Life Science Informatics Publications

Research Journal of Life Sciences, Bioinformatics, Pharmaceutical and Chemical Sciences

Journal Home page http://www.rjlbpcs.com/

RJLBPCS

ISSN 2454-6348

Original Research Article **DOI: 10.26479/2019.0502.10**

DIFFRACTION PRECISION INDEX OF MACROMOLECULAR STRUCTURES: A WEB BASED DATABASE

R. Santhosh[1], E. Velayudham[2], Daliah Michael[2], Namrata Bankoti[2], P. Chandrasekaran[2], D. Karthikeyan[2], S. Pavithra[2], M. Gurusaran[2], K. P. R. Nisha[2], S. N. Satheesh[2], K. Rangachari[2], J. Jeyakanthan[1], K. Sekar[2]*

1. Department of Bioinformatics, Alagappa University, Karaikudi, India.

2. Department of Computational and Data Sciences, Indian Institute of Science, Bangalore, India.

ABSTRACT: Diffraction Precision Index (DPI) database provides the DPI value for all the macromolecular structures available in its archive, Protein Data Bank (PDB). It also provides the atomic coordinate error for each atom of the chosen three-dimensional structure. Further, the database has an option for the users to compute the above values for non-redundant structures or a group of structures. The database provides a modified PDB file which contains the DPI value for the given structure and the computed atomic coordinate error for each atom. The proposed DPI database allows the user to easily access all the entities available in the PDB or categorize the structures based on the experimental methods, resolution and DPI values. The database will be updated weekly, hence enabling users to access up-to-date information available in the PDB. DPI database is freely available round the clock and can be accessed through http://pranag.physics.iisc.ac.in/dpi_db/.

KEYWORDS: Diffraction precision index; database; macromolecular study; crystal data quality; atomic coordinate error.

Corresponding Author: Dr. K. Sekar* Ph.D.

Department of Computational and Data Sciences, Indian Institute of Science, Bangalore, India.

Email Address: sekar@iisc.ac.in

1. INTRODUCTION

More than 100,000 macromolecule structures have been determined from crystalline samples using X-ray crystallography, thus providing valuable information for understanding the biological processes at atomic level [2, 10, 20]. In principle, proteins unveil their functions in aqueous

environment, but their structures are more frequently examined inside crystals [15, 18]. It is well understood that proteins are not static structures [3], rather they are highly dynamic molecules, thereby governing most of the biological functions [7, 5, 17]. However, X-ray crystallography can only provide a single (or sometimes alternative) position for each atom [6, 24, 26] and further, it was found that the structures of the same protein exhibited coordinate differences of 1 Å or greater, by independent research groups. Thus, such disparity in the coordinate positions can be considered as 'coordinate uncertainties' or 'positional uncertainties', which may reflect on the actual flexibility of the protein structure [4, 9, 19]. Therefore, it is important to estimate the uncertainty in the atomic positions, which is crucial to identify genuine features or differences between the structures. The DPI database has been developed based on the experimental crystallographic parameters and is made available to acquire the experimental precision of the atomic coordinates of macromolecules. Using the calculation, derived earlier by our group [8, 14], the expected positional uncertainty for every atom in the macromolecule is predicted from the pertinent information present in the Protein Data Bank (PDB). The study performed by our group is the first of its kind to provide the error in the atomic coordinates and this in turn is useful to compute more reliable hydrogen bond distances. Thus, this is the only database available to estimate the DPI of macromolecules and the atomic coordinate error for each atom.

2. MATERIALS AND METHODS

2.1. Features of the database

DPI database encompasses a total of 122,480 macromolecules including DNA, RNA and protein structures and the numbers are illustrated in Table 1 [1, 29]. Also, the number of structures determined according to the experimental method and the percentage of non-redundant (ranges from 20% to 95%) structures are provided in the "About DPI" section (Fig. 1). DPI database comprises three search options, (a) Cut-off parameters, (b) Single structure and (c) Multiple structures. The "Cut-off parameters", allows the user to either select the non-redundant structures ranging from 20% to 95% or include all the structures known till date, under the "dataset" option. In addition, the search can be restricted to specific experimental methods, such as X-ray diffraction, Hybrid (X-ray diffraction and Neutron diffraction), Neutron diffraction and Electron crystallography. There is another option that can be used to filter the search by quality criteria cut-offs, such as resolution, crystallographic R factor and DPI. Using the second option, users can provide a PDB-id and the resulting page displays a set of vital parameters such as experimental method, resolution, R_{work}, R_{free}, DPI etc. The atomic coordinates of the macromolecular structure can be viewed by clicking "Display DPI added file", along with the calculated DPI value and coordinate errors of all the atoms, which are printed in red color and can be downloaded as text and pdf formats. The three-dimensional structure of the macromolecules can be visualized by choosing the appropriate options provided on the right-hand side of the "results" page, using the visualization plug-in, JSmol. The "Multiple

Structures" option allows the user to provide a list of PDB-ids to download the respective PDB files that have been modified to accommodate the computed DPI values and the atomic coordinate errors.

