*Bioinformatics*, 31(22), 2015, 3712–3714
doi: 10.1093/bioinformatics/btv416
Advance Access Publication Date: 17 July 2015
Applications Note



Systems biology

# RxnSim: a tool to compare biochemical reactions

**Varun Giri[1], Tadi Venkata Sivakumar[1], Kwang Myung Cho[2], Tae Yong Kim[2,*] and Anirban Bhaduri[1,*]**

[1]Bioinformatics Lab, Samsung Advanced Institute of Technology, Bangalore 560037, India and [2]Biomaterials Lab, Materials Center, Samsung Advanced Institute of Technology, Gyeonggi-do 443803, Korea

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on April 6, 2015; revised on July 13, 2015; accepted on July 13, 2015

## Abstract

**Summary:** Quantitative assessment of chemical reaction similarity aids database searches, classification of reactions and identification of candidate enzymes. Most methods evaluate reaction similarity based on chemical transformation patterns. We describe a tool, RxnSim, which computes reaction similarity based on the molecular signatures of participating molecules. The tool is able to compare reactions based on similarities of substrates and products in addition to their transformation. It allows masking of user-defined chemical moieties for weighted similarity computations.
**Availability and implementation:** RxnSim is implemented in R and is freely available from the Comprehensive R Archive Network, CRAN (http://cran.r-project.org/web/packages/RxnSim/).
**Contact:** anirban.b@samsung.com or ty76.kim@samsung.com
**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

Reaction similarity is used to classify biochemical reactions (Egelhofer *et al.*, 2010), assess enzyme similarity (O'Boyle *et al.*, 2007), plan molecular syntheses (Gasteiger *et al.*, 1992), search reaction databases (Hu *et al.*, 2011) and annotate new function to enzymes (Carbonell and Faulon, 2010). These methods were developed analogous to molecular similarity (Johnson and Maggiora, 1990; Willett *et al.*, 1998) wherein presence of key molecular signatures is encoded as bits in a feature vector and compared to compute similarity. For reaction similarity, signatures of transformation pattern are encoded into a reaction fingerprint (or feature vector) and used to compute similarity (de Groot *et al.*, 2009; Carbonell and Faulon, 2010; Rahman *et al.*, 2014; Schneider *et al.*, 2015; *ChemAxon Docs*; *Daylight Theory Manual*).

Present approaches evaluate reaction similarity by comparing transformation patterns. Comparison with respect to substrates requires additional assessment. RxnSim evaluates reaction similarity by comparing molecular signatures of individual molecules involved in the reactions. Molecular signatures implicitly encode transformation patterns. Accuracy of such approaches is limited by presence of large chemical moieties, such as ATP (de Groot *et al.*, 2009; Rahman *et al.*, 2014). To overcome this, RxnSim provides methods to mask user-defined moieties for weighted similarity computations.

## 2 Methods

RxnSim is designed to compute biochemical reaction similarity by comparing molecular signatures of input reactions. Each reaction is associated with a set of binary fingerprint vectors, denoted by $F$, derived from signatures of constituent molecules. This set is constructed using three methods capturing molecular signatures at different levels of granularity. The first method extracts molecular signatures at a level of individual molecules. The set of fingerprints, $F_M$, contains a fingerprint vector corresponding to each molecule in the reaction. $F_M = \{f_a,\ \forall a \in S; f_b, \forall b \in P\}$, $f_a$ represents fingerprint of molecule $a$, $S$ and $P$ represent the substrate set and product set of reaction, respectively. The second method constructs a set, $F_{S,P}$, containing two fingerprint vectors, one each for the substrate and product side. Each fingerprint records molecular signatures obtained from all molecules that are part of the substrate and product set,

respectively. $F_{S,P} = \{\sum f_a, \forall a \in S; \sum f_b, \forall b \in P\}$, where, $\sum$ denotes addition (binary OR) of fingerprint vectors. The third and most granular set, $F_R$, contains only one fingerprint capturing molecular signatures from all the molecules in the reaction. $F_R = \{\sum f_a, \forall a \in S, P\}$. RxnSim uses these sets of fingerprint vectors to compute reaction similarity based on four algorithms, namely, *msim*, *msim_max*, *rsim* and *rsim2*. *msim* and *msim_max* infer reaction similarity with respect to individual molecules and use $F_M$ for similarity computations. *rsim* is based upon similarities of the substrate and product sides while *rsim2* considers the reaction as a whole for similarity assessment. They use $F_{S,P}$ and $F_R$ for similarity computations, respectively.

