BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 17 2011, pages 2465–2467
doi:10.1093/bioinformatics/btr413

Data and text mining

Advance Access publication July 12, 2011

RxnFinder: biochemical reaction search engines using molecular structures, molecular fragments and reaction similarity

Qian-Nan Hu[1,*], Zhe Deng[1], Huanan Hu[2,3], Dong-Sheng Cao[4] and Yi-Zeng Liang[4]

[1]Key Laboratory of Combinatorial Biosynthesis and Drug Discovery (Wuhan University), Ministry of Education, Wuhan University School of Pharmaceutical Sciences, Wuhan 430071, [2]School of Psychology, Southwest University, Chongqing 400715, [3]Normal School, Tibet University, Lhasa 850000 and [4]Research Center of Modernization of Traditional Chinese Medicines, Central South University, Changsha 410083, P. R. China

Associate Editor: Jonathan Wren

ABSTRACT

Summary: Biochemical reactions play a key role to help sustain life and allow cells to grow. RxnFinder was developed to search biochemical reactions from KEGG reaction database using three search criteria: molecular structures, molecular fragments and reaction similarity. RxnFinder is helpful to get reference reactions for biosynthesis and xenobiotics metabolism.

Availability: RxnFinder is freely available via: http://sdd.whu.edu.cn/rxnfinder.

Contact: qnhu@whu.edu.cn

Received on April 2, 2011; revised on July 5, 2011; accepted on July 6, 2011

1 INTRODUCTION

Diverse biochemical reactions participate in various important biosynthesis and metabolic pathways. Those reactions are manually curated and stored in several widely used databases (Croft *et al.*, 2011; Jennen *et al.*, 2010; Kanehisa *et al.*, 2008; Reitz *et al.*, 2004). The KEGG reaction database (Kanehisa *et al.*, 2008) contains >8000 reactions. The KEGG reaction database search tools include SIMCOMP and SUBCOMP (Hattori *et al.*, 2003, 2010), reaction ID, name, reactant entry, pathway and enzyme. In the present work (RxnFinder), three additional search engines are developed to help researchers retrieve KEGG reactions.

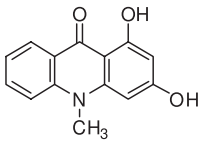
2 DESCRIPTION

In RxnFinder, users can search KEGG reactions using molecular structures, molecular fragments and reaction similarity for different purposes.

2.1 Search reactions with a molecular structure

A molecule matching tool was developed to find reactions with a specific molecule. When a query molecule is input using molecule SMILES string (Weininger, 1988; Weininger *et al.*, 1989), KEGG reactions with the molecule will be listed one by one. The algorithm used in RxnFinder is a string comparison based on canonical SMILES string (Weininger *et al.*, 1989). First, the canonical SMILES strings are calculated and stored in a database for all molecules in KEGG reactions. Secondly, the canonical SMILES string is computed for a query molecule. Thirdly, if the SMILES string of query molecule is matched with a SMILES string in the database, they are regarded as the same molecule. Then, reactions with the molecule will be retrieved. An example on searching biosynthesis reaction with 1, 3-dihydroxy-N-methylacridone is shown in Figure 1. In this search engine example, molecular canonical SMILES strings [for example, Oc1cc(O)c2c(c1)n(C)c1ccccc1c2=O for 1, 3-dihydroxy-N-methylacridone, a metabolite in alkaloids biosynthesis] are applied to both KEGG reaction molecules and the input molecule. In this case, there are four reactions (R07250, R08470, R08471 and R08472) obtained. One (R07250) of the reactions is listed in Figure 1, in which 1, 3-dihydroxy-N-methylacridone is located on the product side.

input: molecule (1,3-dihydroxy-N-methylacridone)
SMILES string: Oc1cc(O)c2c(c1)n(C)c1ccccc1c2=O
structure:



output: reactions with the molecule
4 reactions retrieved: R07250, R08470, R08471 and R08472.
KEGG reaction R07250:

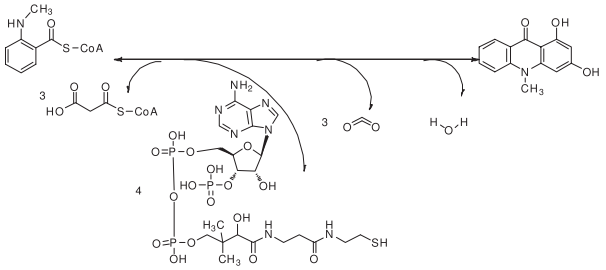


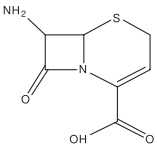
Fig. 1. Examples on searching reactions with a query molecule.

