BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 16 2013, pages 2062–2063
doi:10.1093/bioinformatics/btt322

Gene expression

Advance Access publication June 5, 2013

DIGEP-Pred: web service for *in silico* prediction of drug-induced gene expression profiles based on structural formula

Alexey Lagunin*, Sergey Ivanov, Anastasia Rudik, Dmitry Filimonov and Vladimir Poroikov

Laboratory for Structure-Function Based Drug Design, Orekhovich Institute of Biomedical Chemistry of the Russian Academy of Medical Sciences, Moscow 119121, Russia

Associate Editor: Olga Troyanskaya

ABSTRACT

Summary: Experimentally found gene expression profiles are used to solve different problems in pharmaceutical studies, such as drug repositioning, resistance, toxicity and drug–drug interactions. A special web service, DIGEP-Pred, for prediction of drug-induced changes of gene expression profiles based on structural formulae of chemicals has been developed. Structure–activity relationships for prediction of drug-induced gene expression profiles were determined by Prediction of Activity Spectra for Substances (PASS) software. Comparative Toxicogenomics Database with data on the known drug-induced gene expression profiles of chemicals was used to create mRNA- and protein-based training sets. An average prediction accuracy for the training sets (ROC AUC) calculated by leave-one-out cross-validation on the basis of mRNA data (1385 compounds, 952 genes, 500 up- and 475 down-regulations) and protein data (1451 compounds, 139 genes, 93 up- and 55 down-regulations) exceeded 0.85.

Availability: Freely available on the web at http://www.way2drug.com/GE.

Contact: alexey.lagunin@ibmc.msk.ru

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

Received on December 6, 2012; revised on May 1, 2013; accepted on May 30, 2013

1 INTRODUCTION

Drug-induced gene expression profile (DIGEP) is an important determinant of the drug effect on a cell. Recent rapid advances in DNA microarray technology allow one to use in practice the DIGEPs (Chengalvala *et al*., 2007). Lamb introduced Connectivity Map (CMap) as a phenotype-based drug discovery approach based on comparison of disease gene signature and drug-induced changes in gene expression profiles (Lamb *et al*., 2006). Since then, more than 50 studies using CMap approach for drug repositioning, lead discovery, mechanism of action elucidation and so forth were published (Qu and Rajpal, 2012). It was shown that CMap approach can be used to reveal side effects of drugs (Huang *et al*., 2010; Minowa *et al*., 2012), estimation of drug resistance (Iskar *et al*., 2010) and analysis of the drug synergistic effects (Jin *et al*., 2011).

CMap approach is applicable only for several thousand drugs having experimentally determined DIGEPs but cannot be used for other drugs or new drug candidates. To overcome this limitation, we decided to create a web service to predict DIGEPs for new drug-like compounds on the basis of the existing data on DIGEPs.

There are several freely available databases on experimental DIGEPs, including NCBI Gene Expression Omnibus (Barrett *et al*., 2007), Connectivity Map Database (Lamb *et al*., 2006) and Comparative Toxicogenomics Database (CTD) (Davis *et al*., 2013). Structure–activity relationships are traditionally used to decrease the number of experiments in the process of drug development. The gene expression changes can be considered as a particular type of the biological activity of a drug. Therefore, Prediction of Activity Spectra for Substances (PASS) approach can be applied for prediction of DIGEPs. To the best of our knowledge, the developed web service is the first tool that allows the prediction of drug-induced changes in gene expression profiles based on the structural formulae of chemical compounds.

2 METHODS

The data on drug-induced changes of gene expression were downloaded from CTD (version of October, 2012). The CTD data were derived from scientific literature based on the change in an experimentally measured level of mRNA or protein. CTD contains the data on gene ontology, associations and relationships of genes with diseases and biological pathways that give an insight into the influence mechanisms of chemical compounds on the human health (Davis *et al*., 2013). Despite CTD data not being cell dependent, different studies showed that, based on the gene expression data of one cell line, the active compounds against another cell line can be found (Karube *et al*., 2013; Sirota *et al*., 2011). CTD provides *per se* general drug-induced changes of gene expression that can be used in drug discovery with some limitations and assumptions (Qu and Rajpal, 2012). Such general drug-induced changes of gene expression reflect the intrinsic properties of a drug, and their different fractions may be revealed in the appropriate cell lines or tissues under the certain experimental conditions.