Table 1. Data-set statistics as on September 13, 2018

Macromolecules	**No. of structures available in DPI-Database**
DNA	770
RNA	758
Protein	115066
DNA & RNA	225
DNA & Protein	4424
RNA & Protein	2280
DNA, RNA, & Protein	339

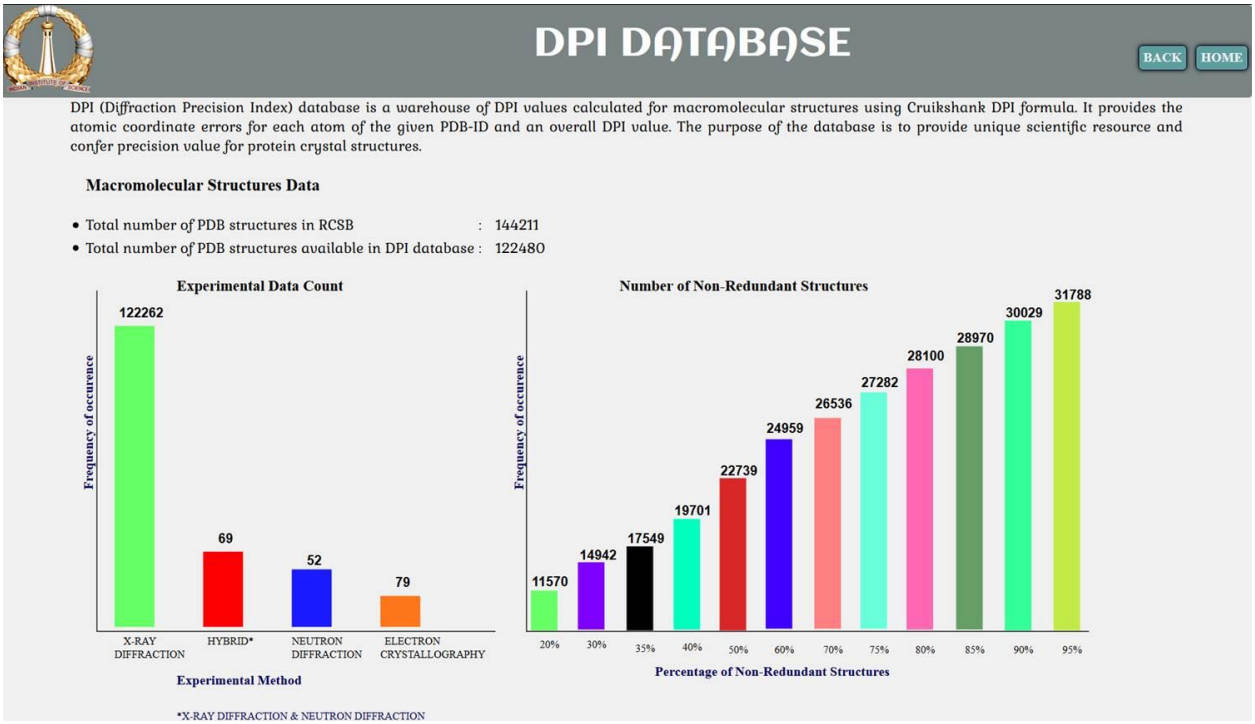

Fig. 1. The number of structures determined according to the experimental method and the percentage of non-redundant (ranges from 20% to 95%) structures.