2.2 Search reactions containing a specific molecular fragment

Users can search reactions with a specific molecular fragment. The key of this engine is substructure search (Xu, 1996). Substructure

*To whom correspondence should be addressed.

input: molecular fragment (7-amino-3-cephem-4-carboxylic acid)
SMILES string: NC1C(=O)N2C1SCC=C2C(=O)O
structure:



output: reactions with the fragment
13 reactions retrieved: R03062, R03063, R03064, R04281, R05228, R05229, R05230, R05301, R05302, R05303, R07400, R07401 and R07402.
KEGG reaction R03062:

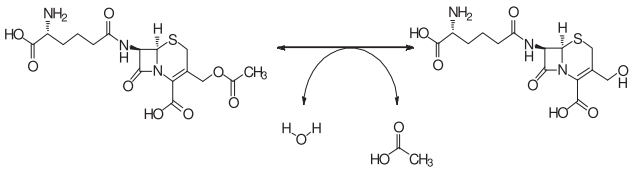


Fig. 2. Examples on searching reactions with a molecular fragment.

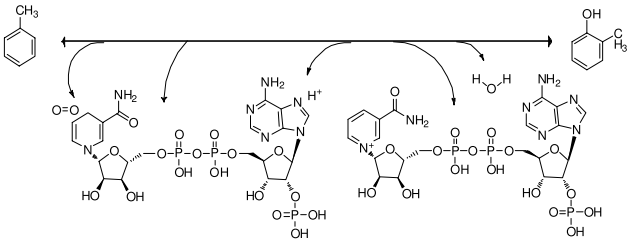
search is in essence a graph comparison algorithm (Hattori *et al*., 2003, 2010). Figure 2 exemplifies searching reactions with 7-amino-3-cephem-4-carboxylic acid, an important structure core of antibiotics. After the SMILES string of the molecular fragment [for instance, NC1C(=O)N2C1SCC=C2C(=O)O of 7-Amino-3-cephem-4-carboxylic acid] is input, reaction molecules will be scanned to identify if any molecule contains the specific fragment. Eventually, there are 13 reactions retrieved. One (R03062) of the reactions is listed in Figure 2, in which 7-amino-3-cephem-4-carboxylic acid can be found on both reactant and product sides.

2.3 Search similar reactions

We propose to search reactions using reactions similarity to find reactions with the same kind of chemical transformation. In this engine, reaction similarity is calculated using reaction difference fingerprints (RDF). RDF is computed using molecular fingerprints. The molecular fingerprint (Swamidass and Baldi, 2007) of a molecule is defined as $M_{FP} = (F_i)$, in which F_i refers to a molecular fragment with real occurrences in a molecule. The fingerprints of reactant molecules minus the fingerprints of product molecules will generate RDF as R_{FP} = (RF_i), in which RF_i represents the difference of a fragment occurring on reactants and products. After calculating RDF of two reactions, the similarity of two reactions can be computed using a Euclidean distance measurement as defined as $D_{i,j}$ = ED($R_{FP,i}$, $R_{FP,j}$). The smaller the distance between two reactions, the more similar they are.

