Bioinformatics, 2018, 1–2
doi: 10.1093/bioinformatics/bty759
Advance Access Publication Date: 31 August 2018
Applications Note

Structural bioinformatics

HitPickV2: a web server to predict targets of chemical compounds

Sabri Hamad[1,‡], Gianluca Adornetto[1,†,‡], J. Jesús Naveja[1,2,3], Aakash Chavan Ravindranath[1], Johannes Raffler[1] and Mónica Campillos [1,4,*]

[1]Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München, Neuherberg 85764, Germany, [2]Faculty of Medicine, PECEM, UNAM, Mexico City 04510, Mexico, [3]Faculty of Chemistry, DIFACQUIM, UNAM, Mexico City 04510, Mexico and [4]German Center for Diabetes Research, Neuherberg 85764, Germany

*To whom correspondence should be addressed.
[†]Present address: Feral GmbH, c/o CoLaborator (Bayer), Building S141, Muellerstr. 178, 13353, Berlin
[‡]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Alfonso Valencia

Received on March 29, 2018; revised on July 24, 2018; editorial decision on August 25, 2018; accepted on August 30, 2018

Abstract

Motivation: The identification of protein targets of novel compounds is essential to understand compounds' mechanisms of action leading to biological effects. Experimental methods to determine these protein targets are usually slow, costly and time consuming. Computational tools have recently emerged as cheaper and faster alternatives that allow the prediction of targets for a large number of compounds.

Results: Here, we present HitPickV2, a novel ligand-based approach for the prediction of human druggable protein targets of multiple compounds. For each query compound, HitPickV2 predicts up to 10 targets out of 2739 human druggable proteins. To that aim, HitPickV2 identifies the closest, structurally similar compounds in a restricted space within a vast chemical–protein interaction area, until 10 distinct protein targets are found. Then, HitPickV2 scores these 10 targets based on three parameters of the targets in such space: the Tanimoto coefficient (Tc) between the query and the most similar compound interacting with the target, a target rank that considers Tc and Laplacian-modified naïve Bayesian target models scores and a novel parameter introduced in HitPickV2, the number of compounds interacting with each target (occur). We present the performance results of HitPickV2 in cross-validation as well as in an external dataset.

Availability and implementation: HitPickV2 is available in www.hitpickv2.com.

Contact: mcampillos@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1. Introduction

The identification of protein targets of novel compounds is essential to understand compounds' mechanisms of action leading to biological effects. To determine protein targets of compounds, experimental and computational approaches are followed. The former methods are usually slow, costly and time consuming, whereas the latter are cheaper and faster alternatives.

Ligand-based approaches are widely used computational approaches that predict targets of compounds considering that similar ligands bind to common proteins. Interestingly, the fast increase in the number of ligand–protein interactions deposited in public databases is expanding the druggable target repertoire predicted by ligand-based approaches, improving their suitability for the systematic molecular analysis of a larger number of compounds.

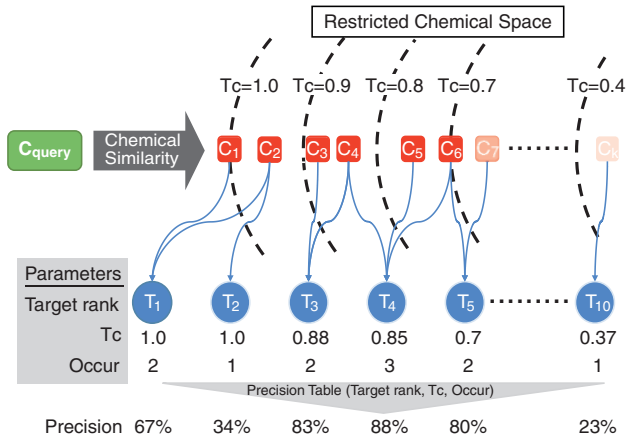

Fig. 1. Illustration of HitPick V2 ligand-based approach. C, compound; T, target

To facilitate these systems pharmacology analyses, we have developed HitPickV2 webserver, which implements an advanced version of the HitPick target prediction method (Liu *et al.*, 2013). HitPick V2 covers 2739 human druggable proteins (1350 more than in the former HitPick version, Liu *et al.*, 2013) and allows the prediction of protein targets (up to 10 per query compound) for multiple ligands, contributing to the expansion our understanding of the biological and phenotypic activity of small molecules.