*msim* pairs each substrate (product) in one reaction with a substrate (products) in the other based on pair-wise molecular similarity values (computed using similarity metric such as Jaccard coefficient) using a greedy algorithm. Reaction similarity is computed by averaging molecular similarity values of identified pairs while considering a '0' value for each unpaired molecule. A modified version of this algorithm, *msim_max*, ignores unpaired molecules while computing the reaction similarity. *rsim* computes reaction similarity as the average of the substrate and product side similarities. *rsim2* directly uses reaction fingerprints to compute the reaction similarity. For reversible reactions, the substrate and product sides are also cross compared and the maximum of the two similarity values is reported.

For brevity discussion in the text is limited to use of fingerprints. The method may be extended to user defined feature-vectors.

## 3 Features and description

RxnSim is implemented as an R package (R Core Team, 2015, http://www.R-project.org/). It imports chemoinformatic functionality from the rcdk and fingerprints packages (Guha, 2007). Input reactions are accepted in the form of reaction SMILES (RSMI) or MDL RXN files. The tool can compute similarity between a pair of reactions or list(s) of reactions that are provided as input. It implements fingerprint caching to reduce computation time on large datasets.

The tool provides four algorithms to compute reaction similarity, viz., *msim*, *msim_max*, *rsim* and *rsim2*. Performance of these algorithms was measured based on their ability to identify similar reactions (following method of Rahman *et al.*, 2014). Reactions sharing similar EC number at level 3 were considered similar. All-by-all comparison of 4828 reactions (with EC number assigned) drawn from the Rhea database (v.59, Morgat *et al.*, 2015) was performed. Distributions of reaction similarities (Fig. 1a) show a median score of ~0.2 for *msim* and *rsim* algorithms. *msim* shows a prediction accuracy of more than 95% for a cut-off of 0.55 and the area under ROC curve (AUC) was found to be 0.90 (Fig. 1b). The shape of the distribution and the accuracy for *msim* is comparable to that of the 'reaction center metric' of EC-BLAST (Rahman *et al.*, 2014). AUCs for *rsim* and *rsim2* were found to be 0.84 and 0.74, respectively. *rsim* and *rsim2* are more useful for analysis of datasets that are known to have common features, such as, catalysis by same enzyme or having same transformation.

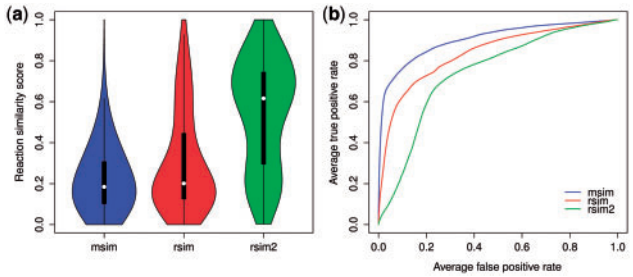


**Fig. 1.** Comparison of *msim*, *rsim* and *rsim2* algorithms implemented in RxnSim. (**a**) Density plots of reaction similarity scores. (**b**) Receiver operator characteristic (ROC) plot showing effectiveness of three algorithms (AUC order msim>rsim>rsim2)

The nature of assessment and data govern the choice of fingerprints and similarity metrics (Todeschini *et al.*, 2012; Riniker and Landrum, 2013). RxnSim allows various customizations for similarity computations as described below. It also allows choice of fingerprints and similarity metric. Within the tool, 12 fingerprints (10 bit and 2 count feature vectors) and 20 similarity metrics are provided (see Section S1 of supplementary materials).A detailed demonstration of the features is available through R's demo utility (section S2)

### 3.1 Substructure masking and reaction similarity

Molecules involving large moieties that are inert to reaction process (not participating in the reaction), may adversely influence the similarity values (de Groot *et al.*, 2009, Rahman *et al.*, 2014). An approach to overcome this is to reduce their weightage towards similarity computation. RxnSim implements a method that masks user-defined chemical substructures, provided as SMARTS or SMILES, present in a molecule. Section S3 discusses the impact of masking on similarity values, considering 50 molecules with Coenzyme A (CoA) moiety.

