Bioinformatics Advance Access published May 28, 2013

Structural Bioinformatics

HitPick: a web server for hit identification and target prediction of chemical screenings

Xueping Liu[1,2,#], Ingo Vogt[1,2,#], Tanzeem Haque[1] and Monica Campillos[1,2,*]
[1]Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München, Neuherberg, Germany.
[2]German Center for Diabetes Research, Helmholtz Zentrum München, Neuherberg, Germany.

Associate Editor: Prof. Anna Tramontano

ABSTRACT

Motivation: High-throughput phenotypic assays reveal information about the molecules that modulate biological processes such as a disease phenotype and a signaling pathway. In these assays, the identification of hits along with their molecular targets is critical to understand the chemical activities modulating the biological system. Here, we present HitPick, a web server for identification of hits in high-throughput chemical screenings and prediction of their molecular targets. HitPick applies the B-score method for hit identification and a newly developed approach combining 1-Nearest-Neighbour (1NN) similarity searching and Laplacian-modified naïve Bayesian target models to predict targets of identified hits. The performance of the HitPick web server is presented and discussed.

Availability: The server can be accessed at http://mips.helmholtz-muenchen.de/proj/hitpick

Contact: monica.campillos@helmholtz-muenchen.de

1 INTRODUCTION

Chemical biology experiments are increasingly used to search for chemical modulators of biological processes in cell-based and even whole-organism assays as illustrated by the thousands of phenotypic screenings stored in public repositories (Seiler et al., 2008; Wang et al., 2010). In these assays, the identification of the molecular targets of hits is essential to understand the molecular basis of the chemical activities in the bioassay. Recently, drug target prediction methods have been applied to the hits of cells (Young et al., 2008) and zebrafish (Laggner et al., 2012) phenotypic screenings showing that computational approaches are suitable tools that facilitate the interpretation of the biological activity of chemicals.

Although diverse *in silico* methods have been proposed to identify hits (Makarenkov et al., 2007; Malo et al., 2006) and predict targets for chemicals (reviewed in (Kuhn et al., 2008)), only few of them are available as easy-to-use online tools (Keiser et al., 2007; Wang et al., 2012). To overcome this situation and assist in the analysis and interpretation of chemical phenotypic screens, we introduce HitPick, the first web server for hit identification and target prediction of chemical screenings. HitPick provides the functionality to detect bioassay hits using the B-score method and predict targets of a chemical of interest using a new integrative approach that combines 1NN similarity searching and a machine learning method. On cross-validation, the target prediction approach of HitPick performs better than each of the methods alone, achieving a sensitivity of 60.94%, a specificity of 99.99% and a precision of 92.11%.

2 METHODS

We apply the widely used B-score method for hit identification, which uses the median polish procedure to remove the bias in rows and columns in a plate (Malo et al., 2006). Hits are determined by a p-value cut-off of 0.05, and replicates of compounds will be considered as hits when all the replicates are identified independently as hits.

For target prediction, HitPick employs a newly developed approach that combines two methods based on 2D molecular fingerprints, namely, 1NN similarity searching (Schuffenhauer et al., 2003) and Laplacian-modified naïve Bayesian target models (Nidhi et al., 2006). For each query compound the most similar compound from a dataset of known ligand-target interactions is determined by calculating the pairwise Tanimoto coefficient (Tc) (Willett, 1998). Then, Laplacian-modified naïve Bayesian target models generate a score for all known targets of the most similar compound (Nidhi et al., 2006), resulting in a list of ranked target predictions.

For the implementation of this approach we used a set of 145,549 human chemical-protein physical interactions extracted from the STITCH 3.1 database (Kuhn et al., 2012). In this study, we restrict the target prediction to human proteins, as it is currently the species with the largest number of known drug targets, enabling thus more accurate predictions. In total, we obtained 99,572 compounds with unique SMILES strings with known interactions for which we generated 2D circular fingerprints based on the Morgan algorithm with feature invariants similar to the FCFP (Rogers and Hahn, 2010) using RDKit (http://rdkit.org). Using these molecular fingerprints we created Bayesian models for 1,375 proteins with at least three known ligands. For benchmarking, we randomly assigned 85% of the known ligands to the training set and the remaining 15% to the validation set. In total, the validation set contained 22,868 positive and 20,779,507 negative compound-target relationships, respectively.

To facilitate the analysis of experiments with many hits, the target prediction for screenings with more than 100 hits is carried out for a structurally diverse subset of 100 compounds obtained by applying the MaxMin-Algorithm (Ashton et al., 2002) implemented in RDKit.

The fingerprint creation for the STITCH compounds, building and application of Bayesian target-specific fingerprint models were implemented in a KNIME (http://www.knime.org) workflow making use of the chemoinformatic functionality provided by KNIME itself as well as by RDKit.

[#] These authors contributed equally to this work.
[*] To whom correspondence should be addressed

3 PERFORMANCE

We assessed the performance of the target prediction method in HitPick using as validation set 15% of all ligands that were not part of the training set. When evaluating the highest scoring target prediction for each compound, HitPick achieves a sensitivity of 60.94% (with 66.16% being the maximum possible sensitivity), a specificity of 99.99% and a precision of 92.11%, an improvement over naïve Bayesian models (sensitivity of 52.95%, specificity of 99.98%, precision of 80.03%) and 1NN similarity searching (precision of 84.72%). HitPick performance is comparable to the target prediction quality achieved by the Similarity Emsemble Approach (SEA), a well-known target fishing application that relates proteins based on the chemical similarity of their ligands (Keiser et al., 2007) (not shown).

