Database, 2020, 1–9
doi:10.1093/database/baaa065
Database tool

Database tool

EnzyMine: a comprehensive database for enzyme function annotation with enzymatic reaction chemical feature

Dandan Sun[1], Xingxiang Cheng[1], Yu Tian[2], Shaozhen Ding[1], Dachuan Zhang[1], Pengli Cai[1,3] and Qian-nan Hu[1,*]

[1]CAS Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200333, P. R. China, [2]School of Biology and Pharmaceutical Engineering, Wuhan Polytechnic University, Wuhan, Hubei 430023, China and [3]Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin 300308, P. R. China

[*]Corresponding author: Tel: +8618920700831; E-mail: qnhu@sibs.ac.cn

Citation details: Sun,D., Cheng,X., Tian,Y., *et al.* EnzyMine: a comprehensive database for enzyme function annotation with enzymatic reaction chemical feature. *Database* (2020) Vol. 00: article ID baaa065; doi:10.1093/database/baaa065

Received 27 July 2020; Revised 19 July 2020; Accepted 24 July 2020

Abstract

Addition of chemical structural information in enzymatic reactions has proven to be significant for accurate enzyme function prediction. However, such chemical data lack systematic feature mining and hardly exist in enzyme-related databases. Therefore, global mining of enzymatic reactions will offer a unique landscape for researchers to understand the basic functional mechanisms of natural bioprocesses and facilitate enzyme function annotation. Here, we established a new knowledge base called EnzyMine, through which we propose to elucidate enzymatic reaction features and then link them with sequence and structural annotations. EnzyMine represents an advanced database that extends enzyme knowledge by incorporating reaction chemical feature strategies, strengthening the connectivity between enzyme and metabolic reactions. Therefore, it has the potential to reveal many new metabolic pathways involved with given enzymes, as well as expand enzyme function annotation.

Database URL: http://www.rxnfinder.org/enzymine/

Introduction

Enzyme function annotation has excellent implications in metabolic engineering, synthetic biology and pathophysiology (1). Along with the rapid expansion of protein sequences, predicting enzymatic reactions of unannotated sequences using computational methodology is widely becoming used (2, 3). This function prediction contains enzyme feature extraction and classification optimization as two main procedures associated with machine learning and deep learning (4). Meanwhile, feature extraction is no longer limited to sequence similarity but includes more conservative features independent of sequence length (3),

Table 1. Detailed comparison of the content of databases

Database	Enzyme	Mechanism	Reaction	Chemical feature		
				Reaction centre	Reaction rule	Core-to-core
ExplorEnz (15)	√	—	—	—	—	—
EzCatDB (16)	√	√	√	—	—	—
BRENDA (17)	√	—	√	—	—	—
BNICE (24)	—	—	√	—	√	—
RetroRules (22)	—	—	√	—	√	—
EnzyMine	√	√	√	√	√	√

such as sequence patterns, structural features and catalytic mechanisms curated in PROSITE, CATH, SCOP and M-CSA (5–8). Sequence information curated in UniProt (9) acted as a vital resource for orphan sequence mining based on the assumption that similar sequences function consistently. In sequence-related features, PROSITE defines protein domains and sequence patterns of biological significance (6). Analysis of the structural domains generated by CATH reveals the prominent features of protein secondary structure (5). Enzyme 3D structural information from Protein Data Bank is also utilized to facilitate functional annotation beyond only sequence information (10). In terms of known structure, SCOP manually provides structural classification to reveal the evolutionary relationships between proteins (8). Structural information can help to identify binding sites and catalytic residues on proteins further. A specific collection of catalytic residues and detailed processes of catalytic mechanisms can be obtained from M-CSA (7). However, pieces of sequence- and structure-related feature information mentioned are scattered in each database, hindering the understanding and integration of enzyme annotation.

At the same time, chemical structure information dramatically improves the accuracy of model prediction and is considered indispensable for accurate prediction in the performance evaluation (11). Various annotation tools have applied reaction-based strategy to assign an Enzyme Commission (EC) number for reactions based on changes in the chemical structure of substrates and products (12–14). However, chemical structure information about enzymatic reactions is limited to compounds in reactions, lacking further feature mining. Currently, several integrated databases that provide detailed profiles about enzymes are available, each having a specific objective towards accelerating metabolic engineering and enzyme annotation. ExplorEnz offers a canonical curation of the International Union of Biochemistry and Molecular Biology (15). EzCatDB prompted a manual classification of enzyme reaction from the perspective of enzyme structure and catalytic mechanisms (16). BRENDA constructs a comprehensive enzyme information system with a focus on enzymatic reactions and relevant pathways (17). Based on enzyme and reaction databases, chemical transformation can be extracted, generalized for similar structural changes and applied to expand possible biosynthetic routes. BNICE defined generalized enzyme reactions for formulating enzyme reaction rules systematically; metabolic in silico network expansion and atlas of biochemistry were therefore depicted based on this computational framework (18–20). RetroRules took another approach to compute rules for available metabolic reactions that describe these reactions and can be plugged into RetroPath2.0 to design bioengineering pathways (21, 22). Reaction rule is gradually showing usefulness in exploring possible biosynthetic routes. However, these specific databases focus on reaction rules, lacking the interaction of enzyme information and place certain demands of chemical literacy on researchers. An enzymatic database with detailed enzyme information and intuitive chemical information extraction is urgently needed at present. A detailed comparison of the content of related databases is provided in Table 1.