As reaction similarity is based on molecular signatures of individual molecules, this effect also percolates to reaction similarity (see Section S4). The presence of large cofactor molecules such as ATP, have similar effect on reaction similarity computations using the *rsim* and *rsim2* algorithms. Masking such molecules would have a similar effect on reaction similarity computations as seen for the molecular similarity (see Section S4).

### 3.2 Querying partial or unbalanced reactions

Enzyme identification for biochemical pathways often requires comparing partial or mass unbalanced reactions. Similarity computations based on the reaction transformation patterns usually require mass balanced reaction and are of limited use. *msim_max* computes reaction similarity based on the similarity of molecules in the input and is particularly designed to address such scenarios (see Section S5). *rsim* and *rsim2* may also be used along with cofactor masking.

### 3.3 User-defined reaction database

RxnSim can compare input reactions against a customized database of biochemical reactions provided as a flat-file. The search on this database can also be restricted based on the EC number range. This feature is helpful in assessing alternate enzymatic pathways. The proposed tool provides a flat-file containing biochemical reactions extracted from the Rhea database (Morgat *et al.*, 2015).

## Acknowledgements

The authors acknowledge support provided by Samsung Advanced Institute of Technology. The authors also thank the anonymous reviewers for their feedback.

Conflict of Interest: none declared.

## References


Carbonell,P. and Faulon,J.L. (2010) Molecular signatures-based prediction of enzyme promiscuity. *Bioinformatics*, **26**, 2012–2019.

de Groot,M.J. *et al*. (2009) Metabolite and reaction inference based on enzyme specificities. *Bioinformatics*, **25**, 2975–2982.

Egelhofer,V. *et al*. (2010) Automated assignment of EC numbers. *PLoS Comput. Biol*., **6**, e1000661.

Gasteiger,J. *et al*. (1992) Similarity concepts for the planning of organic reactions and syntheses. *J. Chem. Inf. Comput. Sci.*, **32**, 700–712.

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

Hu,Q.N. *et al*. (2011) RxnFinder: biochemical reaction search engines using molecular structures, molecular fragments and reaction similarity. *Bioinformatics*, **27**, 2465–2467.

Johnson,A.M. and Maggiora,G.M. (1990) *Concepts and Applications of Molecular Similarity*. Wiley, New York.

Morgat,A. *et al*. (2015) Updates in Rhea—a manually curated resource of biochemical reactions. *Nucl. Acids Res*., **43**, D459–D464.

O'Boyle,N. *et al*. (2007) Using reaction mechanism to measure enzyme similarity. *J. Mol. Biol*., **368**, 1484–1499.

Rahman,S.A. *et al*. (2014) EC-BLAST: a tool to automatically search and compare enzyme reactions. *Nat. Methods*, **11**, 171–174.

Riniker,S. and Landrum,G.A. (2013) Open-source platform to benchmark fingerprints for ligand-based virtual screening. *J. Cheminform*., **5**, 26.

Schneider,N. *et al*. (2015) Development of a novel fingerprint for chemical reactions and its application to large-scale reaction classification and similarity. *J. Chem. Inf. Model*., **55**, 39–53.

Todeschini,R. *et al*. (2012) Similarity coefficients for binary chemoinformatics data: overview and extended comparison using simulated and real data sets. *J. Chem. Inf. Model*., **52**, 2884–2901.

Willett,P. *et al*. (1998) Chemical similarity searching. *J. Chem. Inf. Sci*., **38**, 983–996.