Bioinformatics, 31(20), 2015, 3392–3394
doi: 10.1093/bioinformatics/btv373
Advance Access Publication Date: 17 June 2015
Applications Note

OXFORD

Databases and ontologies

Rchemcpp: a web service for structural analoging in ChEMBL, Drugbank and the Connectivity Map

Günter Klambauer, Martin Wischenbart, Michael Mahr, Thomas Unterthiner, Andreas Mayr and Sepp Hochreiter*

Institute of Bioinformatics, Johannes Kepler University Linz, Altenbergerstr. 69, 4040 Linz, Austria

*To whom correspondence should be addressed.
Associate Editor: Janet Kelso

Received on February 11, 2015; revised on April 27, 2015; accepted on June 11, 2015

Abstract

Summary: We have developed Rchempp, a web service that identifies structurally similar compounds (structural analogs) in large-scale molecule databases. The service allows compounds to be queried in the widely used ChEMBL, DrugBank and the Connectivity Map databases. Rchemcpp utilizes the best performing similarity functions, i.e. molecule kernels, as measures for structural similarity. Molecule kernels have proven superior performance over other similarity measures and are currently excelling at machine learning challenges. To considerably reduce computational time, and thereby make it feasible as a web service, a novel efficient prefiltering strategy has been developed, which maintains the sensitivity of the method. By exploiting information contained in public databases, the web service facilitates many applications crucial for the drug development process, such as prioritizing compounds after screening or reducing adverse side effects during late phases. Rchemcpp was used in the DeepTox pipeline that has won the Tox21 Data Challenge and is frequently used by researchers in pharmaceutical companies.
Availability and implementation: The web service and the R package are freely available via http://shiny.bioinf.jku.at/Analoging/ and via Bioconductor.
Contact: hochreit@bioinf.jku.at
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

One strategy to remedy the steadily declining efficiency of the drug development process is to utilize the vast information on chemical compounds that is already present in public and internal databases (Arrowsmith, 2011; Scannell *et al.*, 2012). To this end, compounds that are structurally similar to the current drug candidate have to be identified in public databases. This 'structural analoging' step can enhance the drug development process at different phases: After high-throughput screening, a number of compounds are selected, of which typically very little information about their biological effects is available. Querying these compounds in a database can yield similar compounds that have already been measured in various biological experiments and provide relevant information (Klambauer *et al.*, 2013; Robert *et al.*, 2008; Rodriguez *et al.*, 2005) (Fig. 1). This information can improve the drug development process at various stages and decreases the risk of late failures (Verbist *et al.*, 2015).

The best-performing similarity measures for identifying structurally similar compounds in a database are molecule kernels. Molecule kernels have been developed for similarity-based machine learning methods, such as Support Vector Machines, and their efficiency and precision have been demonstrated in various tasks (Mahé and Vert, 2009; Mahe (2005); Mohr *et al.*, 2010; Ralaivola *et al.*, 2005). The drawback of molecule kernels is that they are computationally demanding, since they consider all possible common substructures of a compound. To reach a query time of a few seconds, which is desired for web services, we have

© The Author 2015. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

Downloaded from http://bioinformatics.oxfordjournals.org/ at University of California, Los Angeles on August 30, 2016

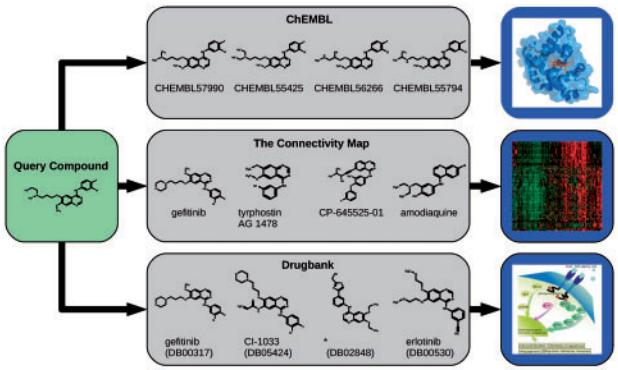

Fig. 1. Structural analoging: compounds in Drugbank (bottom), the Connectivity Map (middle) and ChEMBL (top) that are similar to a query compound (left) are identified. Information on the biological effects of the analogs, like mechanism of action, gene expression and bioassay measurements, from these databases also characterize the query compound

developed a novel and computationally fast criterion that maintains the sensitivity of the molecule kernels. We have extensively tested the presented software in our computational pipeline for prediction of toxic effects of chemical compounds. This computational pipeline 'DeepTox' has won both the Tox21 Data Challenge (http://www.ncats.nih.gov/news-and-events/features/tox21-challenge-winners.html) and the prediction of average cytotoxicity at the NIEHS-NCATS-UNC DREAM Toxicogenetics Challenge (https://www.synapse.org/#!Synapse:syn2208373/wiki/).

