DATABASE
The Journal of Biological Databases and Curation

Database, 2019, 1–8
doi: 10.1093/database/bay142
Database update

Database update

ccPDB 2.0: an updated version of datasets created and compiled from Protein Data Bank

Piyush Agrawal[1,2], Sumeet Patiyal[2], Rajesh Kumar[1,2], Vinod Kumar[1,2], Harinder Singh[3], Pawan Kumar Raghav[2] and Gajendra P.S. Raghava[2,*]

[1]Bioinformatics Center, CSIR-Institute of Microbial Technology, Sector-39A, Chandigarh-160036, India, [2]Department of Computational Biology, Indraprastha Institute of Information Technology, Okhla Industrial Estate, Phase III, New Delhi, 110020, India and [3]J. Craig Venter Institute 9605 Medical Center Drive, Suite 150 Rockville, MD 20850, USA

*Corresponding author: Tel: +91-11-26907444; Email: raghava@iiitd.ac.in

Citation details: Agrawal,P., Patiyal,S., Kumar,R. *et al.* ccPDB 2.0: an updated version of datasets created and compiled from Protein Data Bank. *Database* (2019) Vol. 2019: article ID bay142; doi:10.1093/database/bay142

Received 31 October 2018; Revised ; Accepted 9 December 2018

Abstract

ccPDB 2.0 (http://webs.iiitd.edu.in/raghava/ccpdb) is an updated version of the manually curated database ccPDB that maintains datasets required for developing methods to predict the structure and function of proteins. The number of datasets compiled from literature increased from 45 to 141 in ccPDB 2.0. Similarly, the number of protein structures used for creating datasets also increased from ∼74 000 to ∼137 000 (PDB March 2018 release). ccPDB 2.0 provides the same web services and flexible tools which were present in the previous version of the database. In the updated version, links of the number of methods developed in the past few years have also been incorporated. This updated resource is built on responsive templates which is compatible with smartphones (mobile, iPhone, iPad, tablets etc.) and large screen gadgets. In summary, ccPDB 2.0 is a user-friendly web-based platform that provides comprehensive as well as updated information about datasets.

Database URL: http://webs.iiitd.edu.in/raghava/ccpdb

Introduction

Advancement in the sequencing technology has enormously increased the protein sequence information in the databases. This increase in protein sequence data has widened the gap between the sequences and annotations (1). Therefore, to fill this gap *in silico* tools are required for annotating the function of these proteins since it is very cumbersome to obtain a crystal structure of all the protein sequences. Therefore, to combat this situation, the number of sequence-based tools has been developed in the past few decades (1–16). These methods are developed using the experimentally proven structural information present in the Protein Data Bank (PDB) (17). Hence, cleaned and refined datasets are required for training, testing and validation of the new method and for benchmarking previous ones.

In order to facilitate scientific community working in the field of structural biology, we created a database ccPDB (18) in 2011, where we have collected and compiled the experimentally validated datasets from the literature. In addition, we provide a web-based platform that allows users to create customized datasets as per their requirement from the PDB July 2011. In the past 7 years, the number of structures in PDB has been nearly doubled, as PDB is continually growing over the years. In addition, many methods which use datasets derived from protein structures also increased drastically. There is a significant increase in the number of datasets and the number of protein structures used to create these datasets. In order to provide updated information to the structural biologist, we have developed ccPDB 2.0 which is an updated version of ccPDB.

Materials and methods

Data collection and organization

Since the first version of ccPDB (18) published in 2011, there has been enormous growth in the development of improved methods in the field of secondary structure prediction (9, 19–24), irregular secondary structure prediction (10, 25–28), protein–ligand interactions (7, 15, 16, 29), DNA/RNA–protein interactions (13, 30, 31), protein crystallization and propensity prediction (32–35), dihedral angle prediction (6, 36–38), surface accessibility prediction (39), Rotamer libraries (8) and others (40–43). These methods have been found to annotate protein structure and function in comparison to earlier methods. Therefore, we have performed a major update on the data developed for annotating protein structure and function in the past 7 years. We collected the experimentally validated datasets published in the literature for developing different prediction methods. These datasets were extracted from the PubMed articles and their supplementary materials, websites, databases or directly from the authors.

