Bioinformatics, 33(8), 2017, 1258–1260
doi: 10.1093/bioinformatics/btw801
Advance Access Publication Date: 30 December 2016
Applications Note

Databases and ontologies

cBinderDB: a covalent binding agent database

Jiewen Du†, Xin Yan†, Zhihong Liu, Lu Cui, Peng Ding, Xiaoqing Tan, Xiuming Li, Huihao Zhou, Qiong Gu* and Jun Xu*

Research Center for Drug Discovery, School of Pharmaceutical Sciences, Sun Yat-sen University, Guangzhou 510006, China

*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

Associate Editor: Jonathan Wren

Received on September 22, 2016; revised on November 16, 2016; editorial decision on December 11, 2016; accepted on December 13, 2016

Abstract

Motivation: Small molecule drug candidates with attractive toxicity profiles that modulate target proteins through non-covalent interactions are usually favored by scientists and pharmaceutical industry. In the past decades, many non-covalent binding agents have been developed for different diseases. However, an increasing attention has been paid to covalent binding agents in pharmaceutical fields during recent years. Many covalent binding agents entered clinical trials and exerted significant advantages for diseases such as infection, cancers, gastrointestinal disorders, central nervous system or cardiovascular diseases. It has been recognized that covalent binding ligands can be attractive drug candidates. But, there is lack of resource to support covalent ligand discovery.

Results: Hence, we initiated a covalent binder database (cBinderDB). To our best knowledge, it is the first online database that provides information on covalent binding compound structures, chemotypes, targets, covalent binding types and other biological properties. The covalent binding targets are annotated with biological functions, protein family and domains, gene information, modulators and receptor–ligand complex structure. The data in the database were collected from scientific publications by combining a text mining method and manual inspection processes. cBinderDB covers covalent binder's data up to September 2016.

Availability and Implementation: cBinderDB is freely available at www.rcdd.org.cn/cbinderdb/

Contact: guqiong@mail.sysu.edu.cn orjunxu@biochemomes.com.

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Drugs with covalent binding to their targets have been recognized for a long time in pharmaceutical industry (Bauer, 2015). The drug usually has an appropriate reactive group such as electrophilic warhead, which attacks nucleophilic group at a side chain of a biopolymer (Singh *et al.*, 2011). Aspirin, penicillin and omeprazole are famous covalent binding drugs (Adeniyi *et al.*, 2016), which have strong efficiencies and prolonged durations of the actions (Bradshaw *et al.*, 2015). Nevertheless, in case of off-target, the covalent bonders could significantly increase the risk of the toxicity, and cause serious side effects (Barf and Kaptein, 2012). Therefore, scientists have been encouraged to explore non-covalent binding leads in past decades (Potashman and Duggan, 2009). For example, BindingDB collects 551 097 binding ligands, the 99% of them are non-covalent ligands (Gilson *et al.*, 2016). With the great concern of the difficulty in finding high potent non-covalent binders, covalent binders have been reiterated in recent years (Singh *et al.*, 2011), and the number of reports on covalent binders is increasing (Supplementary Fig. S1). Many covalent binding agents entered clinical trials and exerted significant advantages for diseases such as infection, cancer, gastrointestinal disorders, central nervous system and cardiovascular indications (Gonzalez-Bello, 2016). Covalent binding agents can be attractive drug candidates however, there is a need for a resource to support the therapeutic innovation based upon covalent binding interactions. Hence, we initiated this covalent binder database (cBinderDB). A text mining toolkit was developed to systematically collect, validate and annotate covalent binders from publicly available resources.

2 Methods

2.1 Data collection and processing

The scientific publications were selected by a text mining method, which ranked a literature by calculating the frequencies of the predefined keywords related to co-valent binding terms in the abstracts. The selected publications were further refined by manual inspections. More covalent binder's data were derived from peer-reviewed scientific literature, United States Patent database and other public accessible databases, such as DrugBank 5.0 (Law *et al.*, 2014), ChEMBL_21 (Gaulton *et al.*, 2012), RCSB PDB (Rose *et al.*, 2015), BindingDB (Gilson *et al.*, 2016). The data from different resources were normalized with the same standard (a covalent binder is described in a chemical structure connection table, reactive groups and types, binding residues and locations, related targets, references, etc.). Experimental data consist of binding affinity measurements, bioactivities from bioassays, mutation validations, mass spectra and/or structure biology data. These data are also associated with genomics/proteomics data resources, such as UniproKB (Bateman *et al.*, 2015), HUGO Gene Nomenclature Committee (HGNC) (Gray *et al.*, 2015), NCBI Protein Database (Geer *et al.*, 2010), RCSB PDB (Rose *et al.*, 2015). All public databases we accessed were up to September,2016.

2.2 Online database implementation

cBinderDB was implemented as a relational database in MySQL server and consisted of 7 tables. The ER-model is depicted in Supplementary Figure S2. Tables *Ligand* and *Target* are main entities of the database. Table *Ligand* possesses calculated properties and knowledge associated with medicine. Table *Target* possesses information on protein sequences and structures. The database allows on-line submitting substructure search query which is implemented in the molecular structure sketcher of MarvinJS web component. Protein 3D structure can be rendered with ChemDoodle web component. Chemical substructure search engine is the in-house algorithm GMA (Xu, 1996), and similarity search engine of chemical structure is also the in-house algorithm GSA (Yan *et al.*, 2012). The substructure search and similarity search algorithms were accelerated by GPU technology. The query builder allows to combine chemical substructure search with text search and digital search.