Based on mRNA and protein data, we have created two independent training sets that provide a more definite prediction of the change in mRNA expression and appropriate protein concentration. Structural formulae of chemicals were obtained from ChemIDPlus (http://chem.sis.nlm.nih.gov/chemidplus) and ChemSpider (http://www.chemspider.com). The structures of single organic molecules with molecular weight 50–1250 Da and the data on drug-induced changes of human-specific gene expression were selected from 2638 agents in CTD. mRNA-based training set consists of 1385 compounds changing the expression of 14 700 genes (11 131 up- and 10 980 down-regulations). Protein-based training set consists of 1451 compounds changing the expression of 1950 genes (1357 up- and 1204 down-regulations).

Developed over 20 years ago and brought up-to-date, PASS is a computer program to predict qualitatively different biological activity types of

*To whom correspondence should be addressed.

© The Author 2013. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

Downloaded from https://academic.oup.com/bioinformatics/article/29/16/2062/199662 by guest on 10 April 2024

drug-like organic compounds based on their structural formula (Filimonov and Poroikov, 2008). PASS is based on the use of Bayesian approach and Multilevel Neighborhoods of Atom descriptors for representation of a molecule structure (Lagunin *et al.*, 2010; Poroikov *et al.*, 2000). It was earlier shown that the Bayesian approach used in PASS, being selected by comparison of many mathematical algorithms for analysis of 'structure–activity' relationships, overtakes them by accuracy of prediction (Filimonov and Poroikov, 2008) and robustness (Poroikov *et al.*, 2000). A freely available web service for prediction of the biological activity spectra for drug-like molecules on the basis of PASS technology (http://www.way2drug.com/PASSonline) was developed earlier (Sadym *et al.*, 2003). PASS prediction results are represented by the list of activities with probabilities "to be active" Pa and "to be inactive" Pi. The list is arranged in a descending order of Pa–Pi; thus, the more probable changes of gene expression are at the top of the list. The list can be shortened at any desirable cutoff value, but $Pa > 0.5$ is used by default. In this case, both the chance to confirm the predicted DIGEPs by the experiment and the number of false-positive predictions are reasonable. The detailed description of PASS approach is represented in the Supplementary Material.

3 RESULTS

During PASS training, we selected the activities with three and more active compounds and with >0.75 receiver operating characteristic (ROC) Area under an ROC curve (AUC) values calculated by leave-one-out cross-validation procedure to reach accurate and robust results of DIGEP prediction. The average accuracy of prediction for mRNA-based training set of 952 genes (500 up- and 475 down-regulations) and for protein-based training set of 129 genes (85 up- and 51 down-regulations) was 0.853 and 0.858, respectively (the accuracy of prediction for each activity is represented in the Supplementary Material). Such accuracy is reasonable for application of DIGEP-Pred to make prediction for new compounds. Input data include the structural formula presented as MOL file or SMILES, or can be drawn by MarvinSketch (Fig. 1).

Input data are sent to PASS service that makes DIGEP prediction on the basis of above mentioned training sets. The prediction results are kept in MySQL database. Web service through PHP scripts and appropriate links provides a possibility for a user to save DIGEP prediction results as SDF or CSV files. One can use either mRNA- or protein-based prediction or both. The service provides links between gene names in predicted DIGEPs and CTD site, which simplifies the interpretation of predicted results because of access to the relationships of genes with diseases, side effects and biological pathways.