2.2. System specification and database implementation

The front end (user interface) of the DPI database has been developed using CGI-PERL, HTML, JavaScript, jQuery and CSS, thus making the page more dynamic and easy for the users to effectively access and operate the database. The back end (calculation and data processing) has been created using PERL, MySQL and CPAN modules in order to harness the data. The computing engine has been designed on a high end processor, Intel-based CentOS 7 operating environment (Intel (R) Xeon (R) 10 core CPU E5-2630 V4 @ 2.20 GHz high end server class board with 32 GB main

memory) in order to facilitate speed and enrich the efficacy. The in-house PDB server, used by the computing engine, is being updated weekly with new entities from the parent PDB server (RCSB, University of California, San Diego). Thus, researchers can easily access all the macromolecular structures available in its archives. The graphical view of the chosen macromolecular structure is enabled by visualization plug-in, JSmol.

3. RESULTS AND DISCUSSION

The enormous availability of protein crystal structures in the PDB exhibits that most of the entities share a sequence and structure similarity. Therefore, while doing data mining studies, it is imperative to select a representative structure from a pool of similar or highly homologous structures to obtain unbiased results. In principle, the parameters like sequence identity, resolution and R_{work} assist in the selection of the finest quality model to be used as a reference model. Subsequently, a study [9] stated that the resolution alone cannot be used as a criterion to select a representative protein structure and is considered to be only one of the quantitative measures. Therefore, the DPI value computed based on the dynamic crystallographic parameters would be a more effective quality criterion parameter to select a representative model. To illustrate this, two examples have been discussed below.

3.1. Identifying a structure from a set of homologous structures

To expose the significance of DPI, specifically to a protein, a case study has been demonstrated (Fig. 2), using four calmodulin structures [16, 25, 27, 28], having identical resolutions of 2 Å each. Amongst the four structures, the PDB-id 5J03 [25] has the least DPI value, thereby being the representative structure that could be used as a reference structure to compare with the rest of the three structures. It is worth noting that the R_{work} for the chosen structure is also the lowest.

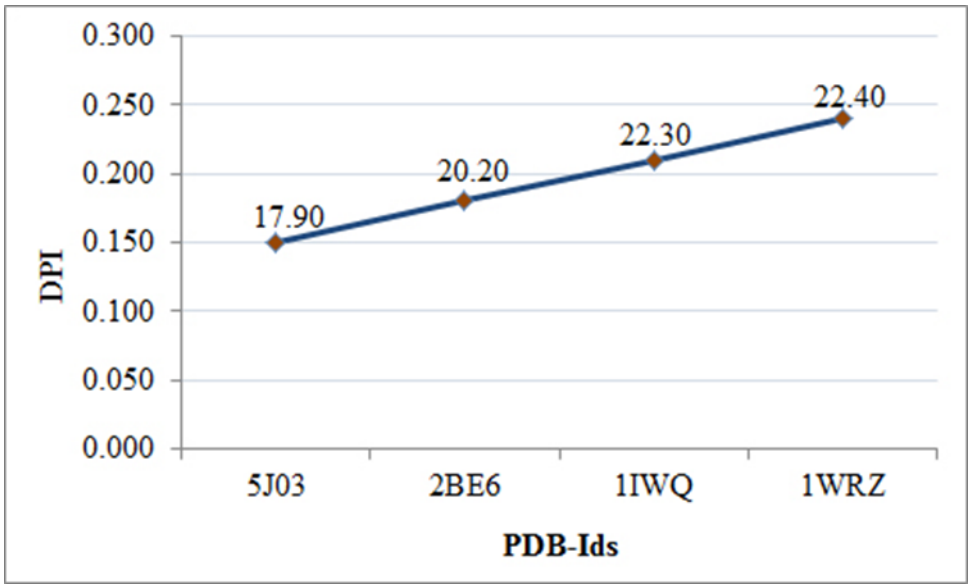

Fig. 2. The DPI value for the selected four calmodulin structures, each having Resolution = 2 Å, has been plotted. The data label represents the R_{work}.

To further illustrate this aspect bovine pancreatic phospholipase A_2 [13, 21, 22, 30] was selected, a previous mining study done on a total of 25 bovine pancreatic phospholipase A2 [12] has been referred too. In this study, the structural and functional role of water molecules in these structures have been examined, for which a fixed or reference molecule had to be selected. The structure, PDB-id 1G4I [23], with the largest number of water molecules and the lowest values for resolution and R_{work} was selected as the reference molecule. Interestingly, among all these structure, 1G4I possesses the lowest DPI value = 0.012 Å. Thus, it is clear that instead of manually correlating various parameters to arrive at the best reference structure, it can be achieved merely by considering only the DPI value.

