BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 11 2008, pages 1413–1415
doi:10.1093/bioinformatics/btn138

Data and text Mining

IDMap: facilitating the detection of potential leads with therapeutic targets

Soyang Ha[1,†], Young-Ju Seo[1,†], Min-Seok Kwon[1], Byung-Ha Chang[1], Cheol-Kyu Han[1] and Jeong-Hyeok Yoon[1,2,*]

[1]BT Solutions Team, R&D Center, Equispharm Co., Ltd, Gyeonggi Bio-Center Bld. 11Fl., 864-1, Iui-dong, Yeongtong-gu, Suwon, Gyeonggi-do [443-766], Republic of Korea and [2]Bioinformatics and Molecular Design Research Center, B138A, Yonsei Engineering Research Complex, Yonsei University 134, Sinchon-dong, Seodaemun-gu, Seoul [120-749], Republic of Korea

Received on December 13, 2007; revised on March 21, 2008; accepted on April 11, 2008

Advance Access publication April 15, 2008

Associate Editor: Jonathan Wren

ABSTRACT

Summary: Pharmaceutical industry has been striving to reduce the costs of drug development and increase productivity. Among the many different attempts, drug repositioning (retargeting existing drugs) comes into the spotlight because of its financial efficiency. We introduce IDMap which predicts novel relationships between targets and chemicals and thus is capable of repositioning the marketed drugs by using text mining and chemical structure information. Also capable of mapping commercial chemicals to possible drug targets and vice versa, IDMap creates convenient environments for identifying the potential lead and its targets, especially in the field of drug repositioning.

Availability: IDMap executable and its user manual including color images are freely available to non-commercial users at http://www.equispharm.com/idmap

Contact: idmap@equispharm.com

1 INTRODUCTION

In an effort to increase productivity in the field of drug development, many companies in biopharmaceutical industry invest heavily in new drug discovery technologies, such as high throughput screening (HTS), genomics and combinatorial chemistry. On the other hand, by offering a better risk-versus-reward trade-off, drug repositioning—finding new uses of the existing drugs—is becoming one of the favorite strategies in the industry (Ashburn and Thor, 2004; O'Connor and Roth, 2005). The most successful example is sildenafil (Viagra; Pfizer). Originally, Pfizer developed sildenafil as an anti-angina medicine, but did not achieve the desired effect. Later, Pfizer researchers found an interesting side effect of sildenafil and repositioned its medical indication for male erectile dysfunction (Ashburn and Thor, 2004). The growing examples of repositioning include antidepressant drugs, neurological drugs and non-neurological drugs. (Ashburn and Thor, 2004). Then, how do you find these off-target effects that lead us to retargeting compounds? In its early stages, drug repositioning came from coincidental observation or platforms established to identify reprofiling opportunities (Ashburn and Thor, 2004). In contrast, Keiser and colleagues (2007) introduced similarity ensemble approach (SEA) algorithm which picks out the implicit relationships between 65 000 ligands and 246 known drug targets at a higher rate. The SEA groups biological targets according to the chemical similarity of their ligands and, therefore, facilitate drug repositioning. It is quite an evident example that such a computer-based prediction of the novel relationship between targets and chemicals before further laboratory experiment would be a cost effective and time saving tool that can increase efficiency at the early drug discovery stages.

We introduce a java-based software called IDMap, which is capable of repositioning marketed drugs to novel targets and vice versa. Additionally, it predicts potential bioactivities for commercial compounds on the basis of text mining and chemical structure information (cheminformatics). In other words, this enables one to map commercial chemicals as well as the existing drugs to possible drug targets and search the results in a very rapid manner. Furthermore, integrating bioactivity and chemical structure information, IDMap relieves researchers' burden of wasting time with querying multiple databases. With this software, researchers can generate new ideas for identifying potential lead and its targets in a highly efficient manner.

2 METHODS

IDMap is built by using the versions of Java from 1.5 through 1.6 under Eclipse Rich Client Platform (http://wiki.eclipse.org/index.php/Rich_Client_Platform). MySQL is used as a Database Management System (DBMS). Chemical data sets are obtained from Elsevier MDL Drug Data Report (MDDR) and the commercially available compound library from ASINEX (http://www.asinex.com). MDDR consists of about 160 000 drug-like chemicals annotated for 703 specific MDDR bioactivity classes including drug targets (MDDR: http://www.mdli.com/products/knowledge/drug_data_report/index.jsp). The commercially available compound library from ASINEX has about 400 000 compounds. Salts and fragments in a chemical structure were removed. Also chemical entries having no structure data were filtered out. In addition, by means of the program "PASS" (Prediction of Activity