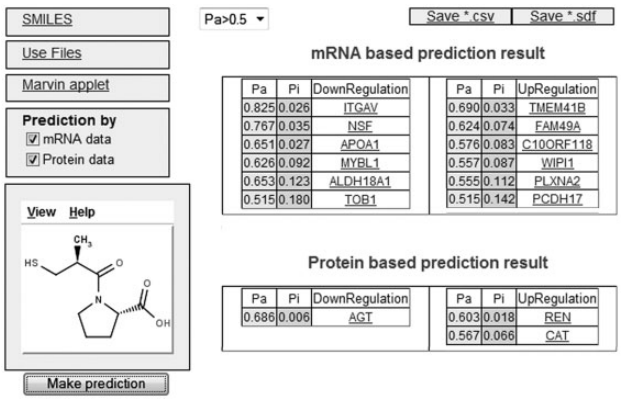

Fig. 1. The prediction results for captopril with the web service

DIGEP prediction for captopril (angiotensin-converting enzyme inhibitor used as antihypertensive drug), that was not included in the training sets, was analyzed by CTD data (Fig. 1). The relationships between genes from predicted DIGEP and hypertension (REN, CAT, AGT and APOA1) and known side effects (MediGuard: https://www.mediguard.org/medication/side-effects/captopril) (fetal death: TMEM41B, PLXNA2, PCDH17, MYBL1; dizziness: WIPI1, AGT, REN; fatigue: TMEM41B, C10ORF118, WIPI1, TOB1, REN; cough: REN) were found. Apart from analysis between the changes of single gene expression and diseases, the results of DIGEP prediction may be used as input data of the methods based on CMap approach for drugs without known experimentally determined microarray data.

Funding: This work was partly supported by the Russian Foundation for Basic Research [12-07-00597-a].

Conflict of Interest: none declared.

REFERENCES

Barrett,T. *et al.* (2007) NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Res.*, **35** (Database issue), D760–D765.

Chengalvala,M.V. *et al.* (2007) Gene expression profiling and its practice in drug development. *Curr. Genomics*, **8**, 262–270.

Davis,A.P. *et al.* (2013) The Comparative Toxicogenomics Database: update 2013. *Nucleic Acids Res.*, **41** (Database issue), D1104–D1114.

Filimonov,D. and Poroikov,V. (2008) Probabilistic approach in activity prediction. In: Varnek,A. and Tropsha,A. (eds) *Chemoinformatics Approaches to Virtual Screening*. RSC Publishing, Cambridge, UK, pp. 182–216.

Huang,H. *et al.* (2010) Bayesian approach to transforming public gene expression repositories into disease diagnosis databases. *Proc. Natl Acad. Sci. USA*, **107**, 6823–6828.

Iskar,M. *et al.* (2010) Drug-induced regulation of target expression. *PLoS Comput. Biol.*, **6**, pii: e1000925.

Jin,G. *et al.* (2011) An enhanced Petri-net model to predict synergistic effects of pairwise drug combinations from gene microarray data. *Bioinformatics*, **27**, i310–i316.

Karube,K. *et al.* (2013) Comprehensive gene expression profiles of NK cell neoplasms identify vorinostat as an effective drug candidate. *Cancer Lett.*, **333**, 47–55.

Lagunin,A. *et al.* (2010) Multi-targeted natural products evaluation based on biological activity prediction with PASS. *Curr. Pharm. Des.*, **16**, 1703–1717.

Lamb,J. *et al.* (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, **313**, 1929–1935.

Minowa,Y. *et al.* (2012) Toxicogenomic multigene biomarker for predicting the future onset of proximal tubular injury in rats. *Toxicology*, **297**, 47–56.

Poroikov,V.V. *et al.* (2000) Robustness of biological activity spectra predicting by computer program PASS for non-congeneric sets of chemical compounds. *J. Chem. Inf. Comput. Sci.*, **40**, 1349–1355.

Qu,X. and Rajpal,D. (2012) Applications of connectivity map in drug discovery and development. *Drug Discov. Today*, **17**, 1289–1298.

Sadym,A *et al.* (2003) Prediction of biological activity spectra via the Internet. *SAR QSAR Environ. Res.*, **14**, 339–347.

Sirota,M. *et al.* (2011) Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Sci. Transl. Med.*, **3**, 96ra77.

Downloaded from https://academic.oup.com/bioinformatics/article/29/16/2062/199662 by guest on 10 April 2024