3.2. Selecting structures for data mining study from 25% non-redundant structures

Here, a more elementary scenario has been described. In many data mining studies, it is a common practice to use structures from the 25% non-redundant or non-homologous dataset, which consists of 6,401 structures (with R_{work} <= 20% and Resolution <= 2 Å). However, when the parameter, DPI <= 0.5 Å was used, the number of structures reduced to 6,349. When the DPI cut-off was made more stringent, <= 0.1 Å, the number of resultant structures further reduced to 3,764, thus yielding a more accurate set of protein structures. This difference in the number of structures undeniably exemplifies the power of DPI used to obtain a true and more accurate set of non-redundant structures for data mining studies, among the 25% non-redundant dataset, given the desired cut-off values. From these case studies, the emphasis that DPI is a vital parameter for selecting the representative model for further unbiased structural studies is clearly seen.

4. CONCLUSION

DPI database is an open-access knowledgebase with up-to-date structural information on the macromolecules available in the PDB. It provides the DPI for a given structure and also offers pivotal information on the atomic coordinate error for each atom in the three-dimensional structure. As is evident from the case studies, considering only the conventional parameters for filtering out relevant structures for data mining studies would most definitely lead to a misleading or biased dataset. Thus, the incorporation of DPI is of utmost importance to obtain a more unbiased and truly representative dataset. Undoubtedly, the proposed DPI database will serve as a highly focused knowledgebase to the scientific community studying three-dimensional protein structures, especially structural biologists and bioinformaticians.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the facilities offered by the centre of excellence in structural biology and bio-computing funded by the Department of Biotechnology (DBT) and Department of Computational and Data Sciences. The author (KS) thanks DBT for a possible financial support in the form of a research grant.

CONFLICT OF INTEREST

The authors have not declared any conflict of interests.

REFERENCES

1. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Research. 2000 January 1, 2000; 28(1):235-42.
2. Chaudhary SK, Jeyakanthan J, Sekar K. Cloning, expression, purification, crystallization and preliminary X-ray crystallographic study of thymidylate kinase (TTHA1607) from Thermus thermophilus HB8. Acta Crystallogr Sect F Struct Biol Cryst Commun. 2013; 69(Pt 2):118-21.
3. Chen YY, Ko TP, Chen WH, Lo LP, Lin CH, Wang AH. Conformational changes associated with cofactor/substrate binding of 6-phosphogluconate dehydrogenase from Escherichia coli and Klebsiella pneumoniae: Implications for enzyme mechanism. J Struct Biol. 2010; 169(1):25-35.
4. Cruickshank D. Remarks about protein structure precision. Acta Crystallographica Section D. 1999; 55(3):583-601.
5. Dauter Z, Wlodawer A, Minor W, Jaskolski M, Rupp B. Avoidable errors in deposited macromolecular structures: an impediment to efficient data mining. IUCrJ. 2014; 1(3):179-93.
6. Davis IW, Arendall WB, 3rd, Richardson DC, Richardson JS. The backrub motion: how protein backbone shrugs when a sidechain dances. Structure. 2006; 14(2):265-74.
7. Fisher SJ, Blakeley MP, Cianci M, McSweeney S, Helliwell JR. Protonation-state determination in proteins using high-resolution X-ray crystallography: effects of resolution and completeness. Acta Crystallographica Section D. 2012; 68(7):800-9.
8. Gurusaran M, Shankar M, Nagarajan R, Helliwell JR, Sekar K. Do we see what we should see? Describing non-covalent interactions in protein structures including precision. IUCrJ. 2013; 1(Pt 1):74-81.
9. Gurusaran M, Sivaranjan P, Dinesh Kumar KS, Radha P, Thulaa Tharshan KPS, Satheesh SN, et al. Hydrogen Bonds Computing Server (HBCS): an online web server to compute hydrogen-bond interactions and their precision. Journal of Applied Crystallography. 2016; 49(2):642-5.
10. Handing KB, Niedzialkowska E, Shabalin IG, Kuhn ML, Zheng H, Minor W. Characterizing metal-binding sites in proteins with X-ray crystallography. Nature Protocols. 2018; 13:1062.
11. Hawkins PC, Warren GL, Skillman AG, Nicholls A. How to do an evaluation: pitfalls and traps. J Comput Aided Mol Des. 2008; 22(3-4):179-90.
12. Kanaujia SP, Sekar K. Structural and functional role of water molecules in bovine pancreatic phospholipase A (2): a data-mining approach. Acta Crystallogr D Biol Crystallogr. 2009; 65(Pt 1):74-84.
13. Kanaujia SP, Sekar K. Structures and molecular-dynamics studies of three active-site mutants of bovine pancreatic phospholipase A(2). Acta Crystallogr D Biol Crystallogr. 2008; 64(Pt 10):1003-11.