In order to create datasets, we downloaded all the PDB files till March 2018 release from RCSB-PDB (http://www.pdb.org/) (44) and maintained/mirrored these PDB files at our server. We also maintained DSSP (45, 46) and other PDB-related information at our server. Therefore, a user can create its own customized datasets using these files. Also, we used different software for generating useful information from PDB files. Some of the software includes the following: (i) PROMOTIF (47) for identifying different structural motifs, (ii) LPC (48) for generating protein–ligand interaction data, (iii) HBPLUS (49) for generating protein–DNA/RNA interaction data, (iv) BlastClust for generating protein cluster based on sequence similarity and (v) in-house PERL and Python scripts for analyzing PDB files and different calculations.

Broadly, datasets created/compiled in ccPDB can be divided into two categories: (i) datasets for structure/-function annotation at the protein level and (ii) datasets for annotation of protein at the residue level. In the case of structural annotation at the protein level, the overall function of a protein is estimated like the prediction of ATP-, RNA- and DNA-binding proteins. In case of residue level annotation, we predict the function of each residue in a protein, like the prediction of ATP-, RNA- and DNA-interacting residues in a protein.

Database architecture

A ccPDB 2.0 back end is built using Apache HTTP server 2.2 and MySQL server 5.1.47 and front end using HTML, PHP 5.2.9 and JAVA scripts. We used an HTML5 web template for making a website compatible with mobile and tablet. The abovementioned technologies were used as they are platform independent and open source.

ccPDB 2.0 implementation

ccPDB 2.0 is an updated and comprehensive database which maintains existing datasets obtained from the published literature and datasets derived from the PDB files. Besides, ccPDB 2.0 also allows a user to create its own customized datasets using PDB's latest data. Functioning of the database can be broadly classified into three major sections. Details of these sections are mentioned below.

Collection and compilation of datasets

This section maintains the experimentally validated and published datasets collected from the literature after an extensive search. This section was present in the previous version too. In the current database, we have included the datasets published in the past 7 years (Figure 1). We have added some datasets which were not included in the previous version like metal-interacting residue datasets, antigen–antibody interaction datasets.

Creation of datasets

This is an important module of the ccPDB database that allows users to create customized dataset as per their requirement. This module enables the user to create any type of dataset from the latest release of PDB (March 2018 release). This kind of dataset is very useful for benchmarking different methods and developing new method as the performance of a method largely depends on the dataset

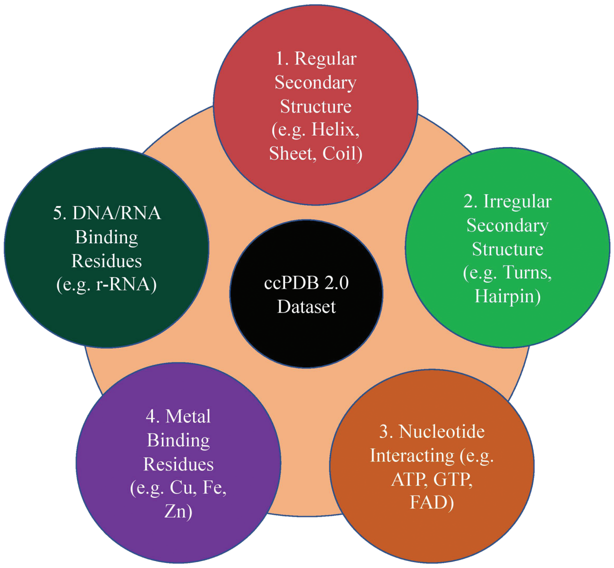
1. Regular Secondary Structure (e.g. Helix, Sheet, Coil)
2. Irregular Secondary Structure (e.g. Turns, Hairpin)
3. Nucleotide Interacting (e.g. ATP, GTP, FAD)
4. Metal Binding Residues (e.g. Cu, Fe, Zn)
5. DNA/RNA Binding Residues (e.g. r-RNA)
ccPDB 2.0 Dataset

Figure 1. Dataset types present in ccPDB 2.0.

size. Different type of datasets compiled from PDB is listed in Figure 1 along with their brief compilation procedure. In order to create a new or customized dataset, the user needs to perform the six steps as explained in Figure 2.

