IDPsBind: a repository of binding sites for intrinsically disordered proteins complexes with known 3D structures

CanZhuang Sun1, YongE Feng1* and GuoLiang Fan2*

Abstract
Background: Intrinsically disordered proteins (IDPs) lack a stable three-dimensional structure under physiological conditions but play crucial roles in many biological processes. Intrinsically disordered proteins perform various biological functions by interacting with other ligands.

Results: Here, we present a database, IDPsBind, which displays interacting sites between IDPs and interacting ligands by using the distance threshold method in known 3D structure IDPs complexes from the PDB database. IDPsBind contains 9626 IDPs complexes and 880 intrinsically disordered proteins verified by experiments. The current release of the IDPsBind database is defined as version 1.0. IDPsBind is freely accessible at http://www.s-bioinformatics.cn/idpsbind/home/.

Conclusions: IDPsBind provides more comprehensive interaction sites for IDPs complexes of known 3D structures. It can not only help the subsequent studies of the interaction mechanism of intrinsically disordered proteins but also provides a suitable background for developing the algorithms for predicting the interaction sites of intrinsically disordered proteins.

Keywords: Intrinsically Disordered Proteins, Intrinsically Disordered Proteins Complexes, Binding sites, PDB

Background
Intrinsically disordered proteins (IDPs) lack a stable secondary or tertiary structure under physiological conditions. Still, they participate in many important biological processes such as cell signal transduction, DNA metabolism, mRNA alternative splicing, protein–protein interaction, and so on [1–3]. Recent studies have shown that IDPs are associated with some diseases when modifications, translations, or expressions of IDPs are abnormal [4–6]. Due to the importance of IDPs in organisms, IDPs have become a hot spot in the current research on protein function. Over the past 20 years, many IDPs have been validated experimentally or computationally. Disprot is the first curated database containing a collection of experimentally validated IDPs and IDP disordered regions [7]. The Disprot database includes a total of 1590 IDPs sequences, excluding the ambiguous and obsolete regions in release 2020_12. These IDPs are from 10 different species. The D2P2 database consists of computationally predicted IDPs from distinct proteomes [8], in which annotations for IDPs are derived from MobiDB. MobiDB3.0 [9] provides information about intrinsically disordered regions (IDRs), related features from various sources, and prediction tools. Different levels of reliability and different features are reported as different and independent annotations. The IDEAL [10] is a database incorporating functional with structural/
disorder annotations for IDPs by manually integrating protein databank (PDB) [11].
These IDPs perform critical biological functions by interacting with other proteins or ligands. Currently, there are some databases about binding sites [12–15], such as Disbind, in which contains 226 IDPs with functional site annotations and binding ligands, including proteins, RNA, DNA, metal ions and others, respectively. However, studying IDPs-ligand interactions is still challenging due to the flexible binding affinity. Lack of enough information on how IDPs interact with other molecules, the biological functions for mostly IDPs are unknown. Although existing databases contain some helpful information, the number of IDPs and IDPs complexes is too tiny to support further study of the IDPs-ligand interaction. With the progress of structural biology, the number of protein structure files in PDB is growing rapidly, among which the structure files of IDPS are also increasing. This motivates us to develop a comprehensive IDPs-ligand interaction database and provide more interacting sites. In this paper, we introduced the IDPsBind database, which contains 9626 IDPs complex interactions. IDPsBind displays binding sites between IDPs and interacting ligands by using the distance threshold method in known 3D structure IDPs complexes from the PDB database. IDPs are selected from the Disprot database (release 2020_12). Each entry in IDPsBind contains a comprehensive list of annotations: primary information of an IDPs, sequence information, and binding sites information in PDB sequences.

**IDPsBind database construction**
The following is the procedure for IDPsBind construction. (a) The intrinsically disordered proteins (IDPs) are derived from the Disprot (http://www.Disprot.org/, release: 2020_12) [7]; (b) Eliminate those IDPs with ambiguous and obsolete regions; (c) IDPs-ligand complex structures with x-ray crystallography resolution of better than 3.5 angstroms in the PDB Database are selected for study; (d) Elimination those IDPs with mutant residues in the IDPs complexes; (e) Binding ligands of IDPs are selected from HETATM in the PDB file. Finally, IDPsBind contains 880 IDPs and 9626 IDPs complexes (from PDB).