2. Methods

We first compiled a database consists of 431 352 unique compounds and 891 629 compound–target associations (Supplementary methods). We generated Morgan fingerprint with feature invariants similar to Functional-Class Fingerprints (FCFP) for all compounds in our database using the RDKit nodes implemented in KNIME 3.3. We then created Laplacian-modified naive Bayesian models (Nidhi *et al.*, 2006) for the 2739 druggable proteins interacting with at least three known ligands using KNIME 3.3 Bayesian model builder. For each protein model, we distributed compounds into an active and an inactive group that include compounds interacting and not with the protein, respectively.

To predict targets of a query compound, HitPickV2 identifies the closest, structurally similar compounds in the chemical–protein interaction space using k-nearest neighbours (k-NN) chemical similarity search (Schuffenhauer *et al.*, 2003) and selects a restricted space comprising compounds interacting with 10 distinct protein targets (Fig. 1). For each protein target, we calculated three parameters, namely, 'Tc', 'Target rank' and 'Target occurrence (occur)'. 'Tc' refers to the Tanimoto coefficient between the query compound and most similar compounds in the restricted space interacting with the target. To determine the 'Target rank' of the 10 proteins, we ranked them by their Tc and afterwards, by scores of Laplacian-modified naive Bayesian target models (Nidhi *et al.*, 2006), for proteins with the same Tc (when the most similar compound interacts with more than one target). 'Occur' parameter refers to the number of compounds that interact with each target in the restricted space. Finally, we scored the interaction between the query compound and each candidate target by assigning the precision values corresponding to the 'Tc', 'Target rank' and 'occur' parameters (Supplementary Table S1).

3. Performance

We evaluated HitPickV2 precision in relation to three parameters of the restricted chemical space in cross-validation of the internal dataset (divided in training and validation set in 85%: 15% ratio; the validation set comprises 110 603 compound–target pairs) (Supplementary Table S1). We observed higher precision for increasing values of 'Tc' and 'occur' and decreasing 'Target rank' values. We used the resulting table (Supplementary Table S1) relating precision values to the three parameters of the restricted chemical space to score the HitPickV2 compound–target predictions. Notably, the incorporation of the 'occur' parameter enables HitPickV2 to predict targets with high precision (>50%) for some compounds with very low chemical similarity to ligands with known targets (Tc < 0.4).

We then tested HitPickV2 performance in an external data of 359 compound–target interaction pairs comprising 55 compounds (Klaeger *et al.*, 2017) not present in our internal dataset. For that, we calculated the precision within two 'occur' intervals (occur ≤ 10, occur > 10) for the first Target ranked across all ranges of Tc in cross-validation. Similar to the observed performance of HitPickV2 in our internal database, we obtained higher precision for increasing values of Target occurrence (occur) (Supplementary Fig. S1), reinforcing the influence of the novel parameter 'occur' on HitPickV2 performance.

4. Implementation

HitpickV2 is freely accessible for non-commercial users in a webserver (www.hitpickv2.com). As input, HitPickV2 requires a list of query compounds pasted on the input window or uploaded in a file, represented as SMILES strings. The output is a table with the Query Compound and its 2D structure, predicted targets as gene symbol and precision. Extra columns (Supplementary methods) include this information: protein complex (from Reactome database, Fabregat *et al.*, 2018) of the predicted targets, most similar compound and its 2D structure; Tc between the query compound and the closest compound in the k-NN chemical space annotated to the predicted target and Target occurrence (occur). The 'Occur' field is enhanced with a scaffold perception feature displaying the three most abundant Murcko scaffolds (Bemis and Murcko, 1996) of the compounds interacting with the same target in the restricted space along with their occurrence. The output table can be downloaded as a text or pdf file.

Funding

This work was supported by German Federal Ministry of Education and Research (BMBF) to the German Center for Diabetes Research (DZD e.V.). J.J.-N is thankful to PECEM and UNAM exchange program for the organization of the fellowship.

Conflict of Interest: none declared.

References

Bemis,G.W. and Murcko,M.A. (1996) The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.*, **39**, 2887–2893.

Fabregat,A. *et al.* (2018) The Reactome pathway Knowledgebase. *Nucleic Acids Res.*, **46**, D649–D655.

Klaeger,S. *et al.* (2017) The target landscape of clinical kinase drugs. *Science*. **358**, PubMed PMID: 29191878.

Liu,X. *et al.* (2013) HitPick: a web server for hit identification and target prediction of chemical screenings. *Bioinformatics*, **29**, 1910–1912.

Nidhi,*et al.* (2006) Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. *J. Chem. Inf. Model.*, **46**, 1124–1133.

Schuffenhauer,A. *et al.* (2003) Similarity metrics for ligands reflecting the similarity of the target proteins. *J. Chem. Inf. Comput. Sci.*, **43**, 391–405.