Web services and availability

ccPDB 2.0 is freely available at http://webs.iiitd.edu.in/raghava/ccpdb/. ccPDB 2.0 provides the same web services which were earlier present in the first version of this database. The website is compatible with different platforms like desktop, smartphone and tablet. We have also retained the previous version of the ccPDB which is maintained at http://crdd.osdd.net/raghava/ccpdb/.

Results and discussion

The ccPDB 2.0 database is an updated version of ccPDB which was created to facilitate the user to access the latest information related to protein structural annotation and function. The first version of ccPDB contains very few datasets which were collected from the literature. In the previous version there was a total of 45 datasets compiled from 37 studies. However, the updated database maintains information of 141 experimentally validated datasets compiled from around 68 studies. There were a total 407 200 chains present at the protein level and 1 928 972 chains at residue level in comparison to 340 864 and 6904 chains at protein and residue level, respectively, present in the previous version. We added a new data type 'Metal and Ions Interacting Residue' which was not there in the earlier version.

In the first version of ccPDB, there were only two datasets for regular and nine for irregular secondary structure prediction whereas in the updated version there are 10 and 13 datasets for regular and irregular secondary structure prediction, respectively. The number of datasets for protein–nucleotide interaction in the first version was four; however, in the updated version it has changed to 22. We also included 13 protein–metal ions and protein–acid radical ligand dataset in the updated version which was previously not present. In the previous version, the dataset number of DNA/RNA-interacting proteins at protein level was 5 and 7 at residue level whereas in the updated version it has increased to 14 and 13, respectively. Likewise, in the updated version dihedral angle prediction dataset has risen from 1 to 20, protein crystallization dataset has increased from 3 to 12, bacterial protein interaction dataset has increased from 4 to 6 and surface accessibility prediction dataset has increased from 1 to 3. This tremendous growth of datasets in the literature clearly shows the importance of protein structural annotation and function and how it could lead to better understanding of the role of proteins in various biological and cellular processes.

In order to provide ready-to-use datasets for developing new prediction methods, we compiled information of 70 customized datasets from PDB using standard protocols. Datasets were assembled using the latest PDB March 2018 release which consists of ∼137 000 PDB tertiary structures which is nearly double to the files present in the first version of ccPDB which consists of around 75 000 PDB structures. Due to an increase in the number of PDB structures, there was a considerable growth in the number of chains in the updated version. In the previous version, there was a total 17 731 chains at the protein level and 66 368 chains at residue level whereas in the updated version the number of chains at protein level increases to 33 488 and at residue level it changed to 139 902. The number of protein chains for regular secondary structure has been increased from 5877 to 17 608. Likewise, for irregular secondary structure there was an enormous increase in the number of protein chains for different beta turns, gamma turns, psiloop and hairpin. DNA- and RNA-interacting protein chains were increased to 560 and 415 from 417 and 282, respectively, in the updated version.

We also observed an increase in the number of ligand-interacting proteins and metal-interacting proteins in the updated version of ccPDB. To assist the scientific community, we created the dataset for some of the widely used protein-interacting ligands and metals. A user can download these datasets by clicking on the desired dataset present at our website http://webs.iiitd.edu.in/raghava/ccpdb/collect.php. Comparison of the statistics between ccPDB and ccPDB 2.0 is given in Table 1.