As in previous studies, an amino acid residue within a protein sequence is designated as a binding site if it contains at least one atom that falls within a cutoff distance from any atoms of the ligand molecule in the complex [16–19]. Binding residues in IDPsBind are determined by a distance cutoff of 3.5 angstroms between any atoms of a protein. All corresponding PDB chains (resolution better than 3.5 angstroms) for an IDP are used for analysis. All binding ligands and binding sites information are derived from ATOM and HETATM in the PDB file. The construction process of the IDPsBind database, the distribution of binding ligands and binding sites in the IDPsBind, are shown in Figs. 1, 2, 3 and 4.

**Web interface**
IDPsBind provides six basic interfaces: Home, Browse, Download, Search, Statistics, and Help. The ‘Home’ page describes the introduction of the IDPsBind database provides links to the three primary associated databases. The ‘Browse’ page displays the summary

![Fig. 1 Distribution of 880 IDPs in the organism](image-url)
Fig. 2 Workflow of the construction of IDPsBind

Fig. 3 Distribution of the top 20 binding ligands
of all entries in the IDPsBind database. This interface lists six components: IDPsBind ID, Disprot ID, UniProt ID, Protein name, Source, and Disordered content. All items collected in IDPsBind numbered from IDP0001 to IDP0880 can be retrieved by clicking the ‘Browse’ option. Clicking any IDPsBind ID will return a display of the detailed information on the target chain. The data stored for each ID has three parts. The first part shows the basic information about the protein. The second part provides the sequence of IDPs in the Disprot database and color-codes the disordered regions. The third part shows the labeled binding sites of the PDB chains (resolution better than 3.5 angstroms) when the protein interacts with ligands. Moreover, the interface displays the abbreviation of the ligand, clicking the ‘?’ label will show the full name of the ligand. Users can also check specific ligand information clicking the abbreviated ligand jump link to a new interface. And clicking the ‘load’ option, the corresponding structure of the complex is visualized on the right side of the page. Users can download them as a whole on the ‘Download’ page. On the Search interface, users enter any keyword or IDPsBind/Disprot/UniProt ID and then click the ‘search’, then the page shows the results similar to the “Browse the entry” page. The ‘Statistics’ interface shows some basic data and information about the IDPsBind database. The ‘Help’ interface answers some questions on the IDPsBind database. The current release of IDPsBind is 1.0, which will be updated in the future during the PDB release update.

Conclusions
We have developed a comprehensive IDPs-ligand interaction database, IDPsBind, in which IDPs are taken from the DisProt database (2020_12), and corresponding IDPs complexes are from the PDB database. Although there are already a handful of ligand-binding databases in the literature, IDPsBind is distinguished from other databases in the following aspects. (a) IDPsBind contains many interactions of IDPs, 3203 binding ligands including proteins, DNA, RNA, et al. (b) The interaction includes not only the disordered regions with the ligand, but also that of the ordered regions in IDPsBind. (c) The IDPs-ligand binding information is based on the PDB file, and all the PDB chains (resolution better than 3.5 angstroms) for IDPs were analyzed. In this way, ligand-binding sites of the target chain cannot be missing in IDPsBind. (d) All data in IDPsBind database are freely available for download. We hope that the IDPsBind can provide helpful information required for specific IDPs-relevant studies.

Acknowledgements
The authors are grateful to the anonymous reviewers for their valuable suggestions and comments, which have led to the improvement of this paper. The authors wish to thank all the members that do this work.
Authors' contributions
F.YE. designed the project and performed the analysis, and drafted the manu-
script. S.CZ. collected the data and carried out the computation of binding
sites. S.CZ. & F.GL. set up IDPsBind web server. The corresponding author:
F.YE. & F.G.L. correspond to yefeng@imau.edu.cn. All authors have read and
approved the final manuscript.

Authors' information
Canzhuang Sun is a Master’s student at the College of Science, Inner Mongolia
Agriculture University. His research interests lie in the field of the interac-
tion between intrinsically disordered proteins and binding ligands.
Yonge Feng is a Full Professor at the College of Science, Inner Mongolia Agri-
culture University. Her research focuses on machine learning and structure of
protein, function, and bioinformatics.
Guoliang Fan is a Full Professor in the Department of Physics, School of Physi-
ical Science and Technology, Inner Mongolia University. His research focuses
on machine learning and epigenetics.

Funding
Funding for the work was provided by the Special Project of the National Nat-
ural Science Foundation of China (62141204) and the National Natural Science
Foundation of China (62063024, 61461038), the Scientific Research Program at
Universities of Inner Mongolia Autonomous Region of China (NJZY20005).

Availability of data and materials
The author can provide compiled executable file on data in this article. Please
send an email to the author (yefeng@imau.edu.cn) to query the relevant data
of this paper. And all data can be downloaded freely in IDPsBind.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors confirm that they have no competing interests.