To conquer the deficiencies mentioned earlier, we proposed EnzyMine, a comprehensive enzyme feature and annotation database. First, to describe the diversity of enzyme data and provide a comprehensive enzyme feature, we collected sequence-structure-related features and catalytic features scattered in the above databases. Next, we replenished a systematic reaction feature mining of chemical structure. Reaction features are expanded beyond reaction compounds to further include reaction centre, reaction rule and core-to-core analysis using our previous work, Rxnblast (23). Core-to-Core is defined as the scaffold transformation made up of atoms that change during the reaction. In particular, core-to-core analysis can show the unique and complex structural scaffold change in an enzymatic reaction, helping to demonstrate the role of the enzyme on the molecular scaffold. This chemical feature mining can intuitively aid the understanding of chemical composition changes and reaction patterns.

In this report, we mined 7767 EC numbers and displayed these enzymes with comprehensive sequence and structural feature visually in 267 345 protein sequences across 8058 organisms, along with deep calculated chemical features in 9831 reactions. This study introduces the process of reaction feature mining and analysis, molecule structure-based searching methods and makes the following contributions: (i) curation of scattered enzyme sequence, structure, catalytic data; (ii) chemical feature mining and analysis of enzymatic reactions and (iii) efficient searching methods based on chemical structure query and text query.

Materials and methods

Data collection and database content

Using EC numbers in EXPASY release (2020/2/27) (ftp://ftp.expasy.org/databases/enzyme/enzyme.dat), EnzyMine displays 7767 enzymes with complete sequences, structural data and vital interface links to referenced databases. The databases include information on enzyme function, sequence information and family classification, as well as protein structure, molecular structure, known reactions and associated literature. Sources and the amount of available data are shown in Table 2.

Database architecture

EnzyMine is built on the Python web development framework, Django, with information stored in the PostgreSQL relational database. The front-end of the site is designed with HTML, CSS, Bootstrap, jQuery and JavaScript. Fuzzy search methods are completed by Haystack (http://haystacksearch.org/), an open-source and integrated full-text search engine in Django. Unique molecule searching methods are completed by RDKit (http://www.rdkit.org/), an open-source chemical informatics and machine learning kit that provides C++ and Python API interface. The architecture of the process is shown in Figure 1.

The algorithm in query methods

Tanimoto coefficient performs remarkably well for similarity-based virtual screening by comparing chemical structures and is an appropriate choice for quantifying molecular similarity calculations. In particular, the Tanimoto coefficient-based similarity algorithm is used to construct a similarity retrieval method (24). Maximum common substructure (MCS) searching is established on the fMCS algorithm (25), providing a flexible and alternative powerful way to search by calculating the distance between two molecules. MCS percentage can be represented as a mathematical equation, T_{MCS} (A, B):

$$T_{MCS}\left(A,B\right)=\frac{\left|MCS\left(A,B\right)\right|_a}{\left|A\right|_a+\left|B\right|_a+\left|MCS\left(A,B\right)\right|_a} \qquad (1)$$

In this equation, |A|a, |B|a represent the number of atoms for the query molecule and another molecule in our database, respectively. |MCS (A, B)|a, in this case, is the number of atoms in the MCS between the two molecules.

Chemical feature extraction in an enzymatic reaction

The chemical feature mining strategy in EnzyMine is to perform regular extraction of enzymatic reactions, making great efforts in feature visualization, including reaction centre and scaffold transformation. In EnzyMine, we adopt SMARTS (SMILES arbitrary target specification), an expanded version of core-SMILES (simplified molecular-input line-entry system), as a code for reaction analysis (https://www.daylight.com/dayhtml/doc/theory/). Generally, atoms that change within a reaction are described as reaction centres, which are equivalent to atoms attached