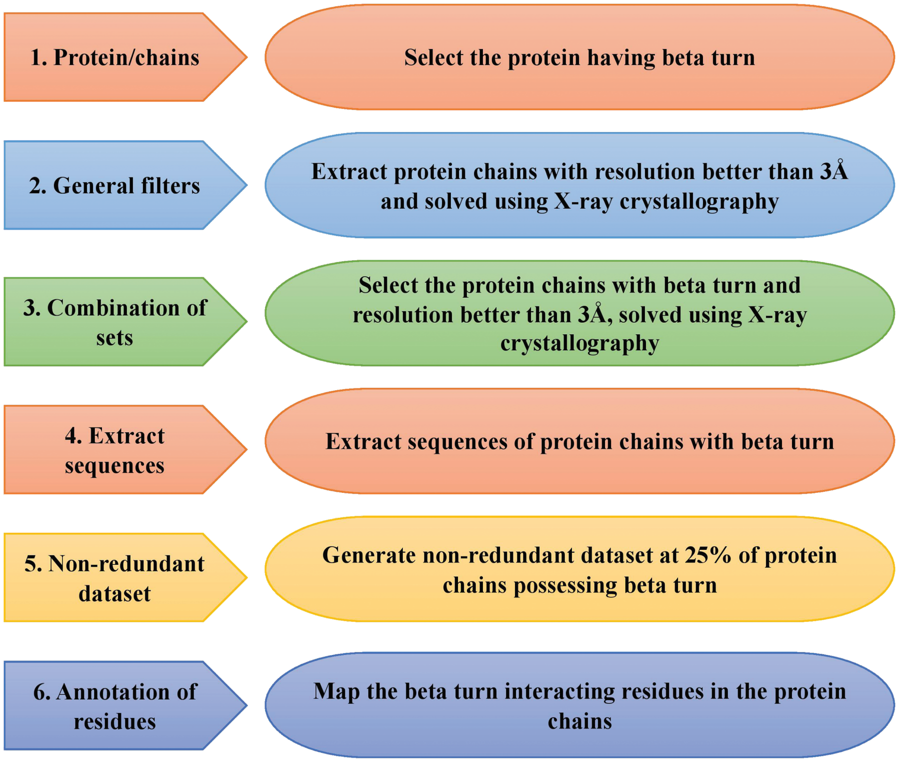
1. Protein/chains
Select the protein having beta turn
2. General filters
Extract protein chains with resolution better than 3Å and solved using X-ray crystallography
3. Combination of sets
Select the protein chains with beta turn and resolution better than 3Å, solved using X-ray crystallography
4. Extract sequences
Extract sequences of protein chains with beta turn
5. Non-redundant dataset
Generate non-redundant dataset at 25% of protein chains possessing beta turn
6. Annotation of residues
Map the beta turn interacting residues in the protein chains

Figure 2. Schematic representation of steps of data set creation module of ccPDB 2.0.