Received: 28 July 2021 Accepted: 14 July 2022
Published online: 26 July 2022

References

1. Csizmok V, Foldi AV, Kriwacki RW, Forman-Kay JD. Dynamic protein inter-
action networks and new structural paradigms in signalling. Chem Rev.
2016;116(11):6424–62.
2. Wright PE, Dyson HJ. Intrinsically disordered proteins in cellular signalling
and regulation. Nat Rev Mol Cell Biol. 2015;16(1):18–29.
3. Binolfi A, Limatola A, Verzini S, Kosten J, Theillet F-X, Rose HM, Bekei B,
Stuiver M, Van Rossum M, Selenko P. Intracellular repair of oxidation-
damaged α-synuclein fails to target C-terminal modification sites. Nat
Commun. 2016;7(1):1–10.
4. Fung HYJ, Birol M, Rhoades E. IDPs in macromolecular complexes: the
roles of multivalent interactions in diverse assemblies. Curr Opin Struct
Biol. 2018;49:36–43.
5. Babu MM. The contribution of intrinsically disordered regions to protein
function, cellular complexity, and human disease. Biochem Soc Trans.
2016;44(5):S185–200.
6. Babu MM, van der Lee R, de Groot NS, Gsponer J. Intrinsically disordered
proteins: regulation and disease. Curr Opin Struct Biol. 2011;21(3):432–40.
7. Hatos A, Hajdu-Soltész B, Monzon AM, Palopoli N, Alvarez L, Aikac-Fas B,
Bassot C, Benitez GI, Bevilacqua M, Chasapí A. DisProt: intrinsic protein
 disorder annotation in 2020. Nucleic Acids Res. 2020;48(D1):D269–76.
8. Oates ME, Romero P, Ishida T, Ghalwash M, Mizianty MJ, Xue B, Dosztányi
Z, Uversky VN, Obradovic Z, Kurian L. D2P2: database of disordered
protein predictions. Nucleic Acids Res. 2012;41(D1):D508–16.
9. Piovesan D, Tabaro F, Paladini L, Necci M, Mičetić I, Camilloni C, Davey N,
Dosztányi Z, Mézírszás B, Monzon AM. MobiDB 3.0: more annotations for
intrinsic disorder, conformational diversity and interactions in proteins.
Nucleic Acids Res. 2018;46(D1):D471–6.
10. Fukushima S, Sakamoto S, Nobe Y, Murakami SD, Amemuya T, Hosoda
K, Koke R, Hiroaki H, Ota M. IDEAL: intrinsically disordered pro-
teins with extensive annotations and literature. Nucleic Acids Res.
2012;40(D1):D507–11.
11. Burley SK, Berman HM, Kleywegt GJ, Markley JL, Nakamura H, Velankar
S. Protein Data Bank (PDB): The Single Global Macromolecular Structure
Archive. Methods: Mol Biol. 2017;1607:627–41.
12. Yu J-F, Dou X-H, Sha Y-J, Wang C-L, Wang H-R, Chen Y-T, Zhang F, Zhou
Y, Wang J-H. DisBind: A database of classified functional binding sites in
disordered and structured regions of intrinsically disordered proteins.
BMC Bioinformatics. 2017;18(1):1–5.
13. Fichó E, Reményi I, Simon I, Mézírszás B. MFIB: a repository of protein
complexes with mutual folding induced by binding. Bioinformatics.
2017;33(22):3682–4.
14. Yang J, Roy A, Zhang Y. BioLP: a semi-manually curated database for
biologically relevant ligand–protein interactions. Nucleic Acids Res.
2012;41(D1):D1096–103.
15. Schad E, Fichó E, Pancsa R, Simon I, Dosztányi Z, Mézírszás B. DiBS: a
repository of disordered binding sites mediating interactions with
ordered proteins. Bioinformatics. 2018;34(3):535–7.
16. Gao M, Skolnick J. A threading-based method for the prediction of DNA-
binding proteins with application to the human genome. PLoS Comput
Biol. 2009;5(11):e1000567.
17. Capra JA, Laskowski RA, Thornton JM, Singh M, Funkhouser TA. Predict-
ing protein ligand binding sites by combining evolutionary sequence
conservation and 3D structure. PLoS Comput Biol. 2009;5(12):e1000585.
18. Zhao H, Yang Y, Zhou Y. Structure-based prediction of RNA-binding
domains and RNA-binding sites and application to structural genomics
targets. Nucleic Acids Res. 2011;39(8):3017–25.
19. Kumar M, Gromiha MM, Raghava GPS. Prediction of RNA binding sites in a
protein using SVM and PSSM profile. Proteins. 2008;71(1):189–94.