BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 22 2012, pages 3002–3003
doi:10.1093/bioinformatics/bts551

Structural bioinformatics Advance Access publication September 8, 2012

ChemBioServer: a web-based pipeline for filtering, clustering and visualization of chemical compounds used in drug discovery

Emmanouil Athanasiadis, Zoe Cournia and George Spyrou*
Biomedical Research Foundation, Academy of Athens, 4 Soranou Ephessiou, 115 27 Athens, Greece

Associate Editor: Anna Tramontano

ABSTRACT

Summary: ChemBioServer is a publicly available web application for effectively mining and filtering chemical compounds used in drug discovery. It provides researchers with the ability to (i) browse and visualize compounds along with their properties, (ii) filter chemical compounds for a variety of properties such as steric clashes and toxicity, (iii) apply perfect match substructure search, (iv) cluster compounds according to their physicochemical properties providing representative compounds for each cluster, (v) build custom compound mining pipelines and (vi) quantify through property graphs the top ranking compounds in drug discovery procedures. ChemBioServer allows for pre-processing of compounds prior to an *in silico* screen, as well as for post-processing of top-ranked molecules resulting from a docking exercise with the aim to increase the efficiency and the quality of compound selection that will pass to the experimental test phase.

Availability: The ChemBioServer web application is available at: http://bioserver-3.bioacademy.gr/Bioserver/ChemBioServer/.

Contact: gspyrou@bioacademy.gr

Received on June 26, 2012; revised on August 6, 2012; accepted on August 31, 2012

1 INTRODUCTION

A rate-limiting step in computer-aided drug design (*CADD*) is often the need for a computational chemist expert to post-process compounds that result from a virtual screening/docking exercise before selecting which ones should be tested experimentally (Dobson, 2004). This post-processing step entails identifying and discarding docking poses that result in intra-ligand steric clashes as well as compounds with undesirable or toxic moieties and unwanted physicochemical properties (Cournia *et al.*, 2009). Cheminformatics web applications play an essential role for searching, filtering and clustering chemical compounds for drug discovery (Backman *et al.*, 2011). In recent years, several chemical compound databases have been developed, including *Zinc*, *chemDB*, *PubChem* and many others (Chen *et al.*, 2007; Irwin and Shoichet, 2005; Li *et al.*, 2010). Nevertheless, although knowledge integration can drastically increase the power and the predictive capability of large-scale computational comparisons of chemical structures, online open access web applications for compound mining are limited in number and, importantly, in pipeline integration level.

Namely, in the EDinburgh University Ligand Selection System (EDULISS) (Hsin *et al.*, 2011), either similarity or pharmacophore search of compounds is performed on commercially available molecules based on a proposed ultrafast shape recognition algorithm or by calculating interatomic distances.

In the SIMilar COMPound and SUBstructure matching of COMPounds web application (Hattori *et al.*, 2010), chemical similarity and substructure searches are computed by means of the maximum common induced subgraph using an atom-based approach and the maximum common edge subgraph or by a bond-based approach. Nevertheless, in both web applications, advanced filtering criteria such as Lipinski Rule of Five (Lipinski *et al.*, 2001) or custom-made filters are not available.

By using the FAF-Drugs2 web application (Lagorce *et al.*, 2011), users are able to process their own compound collections through simple absorption, distribution, metabolism, excretion and toxicity filtering rules to aid the drug discovery process. However, the user is not able to apply sub-graph search with a custom-made Structure Data Format (*SDF*) (Bodenhofer *et al.*, 2011) file. Moreover, van der Waals (*vdW*) energy or geometric criteria are not taken into consideration in the filtering section to discard docking poses with steric clashes. Also, no clustering technique of compounds with similar characteristics is available to any of the previously mentioned web applications.

ChemMine Tools (Backman *et al.*, 2011) is a web service for structure visualization and comparison, similarity searching and compound clustering. However, the user can select only two compounds to compare each time, which is limiting considering the need for massive similarity searches through the compounds of a library. Thus, the process does not facilitate large-scale filtering procedures. Furthermore, neither *vdW* energy or radii restrictions nor toxicity checking is available to the user as a filtering service. Finally, the service does not provide the user with the most representative compound for each cluster.

