similar scaffolds (middle left, green). Tanimoto similarity also links drugs to each other (bottom, blue); this measure is also used to define Louvain communities. In the example depicted, network diffusion is applied on Cluster 1, using active drugs as seeds (circles with orange perimeter). The parameter balance (W_p, W_s, W_t) is varied until an optimal combination is found in terms of recovery capability. After this, each target is prioritized individually, and their performance is assessed.

Using these locally optimized models we were able to evaluate the prioritization performance for targets on the validation set, according to their recovery capability in terms of the number of true positive elements found within the top 5,10,20,50 ranked elements of the prioritization list. Our score is determined as follows:

$$S_{t,c} = \sum_{n}^{(5,10,20,50)} \frac{E_n}{n}$$

where E_n is the number of elements of the validation set that were found within the top n elements of the ranking.

Table 1: Distribution for the performance of individual targets. This score reflects the recovery of excluded interacting drugs within the top 5, 10, 20 and 50 recommendations.



A target may have different performance in different clusters. Thus we analyze two measures of performance: maximum score among all clusters, and average score among them. Out of 2.2k targets, only around 200 have a performance of zero. The remaining ones are shown in Table 1. To this date, this represents a coverage of roughly 40% of our 5k drug targets. Results are promising, showing efficient use of available information that results in a 5-fold increase in reliable predictions.

References

L. Urán Landaburu, A. J. Berenstein, S. Videla, P. Maru, D. Shanmugam, A. Chernomoretz, F. Agüero, Tdr targets 6: driving drug discovery for human pathogens through intensive chemogenomic data integration, Nucleic acids research 48 (2020) D992–D1005.

G. W. Bemis, M. A. Murcko, The properties of known drugs. 1. molecular frameworks, Journal of Medicinal Chemistry 39 (1996) 2887–2893. URL: `https://doi.org/10.1021/jm9602928`. doi:10.1021/jm9602928. `arXiv:https://doi.org/10.1021/jm9602928`, pMID: 8709122.

O. Chapelle, B. Scholkopf, A. Zien, Eds., Semi-supervised learning (chapelle, o. et al., eds.; 2006) [book reviews], IEEE Transactions on Neural Networks 20 (2009) 542–542. doi:10.1109/TNN.2009.2015974.