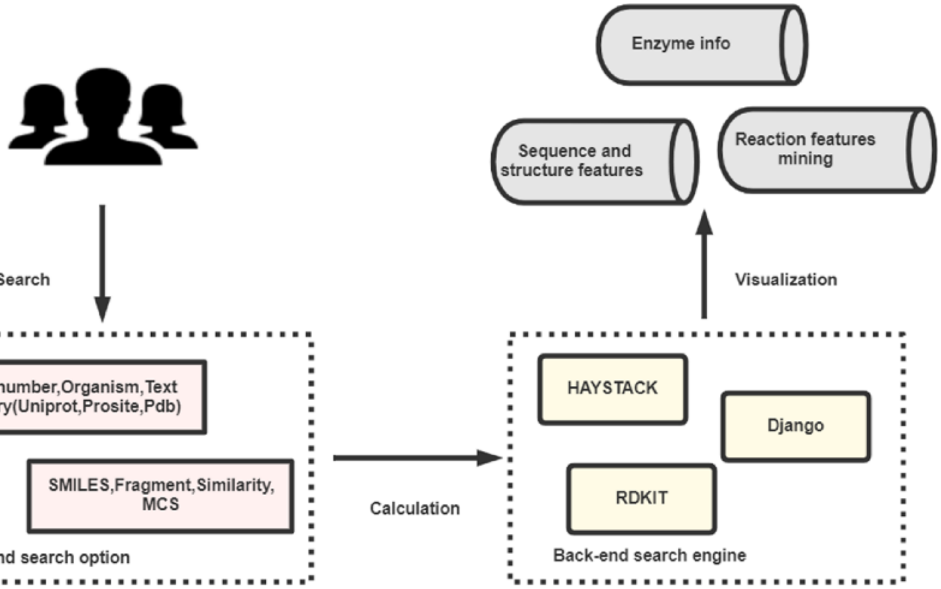
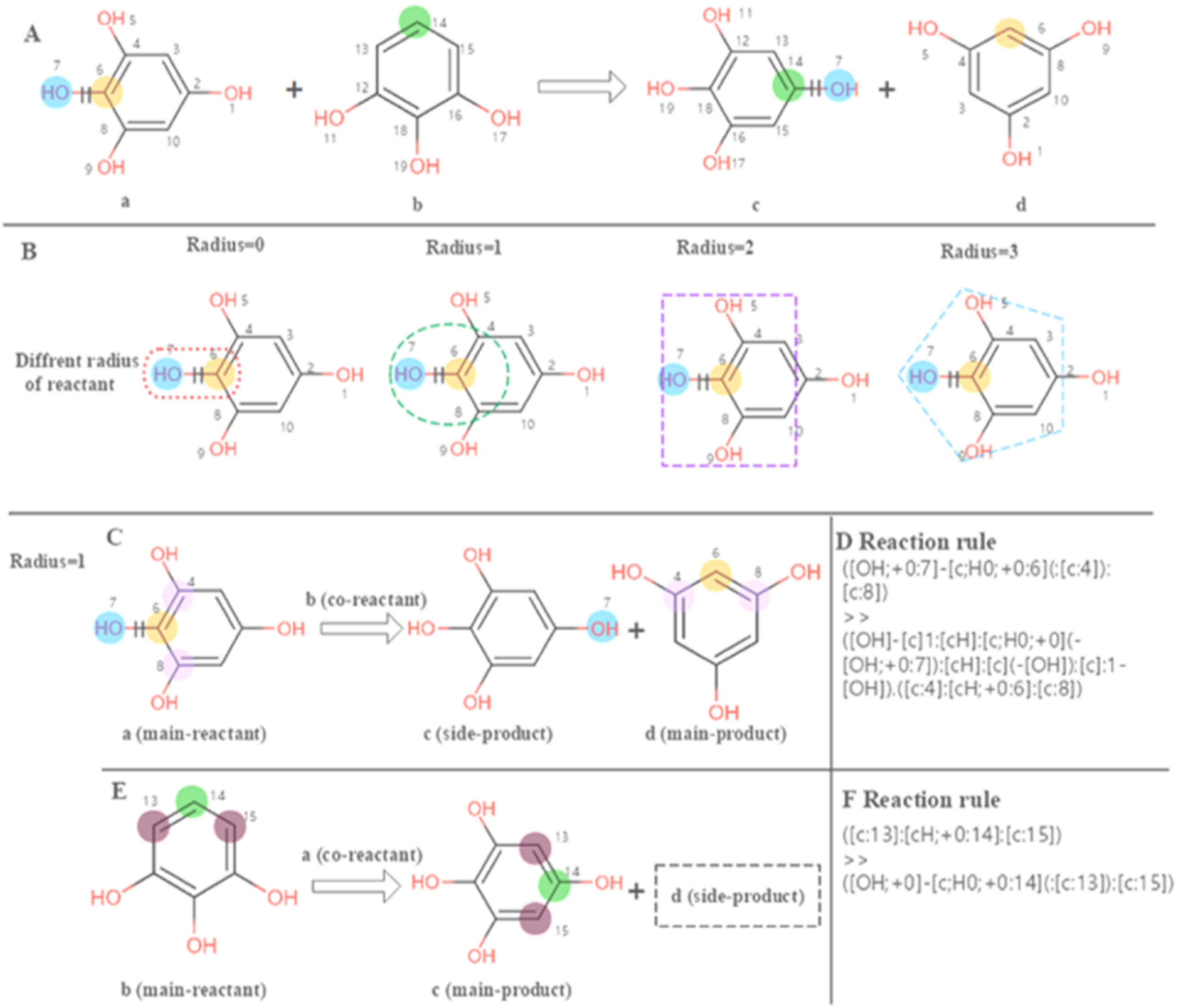


Figure 1. The broad architecture of EnzyMine processing, including the front-end and back-end search engines, which return displayed data.



```
([OH;+0:7]-[c;H0;+0:6](:[c:4]):
[c:8])
>>
([OH]-[c]1:[cH]:[c;H0;+0](-
[OH;+0:7]):[cH]:[c](-[OH]):[c]:1-
[OH]).([c:4]:[cH;+0:6]:[c:8])
```

```
([c:13]:[cH;+0:14]:[c:15])
>>
([OH;+0]-[c;H0;+0:14](:[c:13]):[c:15])
```

Figure 2. Calculation of the reaction rules for EC 1.97.1.2. (A) The reaction transformation is labelled. (B) Different radius can be used to identify adjacent neighbours in the reaction. (C, E) The reaction rules are calculated for two reactants. (D, F) reaction rules encoded with SMART.

to the bonds that are broken or formed (21). Reaction rule summarizes the overall atom-bond pattern change taking place in the reaction and can be different according to the radius taken around the reaction centre (26). We used the RxnBLAST tool to extract the characteristics of reaction environment, including reaction centre, reaction rules and core-to-core, which focuses on extracting scaffold transformation and reactive chemical environment features by analysing atom-atom mapping (23). Our approach for visualizing the reaction centre and reaction rules is described in Figure 2. This generative process is divided into three parts. (1) The reaction centre is generated from atom-atom mapping. (2) Adjacent neighbours are characterized to the reaction centre using different radius; here, we choose a radius of 1 for specificity. (3) For multi-substrates/products reactions, reaction rules are extracted by assigning each reactant as the main reactant molecule at a time to limit the combinatorial complexity.

Results

Database overview

EnzyMine displays 7767 enzyme items complemented with sequences, structural data, family classification, as well as molecular structure, known reactions and associated literature. Different coverage of EC number under each database is described in Figure 3A. With the rapid emergence of sequence data and improvement of sequencing methods, all the enzyme numbers are covered by sequencing data, while reaction data and family classification also reach a high coverage. At the same time, precise 3D structural data of enzymes still need to be revealed further by experimental studies.