We also evaluated the performance of the HitPick target prediction method at different ranges of chemical similarity of the query compound to the closest training compound and for up to five top scoring known targets of this training molecule independently. In order to obtain robust precision estimates we require a minimum of 30 compound-target predictions for each target rank in a given Tc interval (Table. 1).

We observed that the precision increases with increasing Tc. For compounds with a Tc of 0.7 or higher to the training set, the first predicted target was nearly always correct. Furthermore, the precision reached at least 53% for a Tc in the range of 0.4~0.5 (Table. 1) Thus, we chose 50% as default precision threshold for the predicted targets on the web server.

Ranked prediction	[0.2~0.3)	[0.3~0.4)	[0.4~0.5)	[0.5~0.6)	[0.6~0.7)	[0.7~0.8)	[0.8~0.9)	[0.9~1.0)	1
1st	15.7	26.4	53.3	77	89.8	94.8	96.7	97.7	97.3
2nd	15.4	14.5	43.5	54.3	64.1	68.9	75.7	88.2	83
3rd	NA	NA	24.6	39.1	48.3	63.2	71.9	77.9	66.7
4th	NA	NA	15.4	33.3	36.1	62	57.9	77.6	56.5
5th	NA	NA	NA	NA	29.6	46.1	51.8	NA	NA

Table 1. Precision (%) for the first five predicted targets in relation to the Tc similarity of a validation compound to the most similar molecule in the training set. The precision for cells marked as "NA" could not be determined due to the low number of compound-target predictions (less than 30).

4 IMPLEMENTATION

HitPick offers two independent functions. The first function identifies bioassay hits based on the B-score method and predicts targets for up to 100 hits. As input, it requires the data from a bioassay, including plate names, compound identifiers, well positions, activity values and SMILES strings. The output is a table listing the hits and their chemical structures. This table is used as input source for the target prediction method. The output of the target prediction is a list of target predictions for the input compounds ranked by decreasing precision.

In addition, HitPick allows the prediction of targets for up to 100 compounds independently from bioassay data. For this second function only SMILES strings are required as input.

To ensure a high reliability of the target prediction outcome, HitPick reports only those targets per compound for which we can reliably estimate the precision. The precision depends on the similarity to the most similar compound in the set of known interactions as well as on the rank of the target's score. Users can select different precision thresholds for the target prediction results as desired. Under a lower threshold, more chemicals will have predictions at the cost of a lower precision. In addition, an overview of the predicted targets is given in form of pie chart.

The processing time for hit identification depends on the size of the assay data. For bioassays containing less than 5000, 10,000 and 100,000 compounds, the web server returns the results in less than one, two and 30 minutes, respectively. The target prediction takes around five minutes per batch of query.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge Michael Kuhn for providing the STITCH data, the support of the TUM Graduate School's Faculty Graduate Center Weihenstephan at the Technische Universität München, Germany and Jonathan Hoser and Jeanette Prinz for technical support.

Conflict of interest: None declared.

REFERENCES

Ashton, M. et al. (2002) Identification of diverse database subsets using property-based and fragment-based molecular descriptions, *Quant Struct-Act Relat*, **21**, 598-604.

Keiser, M.J. et al. (2007) Relating protein pharmacology by ligand chemistry, *Nat Biotechnol*, **25**, 197-206.

Kuhn, M. et al. (2008) Large-scale prediction of drug-target relationships, *FEBS Lett*, **582**, 1283-1290.

Kuhn, M. et al. (2012) STITCH 3: zooming in on protein-chemical interactions, *Nucleic Acids Res*, **40**, D876-880.

Laggner, C. et al. (2012) Chemical informatics and target identification in a zebrafish phenotypic screen, *Nat Chem Biol*, **8**, 144-146.

Makarenkov, V. et al. (2007) An efficient method for the detection and elimination of systematic error in high-throughput screening, *Bioinformatics*, **23**, 1648-1657.

Malo, N. et al. (2006) Statistical practice in high-throughput screening data analysis, *Nat Biotechnol*, **24**, 167-175.

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.

Rogers, D. and Hahn, M. (2010) Extended-connectivity fingerprints, *J Chem Inf Model*, **50**, 742-754.

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

Seiler, K.P. et al. (2008) ChemBank: a small-molecule screening and cheminformatics resource database, *Nucleic Acids Res*, **36**, D351-359.

Wang, J.C. et al. (2012) idTarget: a web server for identifying protein targets of small chemical molecules with robust scoring functions and a divide-and-conquer docking approach, *Nucleic Acids Res*, **40**, W393-399.

Wang, Y. et al. (2010) An overview of the PubChem BioAssay resource, *Nucleic Acids Res*, **38**, D255-266.

Willett, P. (1998) Chemical Similarity Searching, *J.Chem.Inf.Comput.Sci*, **38**, 983-996.

Young, D.W. et al. (2008) Integrating high-content screening and ligand-target prediction to identify mechanism of action, *Nat Chem Biol*, **4**, 59-68.