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

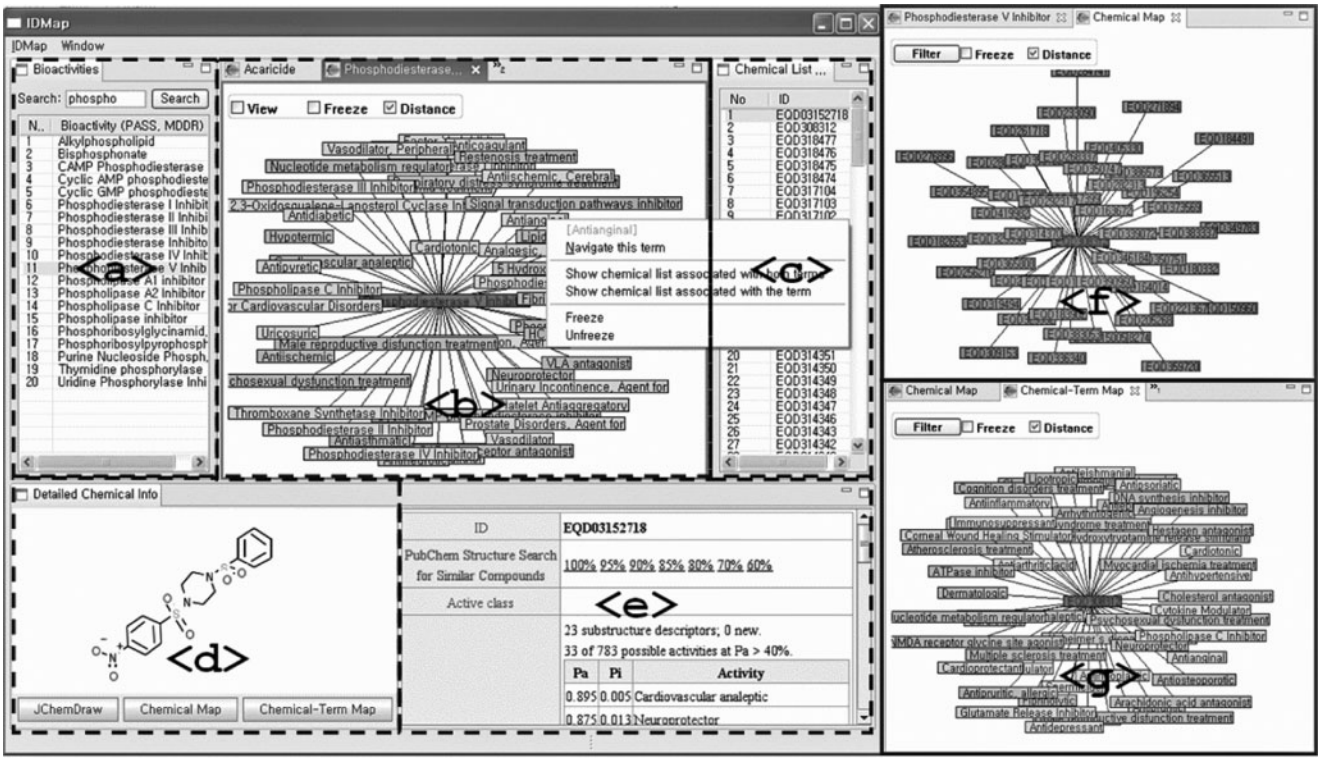

Fig. 1. IDMap Screenshot (**a**) Bioactivity class list table; (**b**) Interaction map viewer showing activity to activity, activity to chemical and chemical to chemical relationships; (**c**) Chemical List View: Chemicals related to selected activity/activities; (**d**) 2D structure for a given chemical; (**e**) Detailed information for a given chemical; (**f**) Screenshot of Chemical Map for EQD308312; and (**g**) Screenshot of Chemical-Term Map for EQD308312.

Spectrum for Substances), PASS-specific biological activities of all compounds are predicted and added to each chemical entry. PASS calculates concurrently several hundreds of biological activities (783 activities) using the chemical structures of compounds (Filimonov and Poroikov, 1996; Gloriozova *et al.*, 1998; Poroikov and Filimonov, 2001). Cerius^2® (http://www.accelrys.com/products/cerius2/) generated structural-physicochemical descriptors for each compound in the database. The principal component analysis (PCA) is used for the multi-dimensional mapping of all datasets, projecting chemicals into a scatter plot which illustrates the cluster of chemically similar chemicals. Thus each compound entry has a principal component representing its unique location in the chemical space. Additionally, the database is designed to facilitate the incorporation and processing of new chemical datasets. For MDDR compounds, only registry numbers (EXTREG) can be provided but not structures, due to license issues. Instead, if you have MDL MDDR database and license to use, EXTREG can be used to search structures in the MDDR Database. We do not allow any user to access or search the MDDR data directly. In case of ASINEX compound we open the chemical structure and users are freely able to edit its structure.