We provide a web service, called Rchemcpp, to identify structurally similar compounds in ChEMBL (Gaulton *et al.*, 2012), Drugbank (Law *et al.*, 2014) and the Connectivity Map (Lamb *et al.*, 2006). This provides researchers with various types of information about chemical compounds and thereby helps to prioritize structures and facilitates decisions based on biological measurements. Rchemcpp combines efficiently implemented molecule kernels with a new prefiltering strategy and offers a user-friendly interface to adjust the kernel parameters. Rchemcpp is a valuable tool for researchers in many areas in particular in drug design, toxicity prediction and cancer research.

The web service Rchemcpp has many advantages over a programming library for molecule kernels. To use molecule kernels requires extensive programming efforts. Therefore, many researchers are restricted to database queries provided by the database interfaces, which are based on simple substructure searches. We provide molecule kernels as a hands-on tool for researchers of all areas and backgrounds. Furthermore, the preparation, curation and clean-up of the databases that is necessary for applying molecule kernels is included the Rchemcpp web service. Additionally, the problem of high computational power that restricts queries in the large databases such as the ChEMBL has been solved by our prefiltering strategy.

2 Methods

The method consists of computation of the molecule kernel and a prefiltering step for speeding up the calculations. For large databases, such as ChEMBL, a prefiltering step is necessary to identify structural analogs in reasonable time for a web service.

2.1 Molecule kernels

A molecule kernel is a function that maps two chemical compounds to a similarity value. The similarity measure is typically based on structural similarity, i.e. the number of substructures that occur in both compounds. If we consider two molecules X and Y, then the kernel K is:

$$K(X,Y) = \sum_{p\in\mathcal{P}} N(p,X)\cdot N(p,Y), \quad (1)$$

where $\mathcal{P}$ is the set of substructures, and the function $N(p, X)$ counts how often the substructure p occurs in molecule graph X. The function $N(p, X)$ is specific for the different kernel types (see Supplementary Section S4).

2.2 Prefiltering

For query compound X and each compound Y from the database, a small set of features is calculated in advance to make the comparison computationally fast. We investigated how this prefiltering strategy changes the result compared to a database query without prefiltering. With the prefiltering strategy, a speed-up is obtained that is necessary for a web service, while at the same, some highly similar molecules might be lost. We calculated how much the sensitivity of the method is decreased by filtering out a certain percentage of molecules from the database (see Supplementary Fig. S2): Overall, the decrease in sensitivity is small, e.g. when removing 99.5% of the compounds in the prefiltering step, i.e. a speed-up of 200-fold, still a sensitivity of 55% is kept.

3 Example: a query compound

To illustrate our structural analoging approach, we use a query compound (left in Fig. 1), of which only the chemical structure is known. In the following, we show how to obtain information on this query compound using our novel web service Rchemcpp.

3.1 Biomolecular targets from CHEMBL

The query for structural analogs in ChEMBL, a database with 1 million molecules, provided compound 'CHEMBL57990' as the closest analog to our query compound (top in Fig. 1). For this compound, ChEMBL reports strong inhibitory activity against the epidermal growth factor receptor (EGFR). This receptor is crucial for the proliferation of cancer cells. A compound with inhibitory activity against EGFR can be a potential drug for certain cancer types. All 10 closest structural analogs to the query compound are strong EGFR inhibitors (see Supplementary Table S1), therefore we can hypothesize that the query compound may act as an EGFR inhibitor.

3.2 Gene expression signature from the Connectivity Map

When we identified the closest structural analogs from the Connectivity Map, gefitinib was ranked first, followed by a tyrphostin AG 1478, a compound named 'CP-645525-01' and amodiaquine (middle in Fig. 1). From compounds with a similarity greater than 0.90, a gene expression signature was derived from the Connectivity Map data (see Supplementary Section S2). The gene expression data of the Connectivity Map were preprocessed with FARMS (Hochreiter *et al.*, 2006; Talloen *et al.*, 2007, 2010) and informative/non-informative (I/NI) calls (Clevert *et al.*, 2011; Hochreiter *et al.*, 2010; Hochreiter, 2013; Klambauer *et al.*, 2012).

