6-2012

P.R.E.S.S. – An R-package for Exploring Residual-Level Protein Structural Statistics

Yuanyuan Huang
Iowa State University

Steve Bonett
Iowa State University

Andrzej Klockowski
Iowa State University

Robert Jernigan
Iowa State University, jernigan@iastate.edu

Zhijun Wu
Iowa State University, zhijun@iastate.edu

P.R.E.S.S. – An R-package for Exploring Residual-Level Protein Structural Statistics

Abstract
P.R.E.S.S. is an R package developed to allow researchers to get access to and manipulate on a large set of statistical data on protein residue-level structural properties such as residue-level virtual bond lengths, virtual bond angles, and virtual torsion angles. A large set of high-resolution protein structures are downloaded and surveyed. Their residue-level structural properties are calculated and documented. The statistical distributions and correlations of these properties can be queried and displayed. Tools are also provided for modeling and analyzing a given structure in terms of its residue-level structural properties. In particular, new tools for computing residue-level statistical potentials and displaying residue-level Ramachandran-like plots are developed for structural analysis and refinement. P.R.E.S.S. will be released in R as an open source software package, with a user-friendly GUI, accessible and executable by a public user in any R environment.

Keywords
Protein structure analysis, protein residual-level structural properties, structural bioinformatics, statistical potentials, structural correlation plots

Disciplines
Biochemistry, Biophysics, and Structural Biology | Bioinformatics | Mathematics | Statistical Models

Comments
This is a manuscript of an Electronic version of an article published as Journal of Bioinformatics and Computational Biology 10, no. 03 (2012): 1242007, DOI: 10.1142/S0219720012420073.

HHS Public Access

Author manuscript

J Bioinform Comput Biol. Author manuscript; available in PMC 2015 March 25.

Published in final edited form as:

J Bioinform Comput Biol. 2012 June ; 10(3): 1242007. doi:10.1142/S0219720012420073.

P.R.E.S.S. – An R-package for Exploring Residual-Level Protein Structural Statistics

Yuanyuan Huang,

Program on Bioinformatics and Computational Biology, Department of Mathematics, Iowa State University, Ames, Iowa 50011

Steve Bonett,

Summer REU Program on Computational Systems Biology, Iowa State University, Ames, Iowa 50011

Andrzej Klockowski,

Department of Biochemistry, Biophysics, and Molecular Biology, Iowa State University, Ames, Iowa 50011

Robert Jernigan, and

Department of Biochemistry, Biophysics, and Molecular Biology, Iowa State University, Ames, Iowa 50011

Zhijun Wu

Program on Bioinformatics and Computational Biology, Department of Mathematics, Iowa State University, Ames, Iowa 50011

Yuanyuan Huang: sunnyuan@iastate.edu; Steve Bonett: sbonett@gmail.com; Andrzej Klockowski: kloczkow@iastate.edu; Robert Jernigan: jernigan@iastate.edu; Zhijun Wu: zhijun@iastat.edu

Abstract

P.R.E.S.S. is an R package developed to allow researchers to get access to and manipulate on a large set of statistical data on protein residue-level structural properties such as residue-level virtual bond lengths, virtual bond angles, and virtual torsion angles. A large set of high-resolution protein structures are downloaded and surveyed. Their residue-level structural properties are calculated and documented. The statistical distributions and correlations of these properties can be queried and displayed. Tools are also provided for modeling and analyzing a given structure in terms of its residue-level structural properties. In particular, new tools for computing residue-level statistical potentials and displaying residue-level Ramachandran-like plots are developed for structural analysis and refinement. P.R.E.S.S. will be released in R as an open source software package, with a user-friendly GUI, accessible and executable by a public user in any R environment.

Correspondence to: Zhijun Wu, `zhijun@iastat.edu`.

Authors' Contributions: YH collected the data, wrote the computer programs for the calculations, performed the analysis, and drafted the paper. SB was responsible for the development of the graphics interface and the public release of the software. AK, RJ, and ZW analyzed the data and interpreted the results. YH, RJ, and ZW finished the writing.

Keywords

Protein structure analysis; protein residual-level structural properties; structural bioinformatics; statistical potentials; structural correlation plots

1. Introduction

The atomic level structural properties of proteins, such as bond lengths, bond angles, and torsion angles, have been well studied and understood based on either chemistry knowledge or statistical analysis [1,2]. For example, we have learned that the bond lengths and bond angles are relatively fixed for given types of bonds, and the torsion angles have preferences. The knowledge on these properties has been crucial for both theoretical and experimental approaches to protein modeling. Potential functions have been defined for bond lengths, bond angles, and torsion angles using their known or preferred values [3]. Energetically favourable structures can be obtained when an energy function that contains these potentials along with some other non-bond potentials is minimized [4]. In either NMR or X-ray crystallography, these properties have been used to refine an initial experimental model, which may otherwise have little atomic details. In particular, in NMR, the experimental data is mainly for hydrogen-hydrogen interactions, which is insufficient for determining a structure if without the pre-existing data on bond lengths and bond angles [5]. The correlations among these properties have been an important source of information as well. For example, a statistical analysis showed that the ϕ-ψ torsion angles around the two bonds of the C_a atom in the backbones of the residues have a special correlation: When the ϕ angle is chosen for some value, the ψ angle has only a restricted range of choice, and vice versa. The information on this correlation has been employed in both experimental determination and theoretical prediction of protein structures [6,7]. The statistical distribution of the ϕ-ψ angles in known proteins has been depicted in a two-dimensional plane called the Ramachandran Plot named after the biophysicist G. N. Ramachandran who first did the statistical survey [8]. The Ramachandran Plot has been widely used for structure evaluation. By evaluating the ϕ-ψ angles for all the residues in a given protein structure and putting them in the Ramachandran Plot, one can tell whether or not the structure is well formed based on how many of the ϕ-ψ angle pairs are in the densest regions of the Plot.