3 Results and discussion

3.1 cBinderDB data collection

6574 scientific papers regarding covalent binding studies were derived from PubMed database. The literature covered the covalent binding studies from 1960 up to 2016. After validating these papers, 527 covalent binders and 190 related protein targets were collected in cBinderDB. More than 95% of covalent binders are inhibitors against their targets, others are activators or modulators. The molecule weights of the covalent binders range from 95 to 1089 Dalton. Supplementary Figure S3 also provides the statistics of the covalent binder Lipinski parameters. Two hundred forty-two covalent binders in cBinderDB are drugs. Functionally, the covalent binders are categorized into four types (Singh *et al.*, 2011): (i) warhead attacker, (ii) Suicide, (iii) Prodrug and (iv) undetermined. Eighty-nine percent binders are warhead attackers, such as tyrosine kinase inhibitors (TKIs) and beta-lactams. These attackers usually are anti-tumor and anti-infection agents, and interact with kinases, enzymes or Penicillin-binding proteins. Eight percent binders are suicide agents, such as acetylcholinesterase inhibitors (AChEI) and alanine racemase inhibitors. These ligands are widely used in the treatment of Parkinson disease or infectious diseases. About 3% binders are prodrugs, such as proton pump inhibitors and P2Y purino-receptor inhibitors, the former can used to treat gastrointestinal disorders, the latter can be used to inhibit blood clots in coronary artery disease, peripheral vascular disease, cerebrovascular disease, or to prevent heart attack and stroke. The majority of electrophilic warheads are alpha-beta unsaturated ketones moiety and beta-lactam. Other functional moieties include epoxides, boron, benzoquinones, vinyl sulfones and so on. On target sides, a covalent binding partner of a ligand is cysteine or serine that possesses a nucleophilic group. Tyrosine, lysine, or glutamine can be a covalent binder's partner as well. It is interesting that several ligands covalently bind at the allosteric sites of their targets (Nussinov and Tsai, 2015). The ligands are mainly peptides, which contain aldehyde moiety or special carbonyl groups. The details can be found in Supplementary Fig. S4.

3.2 Web interface

The web service of cBinderDB was implemented in Golang (golang.org/) language. cBinderDB can be accessed through web browsers. All covalent binders and their targets can be online browsed. The Marvin JS web component is used as a chemical structure query builder, which allows substructure search, simple Markush search and similarity search. Three-dimensional biopolymer structures are rendered with ChemDoodle module. Ligand-receptor binding modes can be examined. Target names, UniProt IDs, gene names, HGNC IDs, ChEMBL IDs, CAS numbers, InChIKeys, PubChem CIDs and SMILES notations are searchable in the database.

3.3 Discussion

A cBinderDB user can be inspired by exploring known binders against a specific target to design a covalent binder. For example, you may find how many covalent inhibitors against EGFR, examine the reactive groups, binding partners and electrophilic warheads; You can also check if your compound is a covalent binder and which targets the binder would bind to. Case studies are listed in Supplementary Figures S5, S6 and S7. Currently, cBinderDB does not support sequence search. This feature is scheduled to support in the next release.

Funding

This work was supported by Natural Science Foundation of China (81473138), Guangdong Frontier & Key Technology Innovation Program (2015B010109004), Guangdong Natural Science Foundation (2016A030310228). Guangdong Province Key Lab. Construction Foundation (2011A060901014), Collaborative Innovation Center of HPC, NUDT, Changsha.

Conflict of Interest: none declared.

References

Adeniyi,A.A. *et al.* (2016) New drug design with covalent modifiers. *Exp. Opin. Drug Discov.*, **11**, 79–90.

Barf,T. and Kaptein,A. (2012) Irreversible protein kinase inhibitors: balancing the benefits and risks. *J. Med. Chem.*, **55**, 6243–6262.

Bateman,A. *et al*. (2015) UniProt: a hub for protein information. *Nucleic Acids Res*., **43**, D204–D212.

Bauer,R.A. (2015) Covalent inhibitors in drug discovery: from accidental discoveries to avoided liabilities and designed therapies. *Drug Discov. Today*, **20**, 1061–1073.

Bradshaw,J.M. *et al*. (2015) Prolonged and tunable residence time using reversible covalent kinase inhibitors. *Nat. Chem. Biol*., **11**, 525–531.

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

Geer,L.Y. *et al*. (2010) The NCBI BioSystems database. *Nucleic Acids Res*., **38**, D492–D496.

Gilson,M.K. *et al*. (2016) BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res*., **44**, D1045–D1053.

Gonzalez-Bello,C. (2016) Designing irreversible inhibitors–worth the effort? *ChemMedChem*, **11**, 22–30.

Gray,K.A. *et al*. (2015) Genenames.org: the HGNC resources in 2015. *Nucleic Acids Res*., **43**, D1079–D1085.

Law,V. *et al*. (2014) DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res*., **42**, D1091–D1097.

Nussinov,R. and Tsai,C.J. (2015) The design of covalent allosteric drugs. *Annu. Rev. Pharmacol*., **55**, 249–267.

Potashman,M.H. and Duggan,M.E. (2009) Covalent modifiers: an orthogonal approach to drug design. *J. Med. Chem*., **52**, 1231–1246.

Rose,P.W. *et al*. (2015) The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. *Nucleic Acids Res*., **43**, D345–D356.

Singh,J. *et al*. (2011) The resurgence of covalent drugs. *Nat. Rev. Drug Discov*., **10**, 307–317.

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

Yan,X. *et al*. (2012) GSA: a GPU-accelerated structure similarity algorithm and its application in progressive virtual screening. *Mol. Divers*, **16**, 759–769.