BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 4 2013, pages 523–524
doi:10.1093/bioinformatics/bts711

Databases and ontologies

Advance Access publication December 20, 2012

ChEMBLSpace—a graphical explorer of the chemogenomic space covered by the ChEMBL database

Nikolas Fechner†, George Papadatos‡, David Evans, John Richard Morphy, Suzanne Clare Brewerton, David Thorner and Michael Bodkin*

Department of Medicinal Chemistry, Lilly Research Laboratories, Eli Lilly and Company, Erl Wood Manor, Surrey GU20 6PH, UK

Associate Editor: Janet Kelso

ABSTRACT

Motivation: The ChEMBLSpace graphical explorer enables the identification of compounds from the ChEMBL database, which exhibit a desirable polypharmacology profile. This profile can be predefined or created iteratively, and the tool can be extended to other data sources.

Contact: m.bodkin@lilly.com

Received on August 30, 2012; revised on October 31, 2012; accepted on December 16, 2012

1 INTRODUCTION

There is much recent interest in designing ligands with well-defined polypharmacology, arising both from the increased awareness of the multiple interactions of existing drugs, and the likely improved efficacy of treatments, which engage at several points in biological networks. (Dudley *et al*., 2010). It is straightforward to search public or proprietary data sources for compounds, which are known to hit predefined sets of targets to provide starting points for a multi-target discovery project. Meanwhile, visualization approaches have been developed that analyse a biologically motivated protein network for molecules that show interactions with more than one component. (Paolini *et al*., 2006). This allows the opportunistic identification of previously unconsidered target combinations. There are several powerful network visualization and analysis tools available for this purpose, including several free/open-source applications [e.g. Cytoscape (Smoot *et al*., 2011)]. However, these tools are designed to support arbitrary network types; thus, it is not directly possible to analyse a chemogenomic network in combination with molecular structures and activity profiles. This integration of visual network exploration and activity profile lookup can be achieved by ChEMBLSpace. The user interface is designed for the visual exploration of a chemogenomic space in which the network vertices correspond to proteins that have at least one ligand in common with other targets. The user can interactively select proteins and list their ligands, subsequently designing an activity profile by adding more targets or anti-targets to the selection and adjusting protein-specific activity thresholds. Furthermore, the user does not have to know the profile in advance but can create it interactively, considering the matching ligands, as well as other associated proteins that might not have been clear choices initially. The molecules that meet the created activity profile are displayed within the application, and the full collection can be saved as an SD file.

2 METHODS

The ChEMBLSpace explorer itself is a database-independent graphical user interface and consequently requires some preprocessing of the original data. We provide all scripts and tools required to create the underlying data collection as open source, but the GUI will also work if the data are compiled using different methods that follow the approaches presented in this section.

Initially, the basic information is extracted from ChEMBL14 (Gaulton *et al*., 2012) and processed into a new database. For each human protein target in ChEMBL, the list of drug-like compounds that have activity data for the target (subject to additional requirements, such as exact activity reported in nM; standard activity type reported as either EC_{50}, IC_{50} or K_i), were extracted. In case of multiple assay measurements per compound and target, only the most active was selected. Subsequently, this resulted in a subset of 81 986 distinct molecule–protein associations (the full list can be obtained from the supporting material).

This subset, which was compiled using the workflow environment KNIME (Berthold *et al*., 2007), was then used to create a new database, consisting of three basic tables: A target table, which contained the non-redundant list of proteins with unique identifiers; a compound table, with unique molecule IDs and canonical SMILES; and an activity table with the numeric activity value (in log units) between a compound ID and a target ID. Columns with target description and measurement method information were also stored. These three tables served as input for a Python script that subsequently generated a fourth table. The latter resembles an adjacency list for the network. For each molecule with reported activities for more than one target, a row was inserted into the table, thus mapping a compound ID to a pair of targets, along with its respective activities.

In the final step, the resulting database was used to create a collection of flat files that contain all necessary information and can be deployed alongside the ChEMBLSpace explorer, thus rendering the application database independent. In more detail, the network was created as graph object in which a vertex corresponds to a target, and two vertices are connected if the respective targets have at least one ligand in common. The resulting network was stored in the GraphML format. Additionally, for each target, a text file was created containing an activity-sorted list of compounds, including IDs, SMILES and full activity profiles. The created collection of text files is all that is required for the visualization and

*To whom correspondence should be addressed.
†Present address: Novartis Institutes for BioMedical Research, Basel CH-4056, Switzerland.
‡Present address: European Bioinformatics Institute (EBI), Hinxton CB10 1SD, UK.

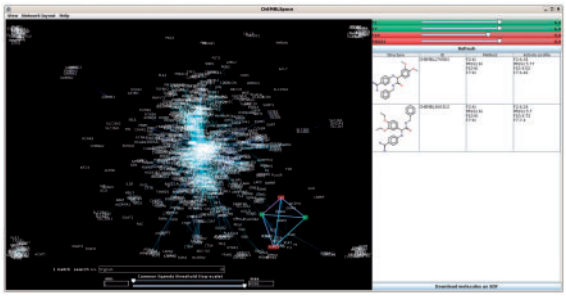

Fig. 1. Screenshot of the ChEMBLSpace explorer with a selected activity profile [Targets: Thrombin (F2), Factor VII >6 pAct; Antitargets: Trypsin (PRSS1) <6. Factor Xa <5 pAct]. The main window is centered on the largest connected component; singletons and small components were placed in the corners by the layouter (Fruchtermann–Reingold in this case)

exploration of the chemogenomic space covered by the underlying database using the presented ChEMBLSpace explorer.