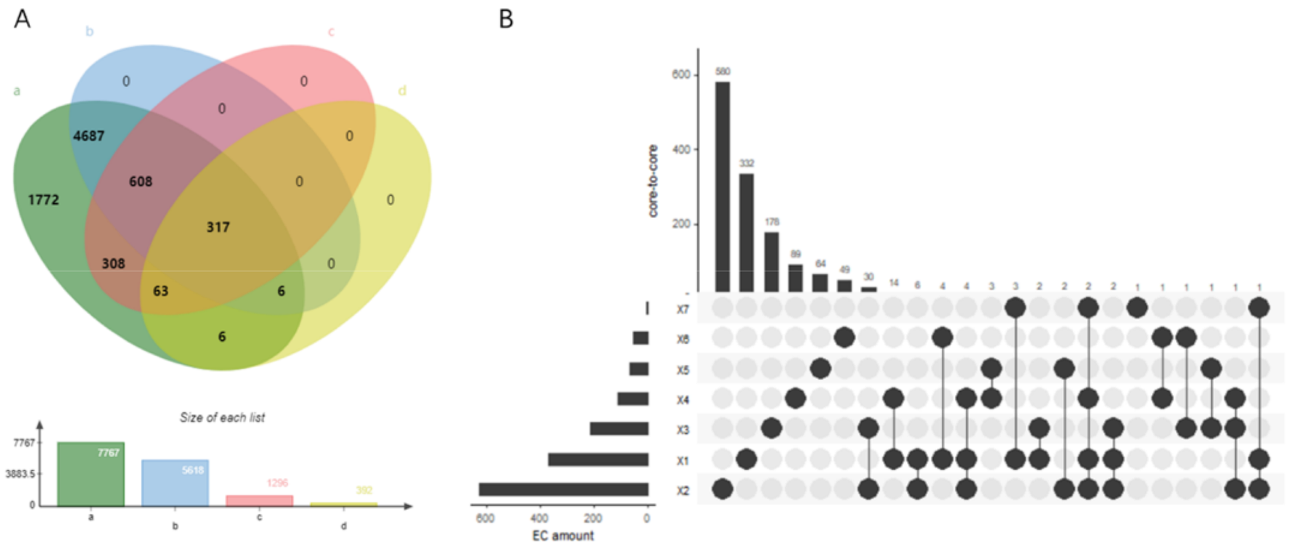


Figure 3. (A) Wynn diagram of four database coverage of EC number, a: sequence amount in UniProt, b: reaction amount in Rhea, c: family classification amount in PROSITE, d: 3D structure amount in Protein Data Bank. (B) The overall distribution of core-to-core across the seven types of enzyme.

At the reaction level, 9831 intuitive display results are found for the centre of enzyme-catalysed reactions, along with the extracted reaction rules and core-to-core characterization. In addition, in order to meet user's design requirements of retrosynthesis, the website provides online calculation of reaction rules of different radii. Seven types of core-to-core distribution are shown in Figure 3B. In the distribution diagram, type two transferases react with a wide range of substrates, having the largest amount of core-to-core types. In addition, a small amount of core-to-core interaction can occur in the reaction across different types of enzymes, and this phenomenon occurs most frequently between the type two transferases and type three hydrolases.