Structural properties similar to those described above can also be found at the residue level such as the distances between two neighboring residues; the angels formed by three residues in sequence; and the torsion angles of four residues in sequence. Proteins are often modelled in a reduced form, with residues considered as basic units. The residue distances and angles then become crucial for the description of the model, and they can be as important as those at the atomic level for structural determination, prediction, and evaluation. The knowledge on these distances and angles can also be used to define residue level potential functions so that potential energy minimization and dynamics simulation can be performed more effectively and efficiently at residue instead of atomic level, because the number of variables may be reduced in magnitudes and the time step may be increased [9,10]. However, the residue distances and angles have not been examined and documented in a similar scale as those at the atomic level. The reason is that they are not easy to measure directly; the physics for the

interactions between residues is not as clear; and they are not as rigid as the bond lengths and bond angles, i.e., their values may vary in a wide range.

While residue distances and angles are difficult to measure experimentally, they can be estimated statistically, based on their distributions in known protein structures. Such approaches have been used for extracting residue contact statistics starting in early 1980s [11]; for developing residue level distance-based mean-force potentials [12] for refining X-ray crystallography determined structures [13,14] and for deriving distance and angle constraints and potentials for NMR structure refinement [15,16,17,18]. Several online databases have also been built for direct access to the statistical data on various types of distances or angles [19,20]. In our recent work [21], we have downloaded a large number of high-resolution X-ray structures from PDB Data Bank [22], and collected and analyzed several important residue-level structural properties including the distances between two neighboring residues; the angles formed by three residues in sequence; and the torsion angles of four residues in sequence. We call them, respectively, the residue level virtual bond lengths, virtual bond angles, and virtual torsion angles. We have examined the statistical distributions of these virtual bonds and virtual angles in known protein structures. In a four-residue sequence, there are two virtual bond angles and one torsion angle in between. We name them, according to their order in the sequence, the α-angle, τ-angle, and β-angle, where τ is the torsion angle (Fig. 1a). In a five-residue sequence, there are three virtual bond angles and two torsion angles. We name them, according to their order in the sequence, the α-angle, τ_1-angle, β-angle, τ_2-angle, γ-angle, where τ_1 and τ_2 are torsion angles (Fig. 1b). For these sequences, we have investigated the correlations among some of associated angles and in particular, the α-τ-β correlations for four-residue sequences and τ_1-β-τ_2 correlations for five-residue sequences. We have shown that the distributions of residue distances and angles may vary with varying residue sequences, but in most cases, are concentrated in some high probability ranges, corresponding to their frequent occurrences in either α-helices or β-sheets in proteins. We have shown that between α and τ angles and τ and β angles, there exist strong correlations, which suggests that proteins follow certain rules to form their residue level angles as well, just like those for their atomic level ϕ-ψ angles. To the authors knowledge, these properties have not been discovered and documented before, but can be very valuable in applications [21]. In this paper, we describe a related piece of work with [21] on developing a software package called P.R.E.S.S. for direct access to the statistical data on the residue-level structural properties we have collected and analyzed. The software is developed in R [23] and will be released as an open source package, with a user-friendly GUI, accessible and executable by a public user in any R environment. With this software, the distributions and correlations of given types of residue distances or angles can all be retrieved and displayed. Tools are also provided in P.R.E.S.S. for modeling and analyzing a given structure in terms of its residue-level structural properties. In particular, tools for computing residue-level statistical potentials and displaying residue-level Ramachandran-like plots are developed for structural analysis and refinement. We describe the organization of the software, the data source, the computational methods, and all the functional modules. We provide examples to demonstrate the use of the software.

2. Graphics Interface and functional module

In the R package there are two major components: the GUI (graphics user interface) and the computational unit. The GUI takes a query from the user and passes it to the computational unit. The computational unit has a collection of routines, responsible for various computational tasks. It retrieves the data from the databases in the back end, performs certain calculations, and returns the results to the GUI. The interface then displays the results. More specifically, the GUI shows a window of six functional panels (Fig. 2), each accepting a specific type of queries: 1). Queries on virtual bond lengths for two residues. 2). Queries on virtual bond angles for three residues. 3). Queries on virtual torsion angles and ATA correlations (α-τ-β angle sequences) for four residues. 4. Queries on TAT correlations (α-τ_2-β-τ_2-γ angle sequences) for five residues. 5. Structural analysis and evaluation. 6. Help information. The computational unit has mainly the following routines:

- Routine for calculating the distribution of the virtual bond length (**B**) between a given pair of residues.
- Routine for calculating the distribution of the virtual bond angle (**A**) for a given sequence of three residues.
- Routine for calculating the distribution of the residue-level 1-3-distance for a given sequence of three residues.
- Routine for calculating the distribution of the virtual torsion angle (**T**)for a given sequence of four residues.
- Routine for calculating the distribution of the residue-level 1-4-distance for a given sequence of four residues.
- Routine for calculating the correlation between the residue-level 1-3- distance and the virtual bond angle for a given sequence of three residues.
- Routine for calculating the correlation between the residue-level 1-4-distance and the virtual torsion angle for a given sequence of four residues.
- Routine for calculating the correlations of the α-τ-β (**ATA**) angle sequence on a given sequence of four residues.
- Routine for calculating the correlation of the α-τ_1-β-τ_2-γ (**TAT**) angle sequence on a given sequence of five residues.
- Routine for evaluating the statistical potentials for the virtual bond lengths or the virtual bond angles for a given protein structure.
- Routine for evaluating the α-τ (**AT**) and τ-β (**TB**) angle pairs for a given protein structure and display them in α-τ (**AT**) and τ-β (**TB**) density distribution contour plots.

The overall system organization of P.R.E.S.S. is shown in Fig. 3. We describe the data source and computational methods used in P.R.E.S.S. in the following section.

3. Functional Modules

3.1. Distribution of Virtual Bond Lengths

One of functions of P.R.E.S.S. is to retrieve the virtual bond lengths for a given pair of residues and find the distribution of the particular bond length over a certain distance range. The found distribution can be displayed in a graph as shown in Fig.4. The residue pair to be searched for can be specified from a pull-down menu. Each residue can be a specific or any type. For the latter, any type is considered for that residue. The bin size of the distribution graph can be adjusted. The graph can be displayed to show either the frequency or density of the bond lengths.