Using the reaction similarity measurement, reactions similar to an input reaction can be retrieved from KEGG reaction database. An example is show in Figure 3. In this example, researchers are seeking for reactions similar to a metabolic reaction making Toluene to be a more soluble compound. After the SMIRKS string (Leach *et al.*, 1999) of a reaction is input, 148 reactions with distance 0 to the input reaction can be retrieved from KEGG reaction database. The distance 0 means the retrieved reaction holds the same chemical transformation to the input reaction. For example, phenol oxidoreductase (KEGG R00815) is retrieved. From the comparison of the two reactions, they hold the same kind of chemical

input: reaction (Toluene oxidoreductase)
SMIRKS string (Leach *et al.,* 1999):
Cc1ccccc1.O=O.O[C@@H]1[C@H](O)[C@@H](COP(=O)(O)OP(=O)(O)OC[C@H]2O[C@H]([C@H](OP(=O)(O)O)[C@@H]2O)n2cnc3c(N)ncnc23)O[C@H]1N1C=CCC(=C1)C(=O)N.[H+]>>Cc1ccccc1O.O[C@@H]1[C@H](O)[C@@H](COP(=O)(O)OP(=O)(O)OC[C@H]2O[C@H]([C@H](OP(=O)(O)O)[C@@H]2O)n2cnc3c(N)ncnc23)O[C@H]1[n+]1cccc(c1)C(=O)N.O
reaction structure:



output: similar reactions
148 reactions (with distance 0 to the input reaction) retrieved: R00815 R01142 R01143 R01295 R01296 R01298 R01628 R01960 R02178 R02253, *et al.* (The first 10 reactions are listed here. The full list is displayed on the web page after users input the reaction SMIRKS string.)
KEGG reaction R00815:

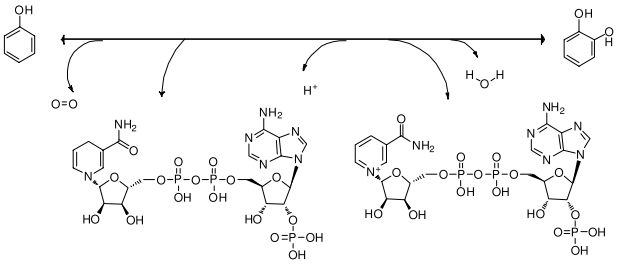


Fig. 3. Examples on searching reactions using reaction similarity.

transformation. Both reactions utilize O=O to add a functional group –OH.

3 DISCUSSION

KEGG provides two ways to search reactions with a molecule structure: KEGG_Mol_A using compound ID (http://www.genome.jp/kegg/compound/); and KEGG_Mol_B using SUBCOMP (http://www.genome.jp/tools/subcomp/) or SIMCOMP (http://www.genome.jp/tools/simcomp/) (Hattori *et al*., 2003, 2010). The results obtained by RxnFinder are the same as KEGG_Mol_A. The difference is that the input in RxnFinder is molecular structure; however, the input in KEGG_Mol_A is compound ID. When molecular structure (such as c1ccccc1Br) is input, RxnFinder retrieves two reactions (R07066 and R07068) with c1ccccc1Br. KEGG_Mol_A gets the same two reactions if the compound ID (C11036) is input. However, users usually do not know the KEGG ID of a specific molecule, for instance C11036 for Bromobenzene. Then, if the molecule structure of Bromobenzene is input to KEGG search engines. KEGG_Mol_B using SUBCOMP will retrieve 30 reactions. KEGG_Mol_B using SIMCOMP will retrieve 571 reactions, in which most of them are similar to c1ccccc1Br, but not identical c1ccccc1Br. The reason might be that SIMCOMP and SUBCOMP are developed to search

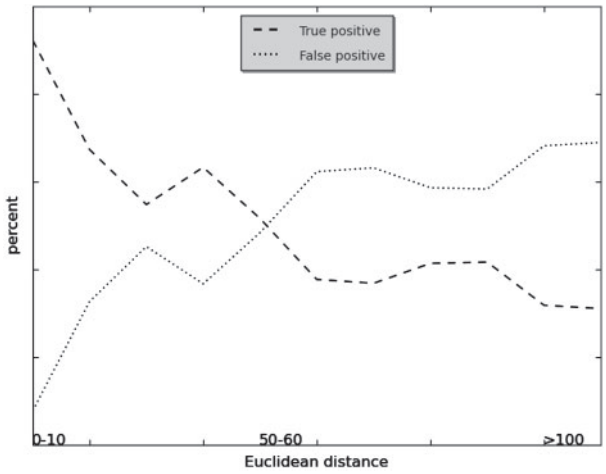


Fig. 4. Prediction performance along with Euclidean distance for statistics significance analysis. True or false positives mean that the retrieved reactions have or have not the same chemical transformations with the query reaction, respectively. We divided the scale of the *X*-axis into 11 parts: 0–10, 11–20 … >100, because some Euclidean distance values are not available for both 'true positive' and 'false positive'. The scale of the *X*-axis starts as 0–10.

similar molecules. In KEGG database, there are two reactions (R07066 and R07068) containing Bromobenzene (c1ccccc1Br).