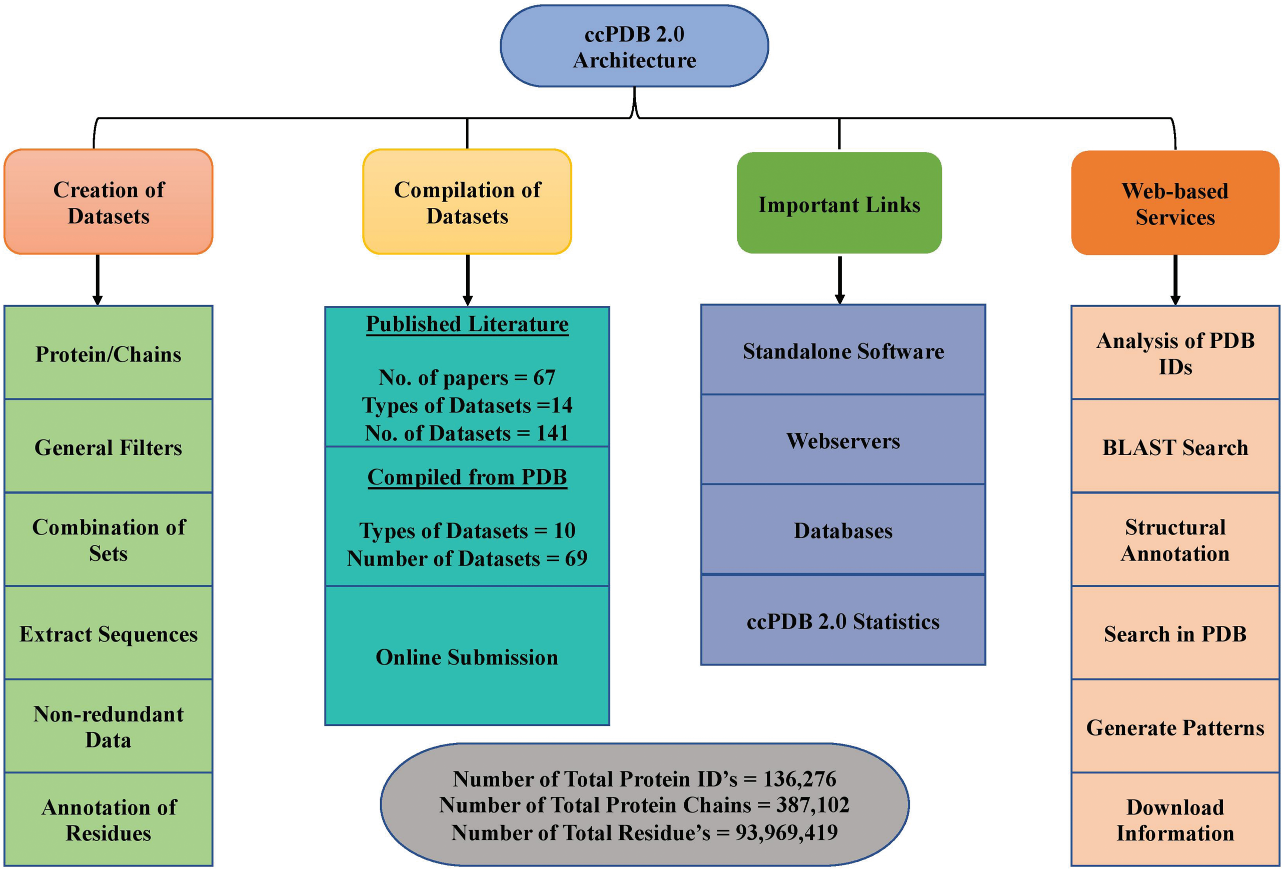

Figure 3. Architecture of ccPDB 2.0.

Table 1. Comparison of datasets compiled from literature and created using PDB at ccPDB and ccPDB 2.0

Sr. No.	Type of dataset	Description	ccPDB (No. of protein chains)	ccPDB 2.0* (No. of protein chains)	Dataset compiled from the literature in ccPDB (no. of protein chains)	Dataset compiled from the literature in ccPDB 2.0 (no. of protein chains)
1	Secondary structure	Eight state	5877	17 608	2	10
		Three state	5877	17 608	(919)	(30 426)
2	Irregular secondary structure	Beta turn I	6691	16 195	9	13
		Beta turn I’	5424	7070	(5045)	(1 913 701)
		Beta turn II	2324	12 429		
		Beta turn II’	6618	5393		
		Beta turn IV	3197	16 183		
		Beta turn VIa1	671	1397		
		Beta turn VIa2	215	406		
		Beta turn VIb	1028	2350		
		Beta turn VIII	4874	11 821		
		Gamma turn C	1059	2889		
		Gamma turn I	5833	12 720		
		Beta buldge-B	271	524		
		Beta buldge-C	4926	9214		
		Beta buldge-G	3694	6931		
		Beta buldge-S	717	1304		
		Beta buldge-W	864	1752		
		Hairpin	4931	12 984		
		Psiloop	1197	2460		
3	DNA/RNA-interacting residues	DNA	417	560	7	13
		RNA	282	415	(1254)	(3958)
4	DNA/RNA-interacting proteins	DNA	417	560	7	13
		RNA	282	415	(1254)	(3958)
5	Nucleotide-interacting residues	ATP	228	313	4	22
		ADP	300	353	(605)	(9213)
		GTP	52	83		
		GDP	88	120		
		NAD	133	140		
		FAD	156	172		
		FMN	103	117		
		UDP	51	68		
6	Nucleotide-interacting proteins	ATP	228	313	4	22
		ADP	300	353	(605)	(9213)
		GTP	52	83		
		GDP	88	120		
		NAD	133	140		
		FAD	156	172		
		FMN	103	117		
		UDP	51	68		
7	Ligand-binding residues	SO4	2604	3312	0	4
		PO4	1002	1299	(0)	(726)
		NAG	488	727		
		HEM	167	176		
		BME	163	191		
		EDO	1095	1507		
		PLP	64	65		

Continued.

Table 1. Continued.