The ChEMBLSpace explorer is a Java application that visualizes and analyses a chemogenomic network (Fig. 1). The visualization is based on the Prefuse (Heer *et al.*, 2005) library. All network visualization components are using the Prefuse API (including layout, colouring, search panel and user interaction). The chemistry engine, which is only required for the molecules rendering and for the SD file export, is provided by the Chemistry Development Kit [(Steinbeck *et al.*, 2003, 2006), version 1.4.8]. The network data are provided as a collection of flat files in the data folder. This folder contains an XML file describing the network topology in the GraphML format and a tab-separated file for each protein containing the molecules that have activity data for the respective protein. The program itself can be run as a stand-alone Java application using the executable jar file or can be used as an applet in a browser by opening the provided HTML file or placing it on a local web server.

3 APPLICATION AND CONCLUSION

When ChEMBLSpace is started, it loads the polypharmacology network from the GraphML file and generates a layout using the Graph API provided by Prefuse. The network is displayed using a radial layout by default. The user can change the layout by selecting one of the methods in the 'Network Layout' menu (the 'Force-directed Layout' continues the energy minimization as long as the mouse is pressed).

The chemogenomic space is visualized as a network of proteins in which two vertices are connected if they share at least one compound that has an activity measured for both of them. The edges are coloured according to their weight, which corresponds to the number of molecules shared by the adjacent vertices. This weight has an effect on some of the layout methods and moreover can be used to filter out edges using the sliders provided on the bottom of the main window. Each node corresponds to a protein, which is defined using the UniProt accession number. Additionally, the first five gene names and a more descriptive name are stored for each protein, whereby the first gene name serves as the node label owing to conciseness and human readability considerations. The descriptive name is displayed on mouseover. All available names (ID, gene names and preferred names) are prefix-searchable using the search panel at the bottom of the main window.

Most of the network visualization can be controlled by intuitive mouse commands (e.g. dragging nodes, zooming and selecting nodes), which are readily provided by the Prefuse library and had only to be associated to some additional scenario-dependent behaviour. The right mouse button opens a context menu that allows the user to restrict the network display to the currently selected nodes. Moreover, it provides an interface to specify an activity profile to retrieve compounds that match this profile. Each protein can get an activity threshold (in log units) assigned, which serves either as a lower limit (target) or as an upper limit (anti-target). The right-hand side of the GUI shows the currently specified profile, as well as a list of compounds that meet these requirements. This list can be exported as an SD file contained the structure and their activities to be used in subsequent experiments.

In conclusion, the ChEMBLSpace explorer provides an intuitive way to explore the chemogenomic space that is covered in a data source and helps the user not only to specify a desired activity profile but also to discover non-obvious relationships amongst proteins and protein families. Furthermore, using an interactive network visualization allows for the exploitation of the powerful pattern recognition capabilities of the human visual cognition, which might help to reveal connections between remote target classes for which a user would not query a database explicitly.

The source code of all developed programs is hosted as a Sourceforge project (https://sourceforge.net/projects/chemblspace) and is available under a Berkeley Software Distribution (BSD)-type license. Moreover, we provide a tutorial, the scripts and the Konstanz Information Miner (KNIME) workflow that have been used to create the local database, an SQL dump of the chemogenomic database at the time the tool was developed, as well as the data files used by the applet. The scripts and workflow also provide a simple starting point to create the polypharmacology database using alternative (e.g. internal) data resources, which subsequently can be used to extract the flat files to allow users to explore different chemogenomic spaces.

Conflict of Interest: none declared.

REFERENCES

Berthold,M. *et al.* (2007) KNIME: The Konstanz Information Miner. *Studies in Classification, Data Analysis, and Knowledge Organization*. Springer, London, UK.

Dudley,J.T. *et al.* (2010) Drug discovery in a multidimensional world: systems, patterns, and networks. *J. Cardiovasc. Trans. Res.*, **3**, 438–447.

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

Heer,J. *et al.* (2005) Prefuse: a toolkit for interactive information visualization. In: *Proceedings of SIGCHI Conference Human Factors in Computing System*. New York, USA, pp. 421–430.

Paolini,G. *et al.* (2006) Global mapping of pharmacological space. *Nat. Biotechnol.*, **24**, 805–815.

Smoot,M. *et al.* (2011) Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, **27**, 431–432.

Steinbeck,C. *et al.* (2003) The chemistry development kit (CDK): an open-source java library for chemo- and bioinformatics. *J. Chem. Inf. Comput. Sci.*, **43**, 493–500.

Steinbeck,C. *et al.* (2006) Recent developments of the chemistry development kit (CDK)—an open-source java library for chemo- and bioinformatics. *Curr. Pharm. Des.*, **12**, 2111–2120.