3.2. Distribution of Virtual Bond Angles

One of functions of P.R.E.S.S. is to retrieve the virtual bond angles for a given sequence of three residues and find the distribution of the particular bond angle over a certain angle range. The found distribution can be displayed in a graph as shown in Fig. 5. The residue triplet to be searched for can be specified from a pull-down menu. Each residue can be a specific or any type. For the latter, any type is considered for that residue. The bin size of the distribution graph can be adjusted. The graph can be displayed to show either the frequency or density of the bond angles.

3.3. Angle-Distance Correlations

When the distribution of a virtual bond angle is queried, an option is available for displaying the correlation between the bond angle and the corresponding residue 1-3-distance. This correlation can be requested for any sequence of three residues, as shown in Fig. 6.

3.4. Distribution of Virtual Torsion Angles

The virtual torsion angles for a given sequence of four residues can be retrieved. The distribution of the particular torsion angle can be displayed over a certain angle range. The residue quadruplet to be searched for can be specified from a pull-down menu. Each residue can be a specific or any type. For the latter, any type is considered for that residue. The bin size of the distribution graph can be adjusted. The graph can be displayed to show either the frequency or density of the torsion angles. The graph can be displayed along with the distributions of the neighboring virtual bond angles (α,β), as shown in Fig. 7. The density distribution of the angle sequence α-τ-β can be displayed as a 3D plot in α-τ-β space, as shown Fig. 8. The correlation between the virtual torsion angle and the corresponding residue 1-4 distance for a given sequence of four residues can also be displayed.

3.5. Correlation of Virtual Torsion Angles

The angle sequence α-τ_1-β-τ_2-γ for a given sequence of five residues can be retrieved. The density distributions of the virtual bond angle β and its neighboring two virtual torsion angles can be displayed. The graphs can be displayed in a matrix of plots, as shown in Fig. 9. The density distribution of τ_1-β-τ_2 can also be displayed as a 3D plot in the τ_1-β-τ_2 space as shown Fig. 10.

3.6. Structural Analysis – Computation of Statistical Potentials

One of important functions of P.R.E.S.S. is to evaluate the statistical potentials on the virtual bonds or virtual bond angles for a given structure (Fig. 11). The potentials are defined in terms of the statistical distributions of the virtual bond lengths and virtual bond angles. The virtual bond length potential can be evaluated for every neighboring pair of residues of the given structure. Therefore, the distribution of the potential energy along the residue sequence of the structure can be obtained and displayed to show how flexible the virtual bonds are along the sequence. The higher the potential energy is for a specific bond, the lower the probability of the bond length is in the distribution of the bond length in known proteins, and hence the more deviated it must be from its average value the bond length (Fig. 12). The virtual bond angle potential can be evaluated for every sequence of three residues of the given structure as well. The distribution of the potential energy along the residue sequence of the structure can also be obtained and displayed to show how flexible the virtual bond angles are along the sequence. It has the same property as that for the bond length energy for structural evaluation (Fig. 13).

3.7. Structural Analysis – Residue Angle-Angle Correlation Plots

One of the most important functions of P.R.E.S.S. is that it can evaluate the correlations of the virtual bond and torsion angles and display a residue-level Ramachandran-like plot for a given structure. Two of the angle-angle correlation plots are proven to be expecially valuable. One is the α-τ correlation plot or the AT-plot for short. Another one is the τ-β correlation plot or the TB-plot for short. Given a protein structure, the α-τ or τ-β angle pairs can be computed along the residue sequence for the structure. Each angle pair can be plotted as a dot in the α-τ or τ-β space. The distribution of the dots over the contour of the general α-τ or τ-β density distribution can then be evaluated to show how the angle pairs in the given structure correlated against to their average correlations in known proteins. These plots can be used effectively to differentiate high-quality structures from low-quality structures at the residue level as the Ramachandran plots for structural evaluation at the atomic level, as shown in Fig. 13 and 14.

We investigated all the obsoleted structures recorded in Protein Data Bank and the comparison of the superseded sturctures with the current sturctures in the general α-τ or τ-β density distribution plots is summarized below in the table regarding to differenct catagories of RMSD values. As of Feb 08, 2012 there are totally 1,654 obsoleted proteins superseded by successors according to the list from PDB (ftp://ftp.wwpdb.org/pub/pdb/data/status/obsolete.dat). We compare these structures in pairs. With the percentages of points fall into allowed region, favoured region, and most favoured region (in Table 1 and Table 2), we check if there was actually an improvement by performing a t-test for the paired data. The p-values of the one-tailed student t-test support the hypothesis that the current structures have more points in the densed region of the angle-angle correlation plots than the preivously obsoleted structures. We observed that for the protein pairs with RMSD less than 3A but bigger than 1A, the gap of the percentage of points in α-τ or β-τ correlation plots is the biggest between the superseded ones and successors. The boxplots (Fig. 16 and 17) indicate strong evidence the median and mean differ in these two groups of structures.

4. Summary and Discussion

In this paper, we have reported our recent work for the development of an R package, called P.R.E.S.S., which allows researchers to get access to and manipulate on a large set of statistical data on protein residue-level structural properties such as residue-level virtual bond lengths, virtual bond angles, and virtual torsion angles. We have downloaded and surveyed a large set of high-resolution protein structures, and calculated and documented an important set of their residue-level structural properties in P.R.E.S.S. With P.R.E.S.S., the statistical distributions and correlations of these properties can be queried and displayed. Tools are also provided for modeling and analyzing a given structure in terms of its residue-level structural properties. In particular, new tools for computing residue-level statistical potentials and displaying residue-level Ramachandran-like plots are developed for structural analysis and refinement. We have discussed the principle for the development of P.R.E.S.S., for statistical analysis on protein structures. We have described the system organization and interface of the software, and provided detailed information on how the structural data was collected and documented in P.R.E.S.S., and how all the statistical results were calculated. We have described the major computational and analysis functions of P.R.E.S.S. and demonstrated them in many examples.