To overcome all previously mentioned limitations, we have developed *ChemBioServer* as a free web-based application aimed to assist hit selection arising from *CADD*. ChemBioServer is a web application that automates pre-/post-processing tasks during virtual screening. Through a customized workflow, molecules are discarded by evaluating parameters such as vdW energy, geometry, physicochemical properties and undesired/toxic moieties. The web application is implemented so that the post-processing procedure can be tailored to the specific needs of the user as every compound query is unique. It also features a clustering section with two clustering methods, the hierarchical (Backman *et al.*, 2011) as well as the modern (Frey and Dueck, 2007) affinity propagation (*AP*) clustering algorithm, grouping together

*To whom correspondence should be addressed.

and proposing a representative compound for each cluster (Bodenhofer *et al.*, 2011). Additionally, visualization of clusters and graphical representations of molecular properties are also available, which provide insights into the compounds' physicochemical similarity level.

2 METHODS AND IMPLEMENTATION

The *ChemBioServer* web application is divided into six main sections: (i) basic search, (ii) filtering, (iii) advanced filtering, (iv) clustering, (v) customize pipeline and (vi) visualize compounds' properties. The application back-end is developed in *R* programming language (http://cran.r-project.org/), while the front-end is implemented with *PHP* (http://www.php.net/). 2D/3D display of compounds is accomplished by means of the open-source Java viewer for chemical structures *JChemPaint* (http://JChemPaint.sourceforge.net) and *Jmol* (http://www.jmol.org/), respectively. Compound Fingerprints are generated with *Open Babel* (12). The format of the files that are uploaded to the *ChemBioServer* is either *SDF* or *MOL*. However, transformation from other file formats is facilitated through proper links in the help page.

The '*Basic Search*' section (i) enables the researcher to browse the contents of a compound file that is uploaded to the server. After the upload, *SDF* files are processed with the use of the *ChemmineR* (Cao *et al.*, 2008) package for *R*. Compounds are displayed with their molecular name in a list form with their corresponding *SDF* information attached such as the molecular name, the connection table, the atom, bond and property block, etc. In addition, 3D visualization of each compound is available by clicking on the atom name.

In the '*Filtering*' section (ii), compound mining can be performed based on a variety of chemical properties. In the predefined queries section, the researcher is able to upload a file and select compounds that comply with the Lipinski Rule of Five. In addition, in the combined search section, searching can be performed by applying more advanced criteria such as partition coefficient *logP* or Polar Surface Area (*PSA*) (Guha, 2007).

Three main filtering methods are provided in the '*Advanced Filtering*' section (iii): (a) the exact substructure filtering can be accomplished by uploading two compound files and identifying whether compounds of the second file can be found in the first. It should be noted that the second file can also contain fragments of unwanted or toxic moieties; the algorithm recognizes whether the fragment is found within any compound of the first file and reports it. This filtering step is accomplished by converting files to Simplified Molecular Input Line Entry Specification (SMILES) with the use of Open Babel (O'Boyle *et al.*, 2011) and applying maximum common substructure searches for pairs of molecules (Guha, 2007). (b) Additional toxicity filtering is performed by screening out compounds that contain a collection of toxic and carcinogenic fragments that is provided on site. (c) The *vdW* filtering to discard molecules with steric clashes is also provided by means of energy and radii tolerance (Jorgensen *et al.*, 1996). Poses that are far from the energy minimum are unlikely to be adopted in nature and hence should be discarded. In several docking exercises with *Glide* (Schrodinger, *LLC*), we have observed that the post-docking poses often suffer from *vdW* clashes (see *ChemBioServer* help page). In particular, we observed that even after *Glide* post-docking minimization, ~20% of the generated poses should be discarded due to unrealistic *vdW* interactions, which required automating this procedure.