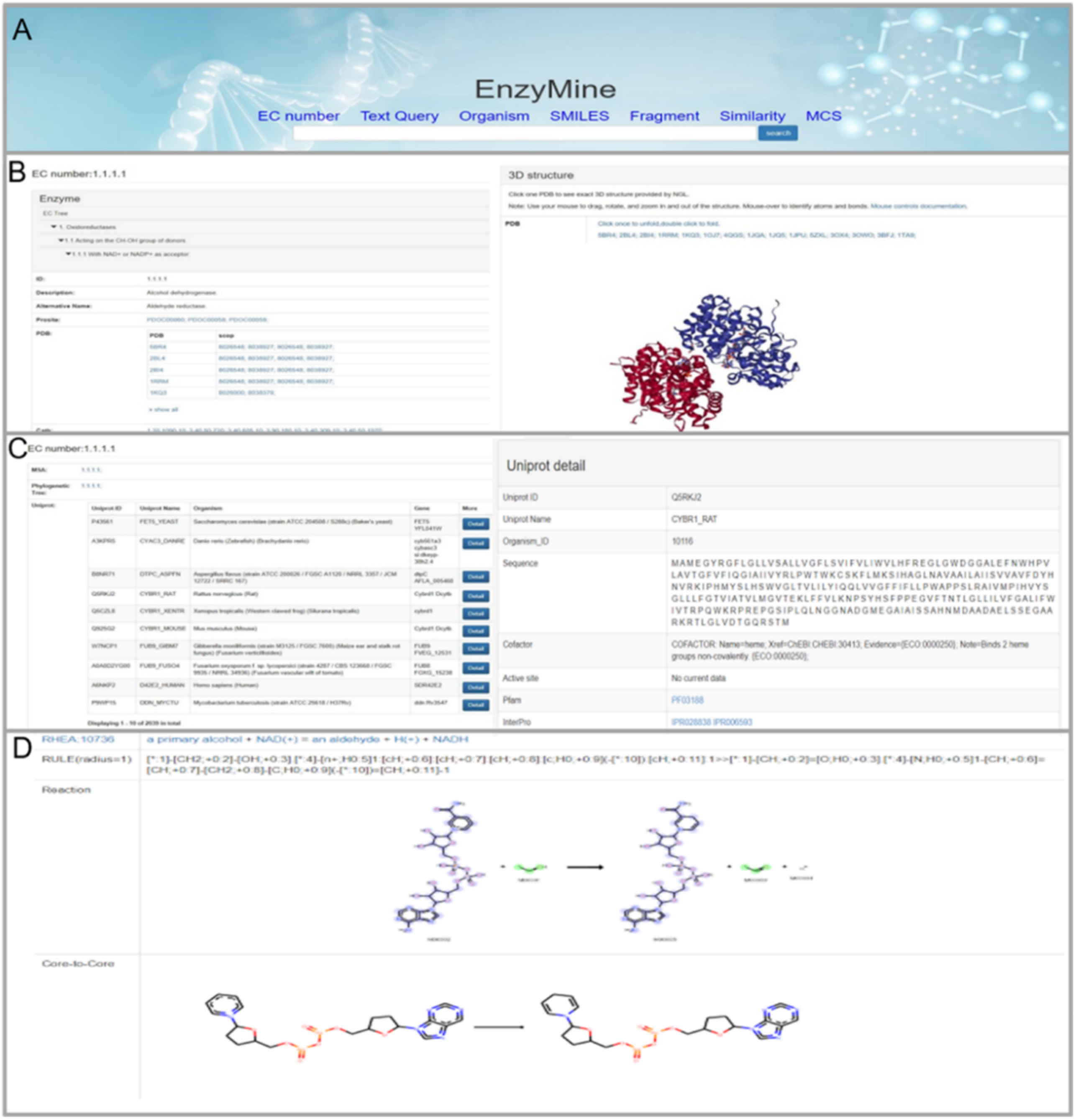
Search results and web usage

At the homepage, EnzyMine provides various search methods based on text and molecular retrieval algorithms When a user enters a query related to an ambiguous enzyme name and organism, the Haystack engine will show all possible results. Additionally, similarity and MCS searches for molecular queries will mainly return a series of outputs with similarity score/MCS percentage ranging from 1.0 or less. The detailed result is calculated using Equation (1). Compounds in the existing reactions can be searched by chemical structure or sequence similarity, and MCS processing provides metabolic results for many more molecules. At the all data page, we have listed the latest changes associated with each piece of data, such as the deletion of item 5.3.99.1, the transfer of 5.3.3.15 to 5.3.2.7 and the continuous increase of seven enzymes. Selecting an organism will return enzymes specific to that organism.

These powerful search methods and result examples are shown in Figure 4A. On the page with details, the result is divided into three parts: Basic info, sequence analysis and reaction feature analysis. First, the basic info part displays enzyme information, 3D structures and related references, as shown in Figure 4B. Second, in sequence analysis, EnzyMine lists UniProt data in all organisms and function patterns in PROSITE under the EC number, as shown in Figure 4C. Multi-sequence alignment and phylogenetic tree facilitated sequence analysis help to provide a better understanding. Finally, 9831 calculated reaction centres, rules and core-to-core are provided in corresponding reaction analyses, as shown in Figure 4D. For enzymes with elucidated catalytic mechanisms in the sequence, we collected catalytic mechanisms from M-CSA, each connected to enzymatic reaction features shown by Sankey charts/Force-layout graphs. Force-layout graphs stress enzymatic catalytic features with active sites registered in UniProt and reaction under the EC number. Sankey diagrams offer more detail into the catalytic mechanism of each enzyme.

Core-to-core analysis

The number of reactions of seven enzymes and the analysis of core-to-core types are shown in Figure 5. The most frequent core-to-core type occurs in oxidoreductases, showing the process that NAD(+) takes hydrogen and turns into NADH in oxidoreductases. Transferases have the broadest range of the core-to-core type. Among these, the most numerous changes catalyse the transfer of glucuronic



EnzyMine

EC number:1.1.1.1

Enzyme

EC Tree

▼1. Oxidoreductases

▼1.1 Acting on the CH-OH group of donors

▼1.1.1 With NAD+ or NADP+ as acceptor

ID:	1.1.1.1
Description:	Alcohol dehydrogenase.
Alternative Name:	Aldehyde reductase.
Prosite:	PDOC00060; PDOC00058; PDOC00059;
PDB:	
Cath:	1.20.1090.10; 3.40.50.720; 3.40.605.10; 3.90.180.10; 3.40.309.10; 3.40.50.1970;

PDB	scop
5BR4	8026548; 8038927; 8026548; 8038927;
2BL4	8026548; 8038927; 8026548; 8038927;
2BI4	8026548; 8038927; 8026548; 8038927;
1RRM	8026548; 8038927; 8026548; 8038927;
1KQ3	8026000; 8038379;

» show all

3D structure

Click one PDB to see exact 3D structure provided by NGL.

Note: Use your mouse to drag, rotate, and zoom in and out of the structure. Mouse-over to identify atoms and bonds. Mouse controls documentation.

PDB Click once to unfold,double click to fold.

5BR4; 2BL4; 2BI4; 1RRM; 1KQ3; 1OJ7; 4QGS; 1JQA; 1JQ5; 1JPU; 5ZXL; 3OX4; 3OWO; 3BFJ; 1TA9;

EC number:1.1.1.1

MSA: 1.1.1.1;

Phylogenetic Tree: 1.1.1.1;

Uniprot:

Uniprot ID	Uniprot Name	Organism	Gene	More
P43561	FET5_YEAST	Saccharomyces cerevisiae (strain ATCC 204508 / S288c) (Baker's yeast)	FET5 YFL041W	Detail
A3KPR5	CYAC3_DANRE	Danio rerio (Zebrafish) (Brachydanio rerio)	cyb561a3 cybasc3 si:dkeyp-38h2.4	Detail
B8NR71	DTPC_ASPFN	Aspergillus flavus (strain ATCC 200026 / FGSC A1120 / NRRL 3357 / JCM 12722 / SRRC 167)	dtpC AFLA_005460	Detail
Q5RKJ2	CYBR1_RAT	Rattus norvegicus (Rat)	Cybrd1 Dcytb	Detail
Q6CZL8	CYBR1_XENTR	Xenopus tropicalis (Western clawed frog) (Silurana tropicalis)	cybrd1	Detail
Q925G2	CYBR1_MOUSE	Mus musculus (Mouse)	Cybrd1 Dcytb	Detail
W7NCP1	FUB9_GIBM7	Gibberella moniliformis (strain M3125 / FGSC 7600) (Maize ear and stalk rot fungus) (Fusarium verticillioides)	FUB9 FVEG_12531	Detail
A0A0D2YG00	FUB9_FUSO4	Fusarium oxysporum f. sp. lycopersici (strain 4287 / CBS 123668 / FGSC 9935 / NRRL 34936) (Fusarium vascular wilt of tomato)	FUB8 FOXG_15238	Detail
A6NKP2	D42E2_HUMAN	Homo sapiens (Human)	SDR42E2	Detail
P9WP15	DDN_MYCTU	Mycobacterium tuberculosis (strain ATCC 25618 / H37Rv)	ddn Rv3547	Detail

Displaying 1 - 10 of 2039 in total

Uniprot detail

Uniprot ID	Q5RKJ2
Uniprot Name	CYBR1_RAT
Organism_ID	10116
Sequence	MAMEGYRGFLGLLVSALLVGFLSVIFVLIWVLHFREGLGWDGGALEFNWHPV LAVTGFVFIQGIAIIVYRLPWTWKCSKFLMKSIHAGLNAVAAILAIISVVAVFDYH NVRKIPHMYSLHSWVGLTVLILYIQQLVVGFFIFLLPWAPPSLRAIVMPIHVYS GLLLFGTVIATVLMGVTEKLFFVLKNPSYHSFPPEGVFTNTLGLLILVFGALIFW IVTRPQWKRPREPGSIPLQLNGGNADGMEGAIAISSAHNMDAADAELSSEGAA RKRTLGLVDTGQRSTM
Cofactor	COFACTOR: Name=heme; Xref=ChEBI:CHEBI:30413; Evidence={ECO:0000250}; Note=Binds 2 heme groups non-covalently. {ECO:0000250};
Active site	No current data
Pfam	PF03188
InterPro	IPR028838 IPR006593

RHEA:10736	a primary alcohol + NAD(+) = an aldehyde + H(+) + NADH
RULE(radius=1)	[*:1]-[CH2;+0:2]-[OH;+0:3].[*:4]-[n+;H0:5]1:[cH;+0:6]:[cH;+0:7]:[cH;+0:8]:[c;H0;+0:9](-[*:10]):[cH;+0:11]:1>>[*:1]-[CH;+0:2]=[O;H0;+0:3].[*:4]-[N;H0;+0:5]1-[CH;+0:6]=[CH;+0:7]-[CH2;+0:8]-[C;H0;+0:9](-[*:10])=[CH;+0:11]-1
Reaction	
Core-to-Core	

Figure 4. Detailed EnzyMine screenshot and search results for EC:1.1.1.1 are shown. (A) Front-end search methods based on text query and molecular searching query. (B) Basic enzyme information is listed, where EC classification, PROSITE and CATH classifications are shown on the left. Various display options related to protein 3D structures are provided by the NGL engine (27). (C) Multi-sequence alignment and phylogenetic trees related to the UniProt sequence under each the EC number are provided in a link, and single UniProt details are included by clicking the button on the right. (D) Reaction centre display, rules and core-to-core are shown. Detailed enzyme information and feature mining information are available for download.

acid from Uridine disphosphate(UDP) to other (usually hydrophobic) molecules, which are quite common in membrane proteins to change the water solubility of the receptor molecules and promote the export of these molecules. In particular, a reaction with the rule may not have a core-to-core type. This phenomenon is relatively common in type 2 enzymes. For example, the reaction with Rhea ID 21824 belongs to transferase with EC number 2.6.1.1. It involves the conversion of side chains on the same skeleton, and core-to-core does not change before and after the reaction. All core-to-core data are available on the EnzyMine download page.