2.1 Identifying associations using text mining

The text mining technique (co-occurrence association) is used to mine hidden associations between bioactivities of chemical sets. Co-occurrence association is a widely accepted method because of its feature to extract relationships between any entities (Bowers *et al.*, 2004; Kuhn *et al.*, 2008; von Mering *et al.*, 2005). Because there were some redundant terms in both PASS and MDDR bioactivity class sets, we firstly removed redundant activity classes found in both class sets. After obtaining 1247 combined activity classes from PASS and MDDR activities, all pairs of activity terms within each chemical entry were compared with each other. The frequency of two activity terms occurring together is used to rank the relationship between bioactivities. The result of the co-occurrence association is a diagonally symmetric 2D matrix in which every element (bioactivity–bioactivity intersection) of the matrix is filled with co-occurrence frequency. Since each chemical is annotated with bioactivity classes, these relationships between bioactivities represented in the matrix intersection can be the mediators for finding new, unidentified relationships between chemicals and bioactivities and among chemicals.

2.2 Integrating chemical structure and bioactivity information

By combining the chemical structure data and the bioactivity data, researchers can draw more valuable information for identifying potential drug candidates and their targets. In IDMap, since each chemical is annotated with bioactivities and at the same time mapped in chemical space by chemical descriptors which represent physical and chemical features of its structure, it is also possible to identify structurally relevant compounds using the relationship between bioactivities. In other words, bioactivities play a role as mediators to integrate a great variety of biological and chemical data.

3 APPLICATION

Figure 1 shows an example that illustrates IDMap's usefulness. Double click the class 'Phosphodiesterase V inhibitor' in the bioactivity class list table, you will find that the interaction network related to the bioactivity class is displayed in the

interaction map viewer (Fig.1b). If you observe the length of the nodes' edges, you will notice that 'Antianginal' is one of the nodes attached nearest to the central 'Phosphodiesterase V inhibitor' node. Right clicking on the node 'Antianginal' displays popup menu. Next, select 'Select chemical list associated with both terms.' The list of all chemicals related to both Antianginal and Phosphodiesterase V inhibitor is displayed in the chemical list table (Fig. 1c). Clicking each row shows the detailed information of the chemical. After examining the EXTREG records of each chemical, we can find that the well-known phosphodiesterase V inhibitors, such as Sildenafil (Viagra), Lodenafil and many other potential candidates for drug repositioning, are detected in the chemical list table. Also, in the Detailed Chemical Information View, IDMap provide the link of NCBI's PubChem Structure Search for Similar Compounds (Wheeler *et al*., 2008). Users can access to the similar compounds of each MDDR and ASINEX chemical automatically at the pre-specified similarity thresholds. The link to PubChem also provides the diagnostic and therapeutic application of the chemicals. This example illustrates that IDMap can provide researchers with a network graph which presents linking relationships between bioactivities. Researchers can save time in their drug discovery with the aid of chemicals presented by the IDMap network. More comprehensive tutorial and the documentations of the IDMap are available at IDMap's website.

4 CONCLUSIONS

Constructing an association network between potential bioactivity classes by text mining technique, IDMap will be a help to researchers for generating new ideas for identifying potential drug candidates and their targets. In other words, IDMap predicts bioactivities including the targets of commercial compounds and therefore can ease the effort of finding leads. Furthermore, providing the integrated information of structural similarities and relevant bioactivities, IDMap increases the reliability of the search result and offers the ways to study the relationship between targets and leads from various angles. In conclusion, IDMap considerably improves the efficiency at the early drug discovery stages by integrating both structure-based and text-based data.

Conflict of Interest: none declared.

REFERENCES

Ashburn,T.T. and Thor,K.B. (2004) Drug repositioning: identifying and developing new uses for existing drugs. *Nat. Rev. Drug Discov.*, **3**, 673–683.

Bowers,P.M. *et al.* (2004) Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol.*, **5**, R35.

Filimonov,D.A. and Poroikov,V.V. (1996) PASS: Computerized prediction of biological activity spectra for chemical substances. In Ford, M.G., Greenwood, R., Brooks, G.T., and Franke, R. (eds) *Bioactive Compound Design: Possibilities for Industrial Use*. BIOS Scientific Publishers, Oxford, pp. 47–56.

Gloriozova,T.A. *et al.* (1998) Evaluation of computer system for prediction of biological activity PASS on the set of new chemical compounds. *Chim.-Pharm. J. (English translation by Consultants Bureau, New York: Pharmaceutical Chemistry Journal)*, **32**, 32–39.

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

Kuhn,M. *et al.* (2008) STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Res.*, **36**, D684–D688.

O'Connor,K.A. and Roth,B.L. (2005) Finding new tricks for old drugs: an efficient route for public-sector drug discovery. *Nat. Rev. Drug Discov.*, **4**, 1005–1014.

Poroikov,V.V. and Filimonov,D.A. (2001) Computer-aided prediction of biological activity spectra. Application for finding and optimization of new leads. In Holtje,H.-D. and Sippl,W. (eds) *Rational Approaches to Drug Design*. Prous Science, Barcelona, pp. 403–407.

von Mering,C. *et al.* (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms, *Nucleic Acids Res.*, **33**, D433–D437.

Wheeler,D.L. *et al.* (2008) Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, **36**, D13–D21.