Sr. No.	Type of dataset	Description	ccPDB (No. of protein chains)	ccPDB 2.0* (No. of protein chains)	Dataset compiled from the literature in ccPDB (no. of protein chains)	Dataset compiled from the literature in ccPDB 2.0 (no. of protein chains)
8	Ligand-binding proteins	SO4	2604	3312	0	4
		PO4	1002	1299	(0)	(726)
		NAG	488	727		
		HEM	167	176		
		BME	163	191		
		EDO	1095	1507		
		PLP	64	65		
9	Metal-interacting residues	Fe	163	215	0	9
		Mg	1384	1908	(0)	(1374)
		Ca	1018	1402		
		Mn	386	521		
		Zn	1118	1660		
		Co	149	201		
		Ni	243	355		
10	Metal-interacting proteins	Fe	163	215	0	9
		Mg	1384	1908	(0)	(1374)
		Ca	1018	1402		
		Mn	386	521		
		Zn	1118	1660		
		Co	149	201		
		Ni	243	355		

*Number of non-redundant PDB chains generated using BlastClust at 25% sequence similarity and resolution in between 0–3 Å.

ccPDB 2.0 also allows the user to create its own customized dataset in six simple steps using the 'CREATION OF DATASET' module (See Materials and methods). These customized datasets can be used for developing new method as well as benchmarking other methods. The 'WEB SERVICES' module is another important module which allows the user to analyze its PDB structure as well as annotate it. Analysis of the PDB_ID option of this module comprises only those web services which are functional. We have removed some of the web services in the updated version whose servers were not functional or have been obsolete. We have tried to compile all the possible links of functional standalone software, web services as well as database related to protein structure annotation and function, molecular dynamics and docking in the 'IMPORTANT LINKS' module. We believe that this module will be of great help to all the researchers working in the field of protein structure annotation, function and drug designing. Complete architecture of the ccPDB 2.0 is given in Figure 3.

Declarations

Ethics approval and consent to participate

This study does not require any ethical clearance or any consent to participate.

Availability of data and materials

All data including datasets are freely available to scientific community and can be downloaded from the webserver.

Authors' contribution

P.A., S.P., R.K., V.K. and P.K.R. collected and compiled the datasets. P.A. and H.S. performed the experiments. P.A., R.K., V.K.P. and S.P. developed the web interface. P.A. and G.P.S.R. analyzed the data and prepared the manuscript. G.P.S.R. conceived the idea and coordinated the project. All authors read and approved the final paper.

Acknowledgements

Authors are thankful to the Department of Science and Technology, Department of Biotechnology and Council of Scientific and Industrial Research and Government of India, for fellowships and the financial support.

Funding

J.C. Bose National Fellowship, Department of Science and Technology, Government of India.

Conflict of interest. None declared.