For reaction search for a given molecular fragment, the algorithm used is a substructure match method (Xu, 1996). Our method can support SMART (http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html) SMILES string search. The implementations in KEGG are based on SUBCOMP algorithms (Hattori *et al.*, 2010). RxnFinder will retrieve 1855 reactions with Enamine fragment ([NX3][CX3]=[CX3], a SMART SMILES string for Enamine). KEGG SUBCOMP (http://www.genome.jp/tools/subcomp/) does not support SMART SMILES string search.

The novelty of this proposed reaction similarity is, at least to our knowledge, the first tool of its kind for retrieving biochemical reactions. The reaction similarity search is based on the Euclidean distance measurement of two reactions. After calculating reaction similarity between KEGG reactions and an input reaction, the KEGG reactions are retrieved based on the similarity rank. The distance values can range from 0 to infinity (theoretically). The distance 0 means the retrieved reaction holds the same chemical transformation to the input reaction. The direct prediction performance measurement of reaction similarity search is to check if the retrieved reactions are with the same chemical transformation pattern of the input reaction. Regarding to the statistical rigorousness of reaction similarity, we applied an indirect way to get raw similarity scores, and found distance <10 is a good threshold. The methodology is as follows: (i) the reactions with the same chemical transformation have the same EC numbers. (ii) Take one reaction out as query reaction, and use the most similar reaction to predict the EC number of the query reaction. (iii) By leave-one-out cross validation, we found that similarity scores (in this case, reaction Euclidean distance) within 10 will get satisfactory accuracy (>85%). (iv) The analysis is shown in Figure 4. In RxnFinder, distance 10 is now used as a statistics threshold to select reactions similar to an input reaction.

4 CONCLUSION

RxnFinder presents a collection of searching engines to search biochemical reactions from KEGG reaction database. With the engines at hand, researchers can get reference reactions for further analysis. RxnFinder also includes online tools to input molecules and reactions, calculate molecular fragments, compute reaction difference fingerprints and measure reaction distance.

ACKNOWLEDGEMENTS

The authors thank Peter Ertl (Novartis) for JME molecular editor and OpenBabel (O'Boyle *et al.*, 2008) for various Chemoinformatics functions used in web server.

Funding: National Natural Science Foundation of China (NSFC).

Conflict of interest: none declared.

REFERENCES

Croft,D *et al.* (2011) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.*, **39**, D691–D697.

Hattori,M. *et al.* (2003) Development of a Chemical Structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J. Am. Chem. Soc.*, **125**, 11853–11865.

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

Jennen,D.G. *et al.* (2010) Biotransformation pathway maps in WikiPathways enable direct visualization of drug metabolism related expression changes. *Drug Discov. Today*, **15**, 851–858.

Kanehisa,M. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36,** D480–D484.

Leach,A.R. *et al.* (1999) Implementation of a system for reagent selection and library enumeration, profiling, and design. *J. Chem. Inf. Comput. Sci.*, **39**, 1161–1172.

O'Boyle,N.M. *et al.* (2008) Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit. *Chem. Cent. J.*, **2**, 5.

Reitz,M. *et al.* (2004) Enabling the exploration of biochemical pathways. *Org. Biomol. Chem.*, **2**, 3226–3237.

Swamidass,S.J. and Baldi,P. (2007) Bounds and algorithms for fast exact searches of chemical fingerprints in linear and sublinear time. *J. Chem. Inf. Model.*, **47**, 302–317.

Weininger,D. (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, **28**, 31–36.

Weininger,D. *et al.* (1989) SMILES. 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.*, **29**, 97–101.

Xu,J. (1996) A Generic match algorithm for structural homomorphism, isomorphism, and maximal common substructure match and its applications. *J. Chem. Inf. Comput. Sci.*, **36**, 25–34.