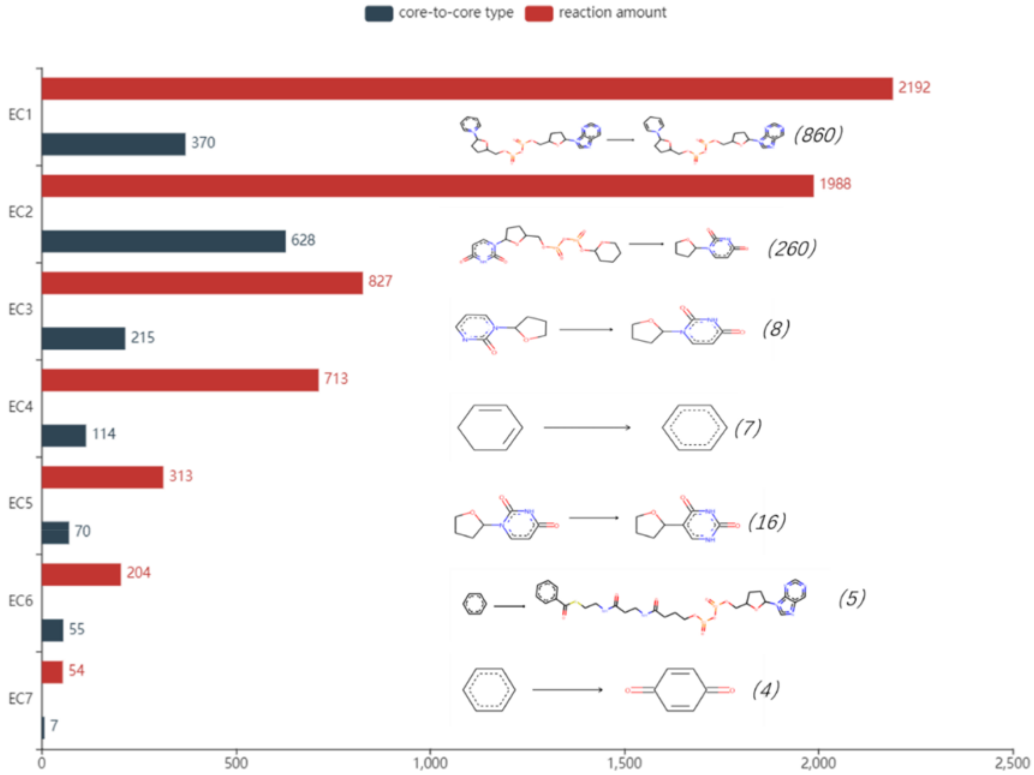


Figure 5. Detailed reaction amount and core-to-core type in seven types of enzymes, the most frequent amount marked in the parentheses and core-to-core pictures are listed on the right of the bar graph. EC1: Oxidoreductases, EC2: Transferases, EC3: Hydrolases, EC4: Lyases, EC5: Isomerases, EC6: Ligases, EC7: Translocases.

Table 2. Current contents of EnzyMine

Related data	Description	Number
EC number	Enzyme commission number	7767
UniProt	Protein sequence data from UniProt (9)	267 345
Organism	Number of species under EC number	8058
PROSITE	Protein domains and functional sites in PROSITE (6)	1450
CATH	Protein superfamily classification (5)	2017
SCOP	Structural classification of protein (8)	9514
PDB	Tertiary structure data from Protein Data Bank (10)	141 414
Catalytic mechanisms	Detailed catalytic mechanisms of an enzyme in M-CSA (7)	864
Compound data	Compound data for substrates/products/cofactors	27 765
Reaction data	Enzymatic reactions with EC number in Rhea (16), KEGG (17)	9831
Reference	Reference about Enzyme and reaction	39 833

Enzyme annotation integrations from sequence to reaction

Extended enzyme-number prediction method based on the core-to-core strategy is also available in EnzyMine as a simple application of those chemical feature data. EnzyMine can predict EC numbers in enzymatic reactions using the logic that different molecules with similar surroundings and the same reactive sites will share the same chemical transformations (28). It allows users to search for and annotate unannotated reactions. To enrich this annotation function for both sequence and reaction, a one-stop annotation integration is offered in the unfolded annotation tools on the EnzyMine homepage. EnzyMine utilizes the HMMER3 algorithm in FunFHMMer to complete the Gene Ontology annotation for protein sequences queries via a web server (29, 30). Bio2Rxn, a user-friendly platform for automatic reaction annotation based on protein sequences, has been integrated into EnzyMine ue to its consensus strategy and high precision in EC number prediction (31). It adopts the voting strategy to achieve high coverage and accuracy, which guarantees the *ab initio* understanding of enzymes and will expand the sequence annotation.

Discussion

The collection of scattered feature data and supplementation of missing information are the direction in which database optimization is headed. Enzyme dataset provides the raw materials needed for insight into biological systems, but its potential can only be realized through high-level analysis. This would include feature exploration of unannotated sequences and reactions that lack protein sequence or identity. Textual description of the biochemical reaction in UniProt is now being replaced by the reaction in Rhea; high-quality biochemical information addition in enzyme annotation gradually shows its trend (32).

Conventional databases have, therefore, been expanded beyond sequence and structure data to include reaction centre visualization, reaction rule extraction and core-to-core analysis. Using current reaction data in EnzyMine, the prediction accuracy of returned EC number output can reach 92%, which can be improved further by a combination with enzyme sequence and structure data and prove the validity of chemical information. Reaction feature has the potential to provide more high-quality and low-latitude features for the feature extraction process of machine learning and is expected to improve the prediction accuracy of function annotation. In future studies, on the one hand, EnzyMine will continue to enrich data resources in line with data development. On the other hand, we will apply this high-valued chemical information to make contributions for further expansion of enzyme functional annotation.

Conclusions

EnzyMine includes compiled and updated data integration in terms of enzyme and replenishes reaction chemical features. Reaction feature data in this study provide the visualization of reaction centre, a summary of reaction rules and analysis of core-to-core, which contributes to a deeper understanding of enzymatic reactions and will offer researchers a clearer view of biochemical processes and enzymative reactions that mediate them. In summary, the overall enzyme feature mining and holistic integration provide advanced and integrated resources to address enzyme function annotation.

Funding

National Key Research and Development Programme of China (grant number 2019YFA0904300, 2018YFA0900700, 2017YFC1601702); National Natural Science Foundation of China (grant number 31700081, 31570092); Scientific Research Conditions and Technical Support System Programme (grant number KFJ-BRP-009); the CAS STS programme (grant number QYZDB-SSW-SMC012); International Partnership Programme of Chinese Academy of Sciences of China (grant number 153D31KYSB20170121) and the Natural Science Foundation of Tianjin (15JCYBJC54300).

Conflicts of interest. Authors declare that there were no conflicts of interest.