References

1. Bromberg,Y., Yachdav,G., Ofran,Y. *et al.* (2009) New in protein structure and function annotation: hotspots, single nucleotide polymorphisms and the "Deep Web". *Curr. Opin. Drug Discov. Devel.*, **12**, 408–419.
2. Yang,J. and Zhang,Y. (2015) Protein structure and function prediction using I-TASSER. *Curr. Protoc. Bioinformatics*, **52**, 5.8.1–5.8.15.
3. Ansari,H.R. and Raghava,G.P.S. (2010) Identification of NAD interacting residues in proteins. *BMC Bioinformatics*, **11**, 160.
4. Mishra,N.K. and Raghava,G.P.S. (2010) Prediction of FAD interacting residues in a protein from its primary sequence using evolutionary information. *BMC Bioinformatics*, **11**, S48.
5. Mizianty,M.J. and Kurgan,L. (2009) Meta prediction of protein crystallization propensity. *Biochem. Biophys. Res. Commun.*, **390**, 10–15.
6. Gao,Y., Wang,S., Deng,M. *et al.* (2018) RaptorX-Angle: real-value prediction of protein backbone dihedral angles through a hybrid method of clustering and deep learning. *BMC Bioinformatics*, **19**, 100.
7. Hu,X., Dong,Q., Yang,J. *et al.* (2016) Recognizing metal and acid radical ion-binding sites by integrating ab initio modeling with template-based transferals. *Bioinformatics*, **32**, 3260–3269.
8. Kirys,T., Ruvinsky,A.M., Tuzikov,A.V. *et al.* (2012) Rotamer libraries and probabilities of transition between rotamers for the side chains in protein–protein binding. *Proteins*, **80**, 2089–2098.
9. Yan,J., Marcus,M. and Kurgan,L. (2014) Comprehensively designed consensus of standalone secondary structure predictors improves Q3 by over 3%. *J. Biomol. Struct. Dyn.*, **32**, 36–51.
10. Singh,H., Singh,S. and Raghava,G.P.S. (2015) In silico platform for predicting and initiating β-turns in a protein at desired locations. *Proteins*, **83**, 910–921.
11. de la Cruz,X., Hutchinson,E.G., Shepherd,A. *et al.* (2002) Toward predicting protein topology: an approach to identifying beta hairpins. *Proc. Natl. Acad. Sci. U. S. A.*, **99**, 11157–62.
12. Yan,J. and Kurgan,L. (2017) DRNApred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues. *Nucleic Acids Res.*, **45**, e84.
13. Muppirala,U.K., Honavar,V.G. and Dobbs,D. (2011) Predicting RNA–protein interactions using only sequence information. *BMC Bioinformatics*, **12**, 489.
14. Chauhan,J.S., Mishra,N.K. and Raghava,G.P.S. (2009) Identification of ATP-binding residues of a protein from its primary sequence. *BMC Bioinformatics*, **10**, 434.
15. Yu,D.-J., Hu,J., Huang,Y. *et al.* (2013) TargetATPsite: a template-free method for ATP-binding sites prediction with residue evolution image sparse representation and classifier ensemble. *J. Comput. Chem.*, **34**, 974–985.
16. Chen,K., Mizianty,M.J. and Kurgan,L. (2012) Prediction and analysis of nucleotide-binding residues using sequence and sequence-derived structural descriptors. *Bioinformatics*, **28**, 331–341.
17. Berman,H.M., Westbrook,J., Feng,Z. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
18. Singh,H., Chauhan,J.S., Gromiha,M.M. *et al.* (2012) ccPDB: compilation and creation of datasets from Protein Data Bank. *Nucleic Acids Res.*, **40**, D486–D489.
19. Drozdetskiy,A., Cole,C., Procter,J. *et al.* (2015) JPred4: a protein secondary structure prediction server. *Nucleic Acids Res.*, **43**, W389–W394.
20. Manikandan,P. and Ramyachitra,D. (2018) PATSIM: prediction and analysis of protein sequences using hybrid Knuth–Morris–Pratt (KMP) and Boyer–Moore (BM) algorithm. *Gene*, **657**, 50–59.
21. Chebrek,R., Leonard,S., de Brevern,A.G. *et al.* (2014) PolyprOnline: polyproline helix II and secondary structure assignment database. *Database (Oxford)*, **2014**, 1–8.
22. Yang,J. and Shen,H.-B. (2017) MemBrain-contact 2.0: a new two-stage machine learning model for the prediction enhancement of transmembrane protein residue contacts in the full chain. *Bioinformatics*, **34**, 230–238.
23. Louis-Jeune,C., Andrade-Navarro,M.A. and Perez-Iratxeta,C. (2012) Prediction of protein secondary structure from circular dichroism using theoretically derived spectra. *Proteins*, **80**, 374–381.
24. Savojardo,C., Fariselli,P., Martelli,P.L. *et al.* (2013) BCov: a method for predicting β-sheet topology using sparse inverse covariance estimation and integer programming. *Bioinformatics*, **29**, 3151–3157.
25. Kountouris,P. and Hirst,J.D. (2010) Predicting beta-turns and their types using predicted backbone dihedral angles and secondary structures. *BMC Bioinformatics*, **11**, 407.
26. Fuchs,P.F.J. and Alix,A.J.P. (2005) High accuracy prediction of beta-turns and their types using propensities and multiple alignments. *Proteins*, **59**, 828–839.
27. Kaur,H. and Raghava,G.P.S. (2003) Prediction of beta-turns in proteins from multiple alignment using neural network. *Protein Sci.*, **12**, 627–634.
28. Freeman,T.C. and Wimley,W.C. (2012) TMBB-DB: a transmembrane β-barrel proteome database. *Bioinformatics*, **28**, 2425–2430.
29. Hu,J., Li,Y., Zhang,Y. *et al.* (2018) ATPbind: accurate protein–ATP binding site prediction by combining sequence-profiling and structure-based comparisons. *J. Chem. Inf. Model*, **58**, 501–510.
30. Tuvshinjargal,N., Lee,W., Park,B. *et al.* (2016) PRIdictor: protein–RNA interaction predictor. *Biosystems*, **139**, 17–22.
31. Shen,W.-J., Cui,W., Chen,D. *et al.* (2018) RPiRLS: quantitative predictions of RNA interacting with any protein of known sequence. *Molecules*, **23**, 540.
32. Mizianty,M.J. and Kurgan,L. (2011) Sequence-based prediction of protein crystallization, purification and production propensity. *Bioinformatics*, **27**, i24–i33.
33. Wang,H., Wang,M., Tan,H. *et al.* (2014) PredPPCrys: accurate prediction of sequence cloning, protein production, purification and crystallization propensity from protein sequences using multi-step heterogeneous feature fusion and selection. *PLoS One*, **9**, e105902.
34. Wang,H., Feng,L., Zhang,Z. *et al.* (2016) Crysalis: an integrated server for computational analysis and design of protein crystallization. *Sci. Rep.*, **6**, 21383.
35. Charoenkwan,P., Shoombuatong,W., Lee,H.-C. *et al.* (2013) SCMCRYS: predicting protein crystallization using an ensemble scoring card method with estimating propensity scores of P-collocated amino acid pairs. *PLoS One*, **8**, e72368.