P.R.E.S.S. will be released in R as an open source software package, with a user-friendly GUI, accessible and executable by a public user in any R environment. The statistical distributions of residue-level distances and angles in known protein structures provide a valuable source of information for estimating these residue level structural properties of proteins, which are not otherwise accessible experimentally. However, these statistical measures rely upon the quality as well as quantity of the sampled known structures. We have downloaded around one thousand high-quality structures from the PDB, which should be sufficient to obtain reliable statistical estimates of the distributions of virtual bond lengths, virtual bond angles, virtual torsion angles, and some of their correlations, but of course there is the possibility that for some cases of specific residue sequences, the values might deviate from the overall characteristic distributions. In P.R.E.S.S., we have provided information about the size of the data set for each estimate. The useful tool from this study is a residue-level Ramamchandran-type of plot for correlations between pairs of neighboring virtual bond angles and virtual torsion angles. Several examples have been given in the present paper, but these differ from the atomic-level Ramachandran Plot in an important way, because the density distribution contours of these residue-level angles show relatively larger deviations. Thus their use requires specifying more precisely what density regions should be permitted for high-quality structures. Further evaluations are needed to decide generally what these evaluation criteria should be.

Acknowledgments

This work is partially supported by the NIH//NIGMS grant R01GM081680 and by the NSF-DMS grant DMS0914354.

References

1. Creighton, TE. Proteins: Structures and Molecular Properties. 2nd. Freeman and Company; 1993.

2. Dunbrack RL. Rotamer libraries in the 21st century. Curr Opin Struct Biol. 2002; 12:431–440. [PubMed: 12163064]
3. Brooks, CL., III; Karplus, M.; Pettitt, BM. Proteins: A Theoretical Perspective of Dynamics, Structure, and Thermodynamics. Wiley; 1989.
4. Schlick, T. Molecular Modeling and Simulation: An Interdisciplinary Guide. Springer; 2003.
5. Wüthrich, K. NMR in Structural Biology. World Scientific Publishing Company; 1995.
6. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structures. Journal of Applied Crystallography. 1993; 26:283–291.
7. Bernasconi A, Segre AM. Ab initio methods for protein structure prediction: A new technique based on Ramachandran plots. ERCIM News. 2000; 43:13–14.
8. Ramachandran GN, Sasiskharan V. Conformation of polypeptides and proteins. Advan Prot Chem. 1968; 23:283–437.
9. Skolnick J, Kolinski A, Ortiz AR. Reduced protein models and their application to the protein folding problem. J Biomol Struct Dyn. 1998; 16:381–396. [PubMed: 9833676]
10. Scheraga HA, Khalili M, Liwo A. Protein-folding dynamics: Overview of molecular simulation techniques. Annual Review of Physical Chemistry. 2007; 58:57–83.
11. Miyazawa S, Jernigan RL. Estimation of effective inter-residue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules. 1985; 18:534–552.
12. Sippl MJ. Calculation of conformational ensembles from potentials of mean force. J Mol Biol. 1990; 213:859–883. [PubMed: 2359125]
13. Rojnuckarin A, Subramaniam S. Knowledge-based potentials for protein structure. Proteins: Structure, Function, and Genetics. 1999; 36:54–67.
14. Wall ME, Subramaniam S, Phillips GN Jr. Protein structure determination using a database of inter-atomic distance probabilities. Protein Science. 1999; 8:2720–2727. [PubMed: 10631988]
15. Kuszewski J, Gronenborn AM, Clore GM. Improving the quality of NMR and crystallographic protein structures by means of a conformational database potential derived from structure databases. Protein Science. 1996; 5:1067–1080. [PubMed: 8762138]
16. Cui F, Jernigan R, Wu Z. Refinement of NMR-determined protein structures with database derived distance constraints. J Bioinform Comput Biol. 2005; 3:1315–1330. [PubMed: 16374909]
17. Cui F, Mukhopadhyay K, Young W, Jernigan R, Wu Z. Improvement of under-determined loop regions of human prion protein by database derived distance constraints. International Journal of Data Mining and Bioinformatics. 2009; 3:454–468. [PubMed: 20052907]
18. Wu D, Jernigan R, Wu Z. Refinement of NMR-determined protein structures with database derived mean force potentials. Proteins: Structure, Function, Bioinformatics. 2007; 68:232–242.
19. Wu D, Cui F, Jernigan R, Wu Z. PIDD: A protein inter-atomic distance distribution database. Nucleic Acid Research. 2007; 35:D202–D207.
20. Sun X, Wu D, Jernigan R, Wu Z. PRTAD: A protein residue torsion angle distribution database. International Journal of Data Mining and Bioinformatics. 2009; 3:469–482. [PubMed: 20052908]
21. Huang Y, Bonett S, Kloczkowski A, Jernigan R, Wu Z. Statistical Measures on Protein Residue-level Structural Properties. J Struct Funct Genomics. 2011 Jul.12(2):119. 18. [PubMed: 21452025]
22. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Research. 2000; 28:235–242. [PubMed: 10592235]
23. Grant, Rodrigues, ElSawy, McCammon, Caves. Bio3D: An R package for the comparative analysis of protein structures. Bioinformatics. 2006; 22:2695–2696. [PubMed: 16940322]
24. Doreleijers JF, Mading S, Maziuk D, Sojourner K, Yin L, Zhu J, Makley JL, Ulrich EL. BioMagResBank database with sets of experimental NMR constraints corresponding to the structures of over 1400 biomolecules deposited in the Protein Data Bank. J Biomol NMR. 2003; 26:139–146. [PubMed: 12766409]
25. Bourne, PE.; Weissig, H. Structural Bioinformatics. John Wiley & Sons, Inc.; 2003.