References

1. Machado, D., Andrejev, S., Tramontano, M. *et al.* (2018) Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic Acids Res.*, **46**, 7542–7553.
2. Zhang, C., Freddolino, P.L. and Zhang, Y. (2017) COFACTOR: improved protein function prediction by combining structure, sequence and protein-protein interaction information. *Nucleic Acids Res.*, **45**, W291–W299.
3. Kumar, N. and Skolnick, J. (2012) EFICAz2.5: application of a high-precision enzyme function predictor to 396 proteomes. *Bioinformatics*, **28**, 2687–2688.
4. Li, Y., Wang, S., Umarov, R. *et al.* (2018) DEEPre: sequence-based enzyme EC number prediction by deep learning. *Bioinformatics*, **34**, 760–769.
5. Dawson, N.L., Lewis, T.E., Das, S. *et al.* (2017) CATH: an expanded resource to predict protein function through structure and sequence. *Nucleic Acids Res.*, **45**, D289–D295.
6. Sigrist, C.J., de Castro, E., Cerutti, L. *et al.* (2013) New and continuing developments at PROSITE. *Nucleic Acids Res.*, **41**, D344–D347.
7. Ribeiro, A.J.M., Holliday, G.L., Furnham, N. *et al.* (2018) Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. *Nucleic Acids Res.*, **46**, D618–D623.
8. Andreeva, A., Howorth, D., Chothia, C. *et al.* (2014) SCOP2 prototype: a new approach to protein structure mining. *Nucleic Acids Res.*, **42**, D310–D314.
9. Bateman, A., Martin, M.J., O'Donovan, C. *et al.* (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.
10. Berman, H.M., Westbrook, J., Feng, Z. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
11. Watanabe, N., Murata, M., Ogawa, T. *et al.* (2020) Exploration and evaluation of machine learning-based models for predicting enzymatic reactions. *J. Chem. Inf. Model*, **60**, 1833–1843.
12. Dalby, A.R., Hu, Q.-N., Zhu, H. *et al.* (2012) Assignment of EC numbers to enzymatic reactions with reaction difference fingerprints. *PLoS One*, **7**, e52901.
13. Yamanishi, Y., Hattori, M., Kotera, M. *et al.* (2009) E-zyme: predicting potential EC numbers from the chemical transformation pattern of substrate-product pairs. *Bioinformatics*, **25**, i179–i186.
14. Moriya, Y., Shigemizu, D., Hattori, M. *et al.* (2010) PathPred: an enzyme-catalyzed metabolic pathway prediction server. *Nucleic Acids Res.*, **38**, W138–W143.
15. McDonald, A.G., Boyce, S. and Tipton, K.F. (2009) ExplorEnz: the primary source of the IUBMB enzyme list. *Nucleic Acids Res.*, **37**, D593–D597.

16. Nagano, N., Nakayama, N., Ikeda, K. *et al.* (2014) EzCatDB: the enzyme reaction database, 2015 update. *Nucleic Acids Res.*, **43**, D453–D458.
17. Schomburg, I., Jeske, L., Ulbrich, M. *et al.* (2017) The BRENDA enzyme information system–from a database to an expert system. *J. Biotechnol.*, **261**, 194–206.
18. Hatzimanikatis, V., Li, C., Ionita, J.A. *et al.* (2005) Exploring the diversity of complex metabolic networks. *Bioinformatics*, **21**, 1603–1609.
19. Jeffryes, J.G., Colastani, R.L., Elbadawi-Sidhu, M. *et al.* (2015) MINEs: open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics. *J. Cheminform.*, **7**, 44.
20. Hadadi, N., Hafner, J., Shajkofci, A. *et al.* (2016) ATLAS of Biochemistry: a repository of all possible biochemical reactions for synthetic biology and metabolic engineering studies. *ACS Synth. Biol.*, **5**, 1155–1166.
21. Delepine, B., Duigou, T., Carbonell, P. *et al.* (2018) RetroPath2.0: a retrosynthesis workflow for metabolic engineers. *Metab. Eng.*, **45**, 158–170.
22. Duigou, T., Du Lac, M., Carbonell, P. *et al.* (2019) RetroRules: a database of reaction rules for engineering biology. *Nucleic Acids Res.*, **47**, D1229–D1235.
23. Cheng, X., Sun, D., Zhang, D. *et al.* (2020) RxnBLAST: molecular scaffold and reactive chemical environment feature extractor for biochemical reactions. *Bioinformatics*, **36**, 2946–2947.
24. Bajusz, D., Racz, A. and Heberger, K. (2015) Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminform.*, **7**, 20.
25. Cao, Y., Jiang, T. and Girke, T. (2008) A maximum common substructure-based algorithm for searching and predicting drug-like compounds. *Bioinformatics*, **24**, i366–i374.
26. Tian, Y., Wu, L., Yuan, L. *et al.* (2020) BCSExplorer: a customized biosynthetic chemical space explorer with multifunctional objective function analysis. *Bioinformatics*, **36**, 1642–1643.
27. Rose, A.S. and Hildebrand, P.W. (2015) NGL Viewer: a web application for molecular visualization. *Nucleic Acids Res.*, **43**, W576–W579.
28. Rahman, S.A., Cuesta, S.M., Furnham, N. *et al.* (2014) EC-BLAST: a tool to automatically search and compare enzyme reactions. *Nat. Methods.*, **11**, 171–174.
29. Eddy, S.R. (2011) Accelerated profile HMM searches. *PLoS Comput. Biol.*, **7**, e1002195.
30. Das, S., Sillitoe, I., Lee, D. *et al.* (2015) CATH FunFHMMer web server: protein functional annotations using functional family assignments. *Nucleic Acids Res.*, **43**, W148–W153.
31. Zhang, T., Tian, Y., Yuan, L. *et al.* (2020) Bio2Rxn: sequence-based enzymatic reaction predictions by a consensus strategy. *Bioinformatics*, **36**, 3600–3601.
32. Morgat, A., Lombardot, T., Coudert, E. *et al.* (2019) Enzyme annotation in UniProtKB using Rhea. *Bioinformatics*, **36**, 1896–1901.