3.3 Mechanism of action from Drugbank

The closest structural analogs in Drugbank are gefitinib, compounds with the identifiers DB03365, DB02984, DB02848 and erlotinib

Downloaded from http://bioinformatics.oxfordjournals.org/ at University of California, Los Angeles on August 30, 2016

(bottom in Fig. 1). The Drugbank contains information on the mechanism of action of these compounds. The database reports that both gefitinib and erlotinib are marketed drugs for the treatment of certain tumor types. Both have EGFR as biomolecular target. For gefitinib, the mechanism of action is known: it binds to the ATP-binding site of EGFR, inactivates the Ras signal transduction cascade and thereby inhibits malignant cells. We have now obtained a second hint that the query compound may act as an EGFR inhibitor.

In summary, we have acquired information about a query compound: bioassay measurements from ChEMBL, a potential gene expression signature from the Connectivity Map and a potential mechanism of action from Drugbank. The results indicate that the query compound could be an EGFR inhibitor.

Funding

This work was supported by the Institute for the Promotion of Innovation by Science and Technology in Flanders [IWT 135122].

Conflict of Interest: none declared.

References

Arrowsmith,J. (2011) Trial watch: phase III and submission failures: 2007–2010. *Nat. Rev. Drug Discov.*, **10**, 87.

Clevert,D.-A. *et al.* (2011) cn.FARMS: a latent variable model to detect copy number variations in microarray data with a low false discovery rate. *Nucleic Acids Res.*, **39**, e79.

Gaulton,A. *et al.* (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.*, **40**, D1100–D1107.

Hochreiter,S. (2013) Hapfabia: Identification of very short segments of identity by descent characterized by rare variants in large sequencing data. *Nucleic Acids Res.*, **41**, e202.

Hochreiter,S. *et al.* (2006) A new summarization method for affymetrix probe level data. *Bioinformatics*, **22**, 943–949.

Hochreiter,S. *et al.* (2010) FABIA: factor analysis for bicluster acquisition. *Bioinformatics*, **26**, 1520–1527.

Klambauer,G. *et al.* (2012) cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res.*, **40**, e69.

Klambauer,G. *et al.* (2013) DEXUS: identifying differential expression in RNA-seq studies with unknown conditions. *Nucleic Acids Res.*, **41**, e198–e198.

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

Law,V. *et al.* (2014) Drugbank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.*, **42**, D1091–D1097.

Mahé,P. and Vert,J.-P. (2009) Graph kernels based on tree patterns for molecules. *Mach. Learn.*, **75**, 3–35.

Mahé,P. *et al.* (2005) Graph kernels for molecular structure-activity relationship analysis with support vector machines. *J. Chem. Inf. Model.*, **45**, 939–51.

Mohr,J. *et al.* (2010) A maximum common subgraph kernel method for predicting the chromosome aberration test. *J. Chem. Inf. Model.*, **50**, 1821–1838.

Ralaivola,L. *et al.* (2005) Graph kernels for chemical informatics. *Neural Netw.*, **18**, 1093–1110.

Robert,R. *et al.* (2008) Structural analog of sildenafil identified as a novel corrector of the f508del-cftr trafficking defect. *Mol. Pharmacol.*, **73**, 478–489.

Rodriguez,A.L. *et al.* (2005) A close structural analog of 2-methyl-6-(phenylethynyl)-pyridine acts as a neutral allosteric site ligand on metabotropic glutamate receptor subtype 5 and blocks the effects of multiple allosteric modulators. *Mol. Pharmacol.*, **68**, 1793–1802.

Scannell,J.W. *et al.* (2012) Diagnosing the decline in pharmaceutical r&d efficiency. *Nat. Rev. Drug Discov.*, **11**, 191–200.

Talloen,W. *et al.* (2007) I/NI-calls for the exclusion of non-informative genes: a highly effective filtering tool for microarray data. *Bioinformatics*, **23**, 2897–2902.

Talloen,W. *et al.* (2010) Filtering data from high-throughput experiments based on measurement reliability. *Proc. Natl. Acad. Sci. USA.*, **107**, 173–174.

Verbist,B. *et al.* (2015) Using transcriptomics to guide lead optimization in drug discovery projects: Lessons learned from the QSTAR project. *Drug Discov. Today.*, **20**, 505–513.

Downloaded from http://bioinformatics.oxfordjournals.org/ at University of California, Los Angeles on August 30, 2016