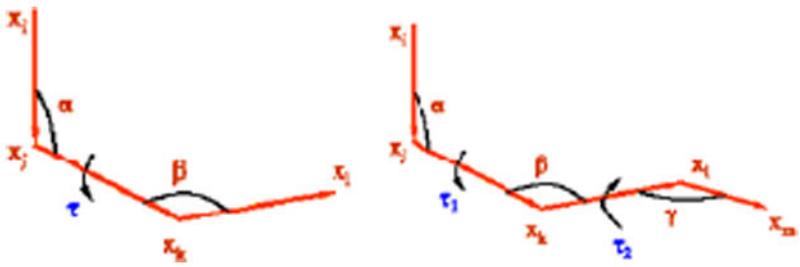

Fig. 1. Residue distances and angles

(a) The α-τ-β angle triplet in a four-residue sequence. (b) The α-τ_1-β-τ_2-γ angle quadruple in a five-residue sequence. The residues are assumed to be located at x_i, x_j, x_k, x_l, x_m.

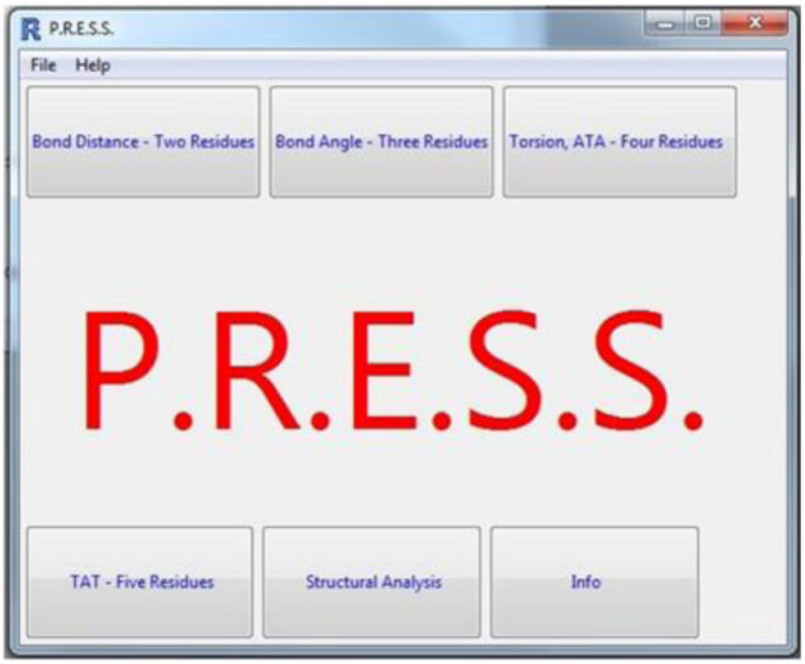

Fig. 2. P.R.E.S.S. graphics interface

PRESS has a graphics interface with six functional panels corresponding six functional routines, each providing a specific structural computing or analysis function.

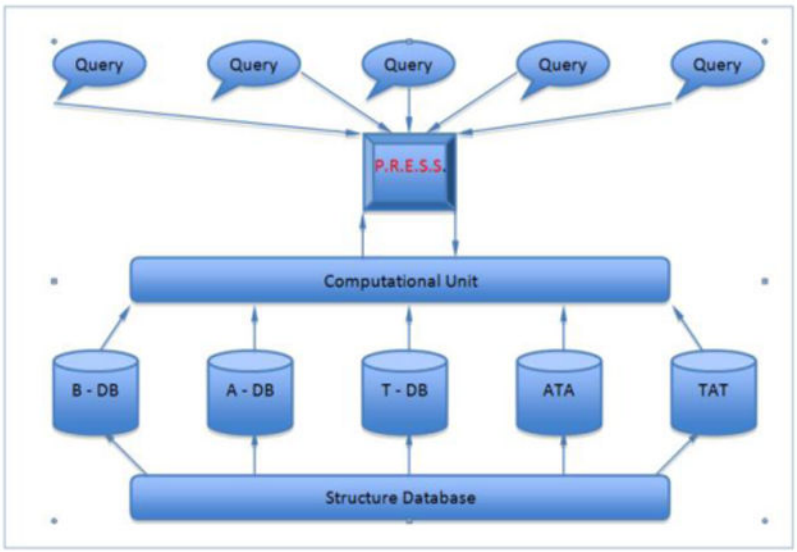

Fig. 3. System organization and interface

The system has a GUI, which takes queries on distributions or correlations of residue-level distances or angles and passes them to the computational unit. The computational unit retrieves the data from the distance or angle databases documented from the structural database, and computes the requested distributions or correlations. The results are returned to the interface and displayed.

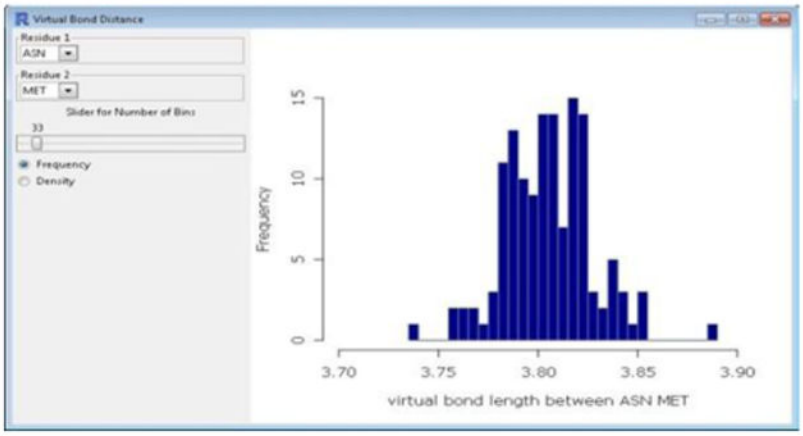

Fig. 4. Distribution of virtual bond lengths

This snapshot shows the distribution graph for the virtual bond lengths between ASN and MET. The users can not only move the slider to adjust the bin size of the histogram, but also switch between frequency and density displays.