14. Kumar KSD, Gurusaran M, Satheesh SN, Radha P, Pavithra S, Thulaa Tharshan KPS, et al. Online_DPI: a web server to calculate the diffraction precision index for a protein structure. Journal of Applied Crystallography. 2015; 48(3):939-42.
15. Kurauskas V, Izmailov SA, Rogacheva ON, Hessel A, Ayala I, Woodhouse J, et al. Slow conformational exchange and overall rocking motion in ubiquitin protein crystals. Nat Commun. 2017; 8(1):017-00165.
16. Kursula P, Vahokoski J, Wilmanns M. The mode of binding of calmodulin to death-associated protein kinases: http://www.rcsb.org/structure/1WRZ.
17. Phillips GN, Jr. Describing protein conformational ensembles: beyond static snapshots. F1000 Biol Rep. 2009; 1(38):B1-38.
18. Rajakannan V, Yogavel M, Poi M-J, Jeyaprakash AA, Jeyakanthan J, Velmurugan D, et al. Observation of Additional Calcium Ion in the Crystal Structure of the Triple Mutant K56,120,121M of Bovine Pancreatic Phospholipase A2. Journal of Molecular Biology. 2002 2002/12/06/; 324(4):755-62.
19. Rashin AA, Rashin AH, Jernigan RL. Protein flexibility: coordinate uncertainties and interpretation of structural differences. Acta Crystallogr D Biol Crystallogr. 2009; 65(Pt 11):1140-61.
20. Ronda L, Bruno S, Bettati S, Storici P, Mozzarelli A. From protein structure to function via single crystal optical spectroscopy. Frontiers in Molecular Biosciences. 2015; 2:12.
21. Sekar K, Eswaramoorthy S, Jain MK, Sundaralingam M. Crystal structure of the complex of bovine pancreatic phospholipase A2 with the inhibitor 1-hexadecyl-3-(trifluoroethyl)-sn-glycero-2-phosphomethanol. Biochemistry. 1997; 36(46):14186-91.
22. Sekar K, Sundaralingam M. High-resolution refinement of orthorhombic bovine pancreatic phospholipase A2. Acta Crystallogr D Biol Crystallogr. 1999; 55(Pt 1):46-50.
23. Steiner RA, Rozeboom HJ, de Vries A, Kalk KH, Murshudov GN, Wilson KS, et al. X-ray structure of bovine pancreatic phospholipase A2 at atomic resolution. Acta Crystallogr D Biol Crystallogr. 2001; 57(Pt 4):516-26.
24. Stroud RM, Fauman EB. Significance of structural changes in proteins: expected errors in refined protein structures. Protein Sci. 1995; 4(11):2392-404.
25. Strulovich R, Tobelaim WS, Attali B, Hirsch JA. Structural Insights into the M-Channel Proximal C-Terminus/Calmodulin Complex. Biochemistry. 2016; 55(38):5353-65.
26. Van den Bedem H, Lotan I, Latombe JC, Deacon AM. Real-space protein-model completion: an inverse-kinematics approach. Acta Crystallogr D Biol Crystallogr. 2005; 61(Pt 1):2-13.
27. Van Petegem F, Chatelain FC, Minor DL, Jr. Insights into voltage-gated calcium channel regulation from the structure of the CaV1.2 IQ domain-Ca2+/calmodulin complex. Nat Struct Mol Biol. 2005; 12(12):1108-15.

28. Yamauchi E, Nakatsu T, Matsubara M, Kato H, Taniguchi H. Crystal structure of a MARCKS peptide containing the calmodulin-binding domain in complex with Ca2+-calmodulin. Nat Struct Biol. 2003; 10(3):226-31.
29. Yang H, Tan L, Sala R, Burley SK, Hudson BP, Bhikadiya C, et al. RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. Nucleic Acids Research. 2018; 47(D1):D464-D74.
30. Yu B-Z, Poi MJ, Ramagopal UA, Jain R, Ramakumar S, Berg OG, et al. Structural Basis of the Anionic Interface Preference and Activation of Pancreatic Phospholipase A2. Biochemistry. 2000 2000/10/01; 39(40):12312-23.