The '*Clustering*' section (iv) includes a classical as well as a modern clustering algorithm. Compound fingerprints are either provided by the user or generated using the 166 bit MACCS Open Babel fingerprint. In the case of hierarchical clustering, the user is able to select the distance, the linkage and a threshold value. In the case of AP clustering, the algorithm takes as input a set of pairwise similarities among compound fingerprints, considering them as potential representative compounds (exemplars). The clusters are calculated by exchanging messages between data points until a maximization process converge. Thus, exemplars for each cluster are proposed to the researcher for further investigation. Additionally, visualization of clusters is also available as a *PDF* dendrogram plot.

In section (v), a pipeline workflow that combines all or part of the previously described filtering services is provided by the ChemBioServer so as to speed up the filtering process. Compound sets are successively tested on enabled filtering modules and molecules that fail are discarded. When the process in completed, the user is provided with a single file that contains compounds that have passed all previously enabled filtering tests.

In the final section (vi), graphical representations of molecular properties can be created by means of the Raphaël javascript library (http://raphaeljs.com/). More precisely, Principal Component Analysis (Wehrens and Buydens, 2007) of the first principal component against the second, based on the tanimoto coefficient can be applied. In addition, graphical representation of *logP* against *PSA* is also available. Finally, *MW* against *PSA* and *logP* plots can be performed. Thus, compounds that have survived from the extensive filtering are then explored for the optimum subset that will pass to the experimental test phase. We have tested this pipeline in several test datasets and found that the Server produced accurate results and tremendously helped our CADD process in one automated step (see Help section of the server).

Funding: This work was funded by the NSRF 2007–2013, cofunded by the European Regional Development Fund and national resources ['Cooperation' No. 09ΣΥΝ-11-675].

Conflict of Interest: none declared.

REFERENCES

Backman,T.W. *et al.* (2011) ChemMine tools: an online service for analyzing and clustering small molecules. *Nucleic Acids Res.*, **39**, W486–W491.

Bodenhofer,U. *et al.* (2011) APCluster: an R package for affinity propagation clustering. *Bioinformatics*, **27**, 2463–2464.

Cao,Y. *et al.* (2008) ChemmineR: a compound mining framework for R. *Bioinformatics*, **24**, 1733–1734.

Chen,J.H. *et al.* (2007) ChemDB update full-text search and virtual chemical space. *Bioinformatics*, **23**, 2348–2351.

Cournia,Z. *et al.* (2009) Discovery of human macrophage migration inhibitory factor (MIF)-CD74 antagonists via virtual screening. *J. Med. Chem.*, **52**, 416–424.

Dobson,C.M. (2004) Chemical space and biology. *Nature*, **432**, 824–828.

Frey,B.J. and Dueck,D. (2007) Clustering by passing messages between data points. *Science*, **315**, 972–976.

Guha,R. (2007) Chemical informatics functionality in R. *J. Stat. Softw.*, **18**, 1–16.

Hattori,M. *et al.* (2010) SIMCOMP/SUBCOMP: chemical structure search servers for network analyses. *Nucleic Acids Res.*, **38**, W652–W656.

Hsin,K.Y. *et al.* (2011) EDULISS: a small-molecule database with data-mining and pharmacophore searching capabilities. *Nucleic Acids Res.*, **39**, D1042–D1048.

Irwin,J.J. and Shoichet,B.K. (2005) ZINC—a free database of commercially available compounds for virtual screening. *J. Chem. Inform. Model.*, **45**, 177–182.

Jorgensen,W.L. *et al.* (1996) Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.*, **118**, 11225–11236.

Lagorce,D. *et al.* (2011) The FAF-Drugs2 server: a multistep engine to prepare electronic chemical compound collections. *Bioinformatics*, **27**, 2018–2020.

Li,Q. *et al.* (2010) PubChem as a public resource for drug discovery. *Drug Discov. Today*, **15**, 1052–1057.

Lipinski,C.A. *et al.* (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.*, **46**, 3–26.

O'Boyle,N.M. *et al.* (2011) Open Babel: an open chemical toolbox. *J. Cheminform.*, **3**, 33.

Wehrens,R. and Buydens,L.M.C. (2007) Self- and super-organising maps in R: the kohonen package. *J. Stat. Softw.*, **21**, 1–19.