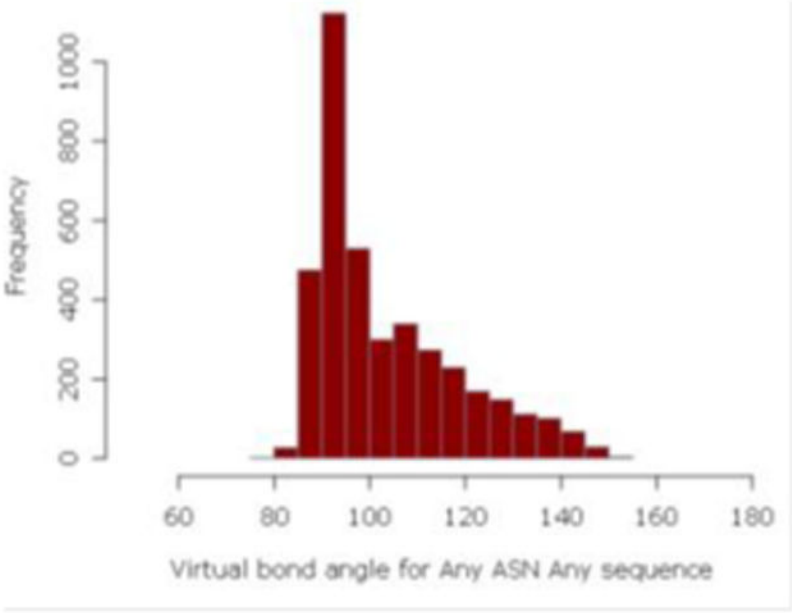

Fig. 5. Distribution of virtual bond angles formed by Any, ASN, and Any

This snapshot shows the distribuion graph for the virtual bond angles formed by a residue sequence Any, ASN, and Any.

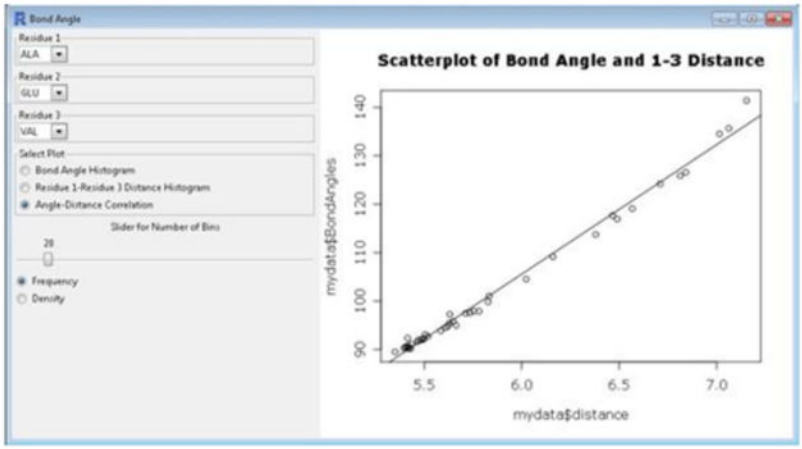

Fig. 6. Scattered plot of the virtual bond angles against their residue 1-3 distances

This snapshot shows the distribution graph for the angle-distance pairs for residues ALA, GLU, and VAL.

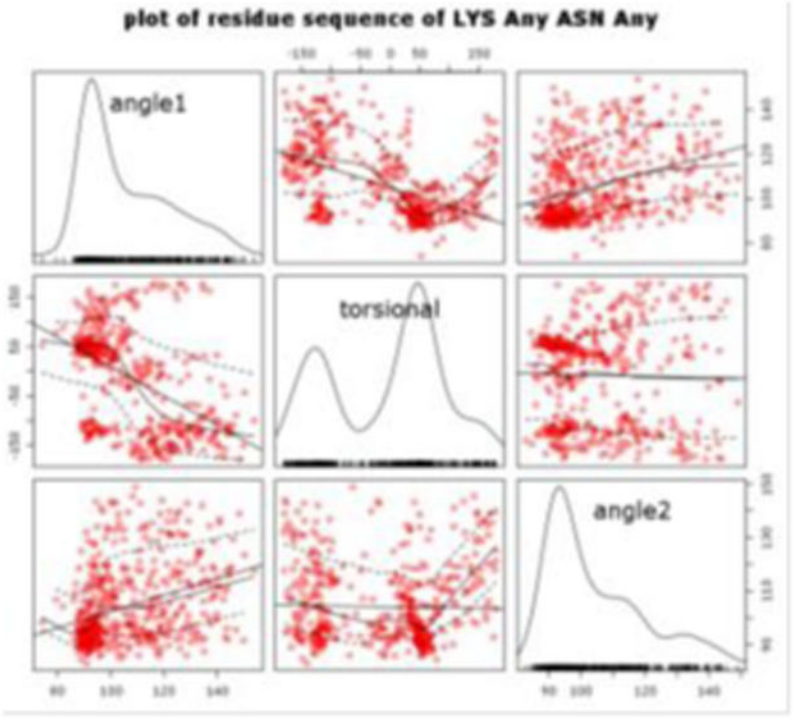

Fig. 7. A matrix of scattered plots of the distributions and correlations of the virtual torsion angle and its two neighboring virtual bond angles

The plots show the distribution and correlation graphs for the virtual torsion angle and its two neighboring virtual bond angles for residues LYS, Any, ASN, Any. The matrix of plots is 3 by 3. The graph in each is defined as follows: Square(1,1) = distribution of virtual bond angle 1; Square(2,2) = distribution of the virtual torsion angle; Square(3,3) = distribution of virtual bond angle 2; Square(1,2) = correlation between virtual bond angle 1 and the virtual torsion angle; Square(1,3) = correlation between the bond angle 1 and virtual bond angle 2; Square(2,3) = correlation between the virtual torsion angle and virtual bond angle 2.