36. Berjanskii,M.V., Neal,S. and Wishart,D.S. (2006) PREDITOR: a web server for predicting protein torsion angle restraints. *Nucleic Acids Res.*, **34**, W63–W69.
37. Wu,S. and Zhang,Y. (2008) ANGLOR: a composite machine-learning algorithm for protein backbone torsion angle prediction. *PLoS One*, **3**, e3400.
38. Li,H., Hou,J., Adhikari,B. *et al.* (2017) Deep learning methods for protein torsion angle prediction. *BMC Bioinformatics*, **18**, 417.
39. Ma,J. and Wang,S. (2015) AcconPred: predicting solvent accessibility and contact number simultaneously by a multitask learning framework under the conditional neural fields model. *Biomed Res. Int.*, **2015**, 678764.
40. Chakraborty,A., Ghosh,S., Chowdhary,G. *et al.* (2012) DBETH: a database of bacterial exotoxins for human. *Nucleic Acids Res.*, **40**, D615–D620.
41. Wang,J., Li,Y., Wang,Q. *et al.* (2012) ProClusEnsem: predicting membrane protein types by fusing different modes of pseudo amino acid composition. *Comput. Biol. Med.*, **42**, 564–574.
42. Huang,J., Ru,B., Zhu,P. *et al.* (2012) MimoDB 2.0: a mimotope database and beyond. *Nucleic Acids Res.*, **40**, D271–D277.
43. Ferdous,S. and Martin,A.C.R. (2018) AbDb: antibody structure database—a database of PDB-derived antibody structures. *Database (Oxford)*, **2018**, 1–9.
44. Rose,P.W., Prlić,A., Altunkaya,A. *et al.* (2017) The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Res.*, **45**, D271–D281.
45. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
46. Touw,W.G., Baakman,C., Black,J. *et al.* (2015) A series of PDB-related databanks for everyday needs. *Nucleic Acids Res.*, **43**, D364–D368.
47. Hutchinson,E.G. and Thornton,J.M. (1996) PROMOTIF—a program to identify and analyze structural motifs in proteins. *Protein Sci.*, **5**, 212–220.
48. Sobolev,V., Sorokine,A., Prilusky,J. *et al.* (1999) Automated analysis of interatomic contacts in proteins. *Bioinformatics*, **15**, 327–332.
49. McDonald,I.K. and Thornton,J.M. (1994) Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.*, **238**, 777–793.