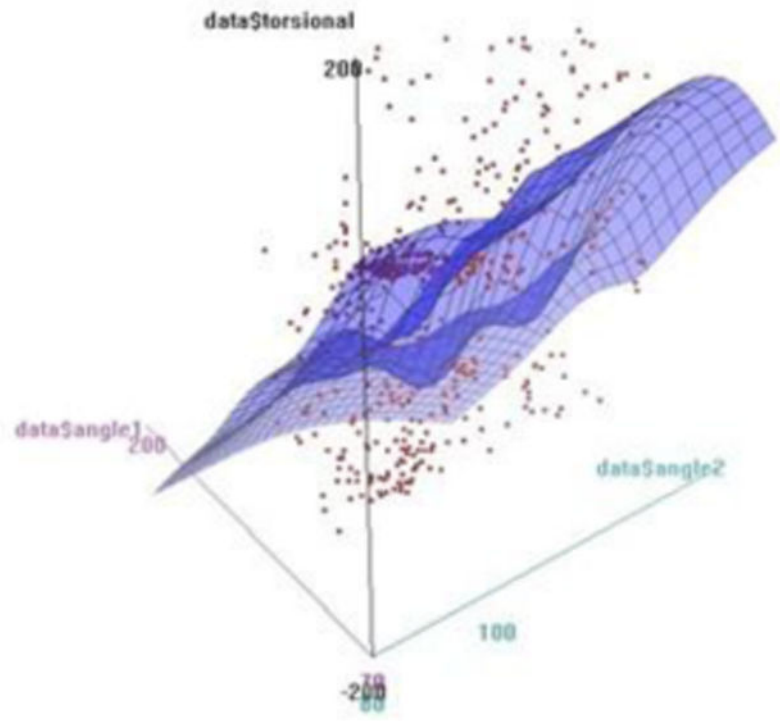

Fig. 8. 3D scattered plot for the virtual torsion angle and its two neighboring virtual bond angles

This snapshot shows the density distribution of the α-τ-β angle triplets for residue sequence LYS, Any, ASN, and Any in the α-τ-β space. A lowess approximation to the distribution is also plotted.

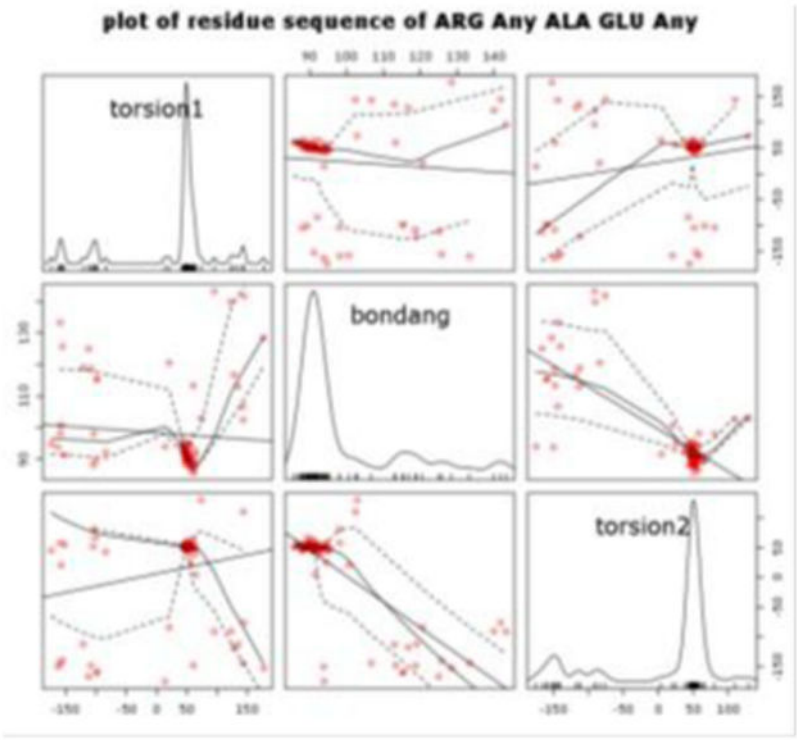

Fig. 9. A matrix of scattered plots for the distributions and correlations of virtual bond and torsion angles τ_1-β-τ_2

The plots show the distribution and correlation graphs for the virtual bond and torsion angles τ_1-β-τ_2 for residues ARG, Any, ALA, GLU, and Any. The matrix of plots is 3 by 3. The graph in each is defined as follows: Square(1,1) = distribution of virtual torsion angle τ_1; Square(2,2) = distribution of virtual bond angle β; Square(3,3) = distribution of virtual torsion angle τ_2; Square(1,2) = correlation between τ_1 and β; Square(1,3) = correlation between τ_1 and τ_2; Square(2,3) = correlation between β and τ_2.

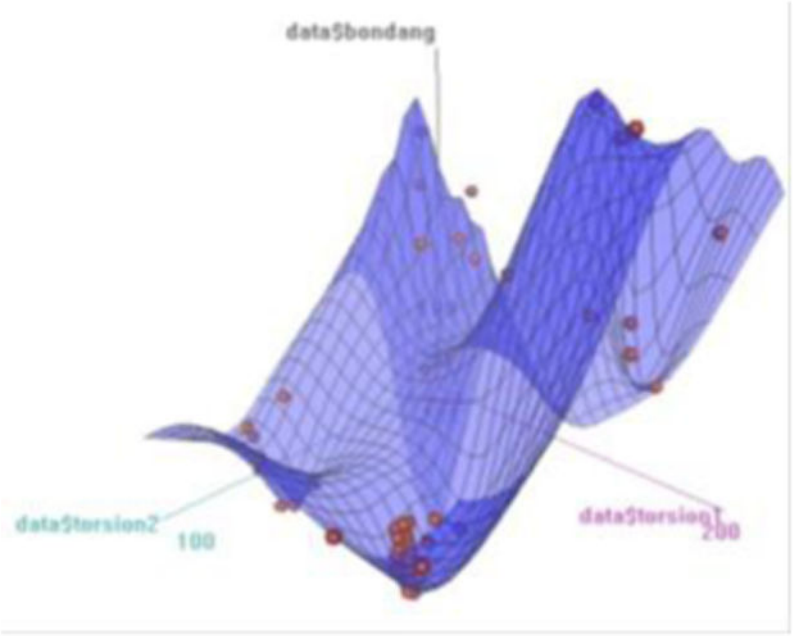

Fig. 10. scattered plot of virtual bond and torsion angles

This snapshot shows the density distribution of the τ_1-β-τ_2 angle sequence in a τ_1-β-τ_2 space for a residue sequence ARG, Any, ALA, GLU and Any. A lowess approximation to the distribution is also plotted.

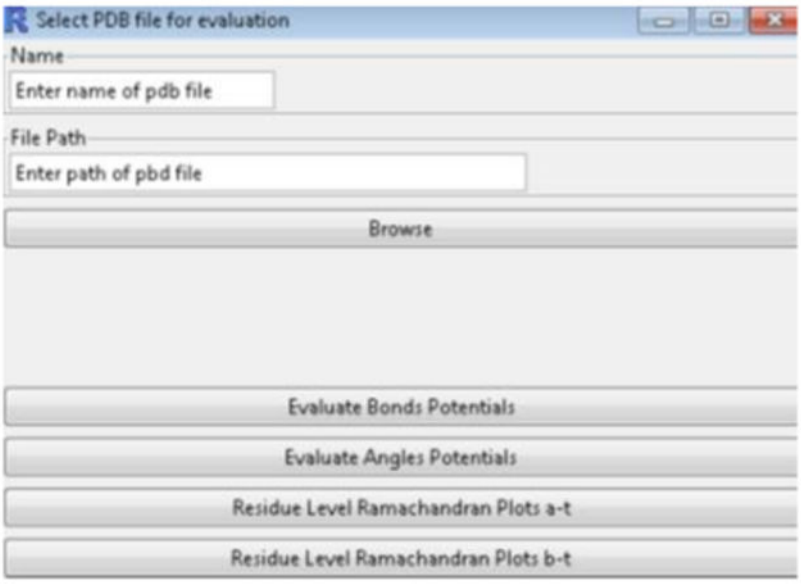

Fig. 11. Virtual bond length potentials

A window is popped out for the user to upload the structural file. The system can then evaluate the virtual bond length potential for each neighboring pair of residues in the protein sequence and display the distribution of the potential energy over the residue sequence.

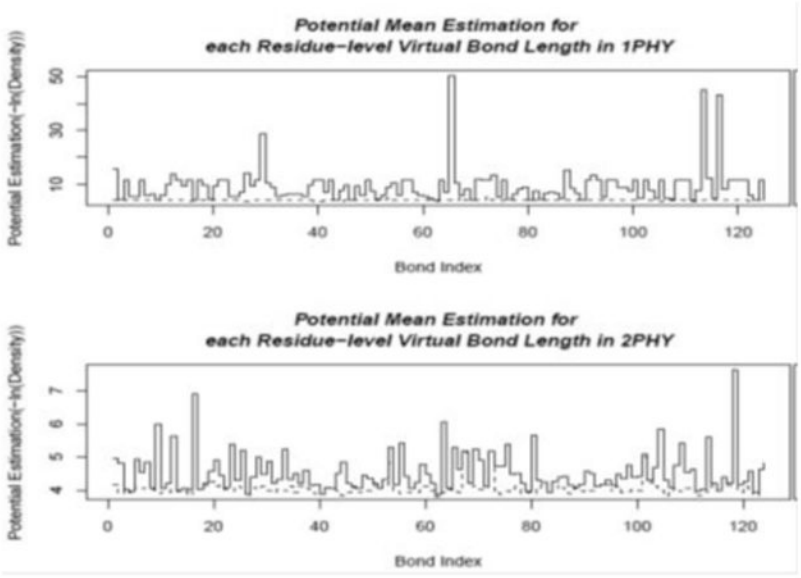

Fig. 12. Distributions of virtual bond energies for 1PHY (2.4Å) and 2PHY (1.4Å)

The energy levels of the virtual bond lengths of two structures 1PHY and 2PHY are shown in solid lines. The minimal possible energies are plotted as the dashed line. If there is no distribution data for some virtual bond, such as the bond at index 98, the potential function is not defined, and there is a gap in the energy plot for that bond. These two structures are determined with different resolutions for the same protein. The better-resolved structure (2PHY) has lower potential energies in average than the poorly determined one (1PHY).

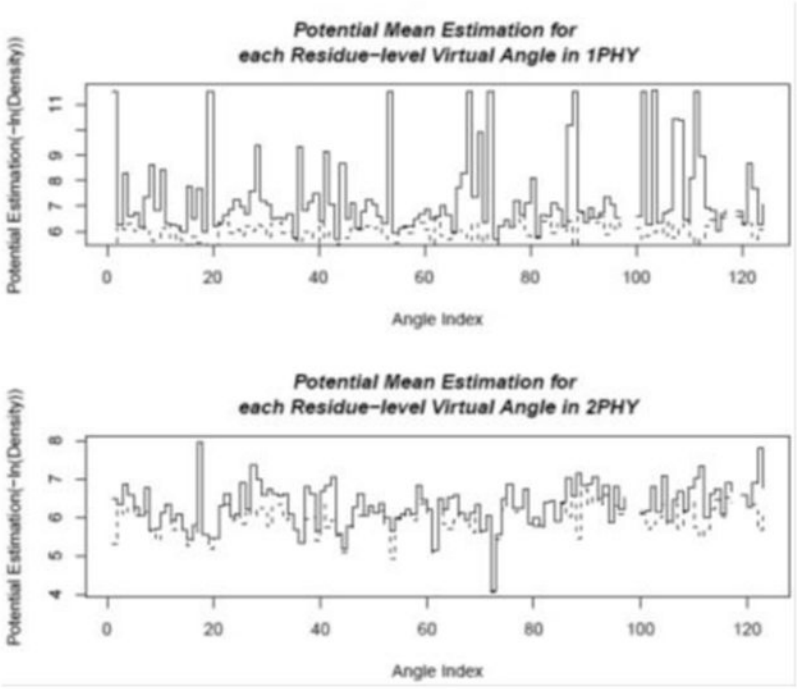

Fig. 13. Distribution of virtual bond length and bond angle energies for 1PHY (2.4Å) and 2PHY (1.4Å)

The energy levels of the virtual bond angles of two structures 1PHY and 2PHY are plotted in solid lines. The minimal possible energies are shown as the dashed line. These two structures are determined with different resolutions for the same protein. The better-resolved structure (2PHY) has lower potential energies in average than the poorly determined one (1PHY).

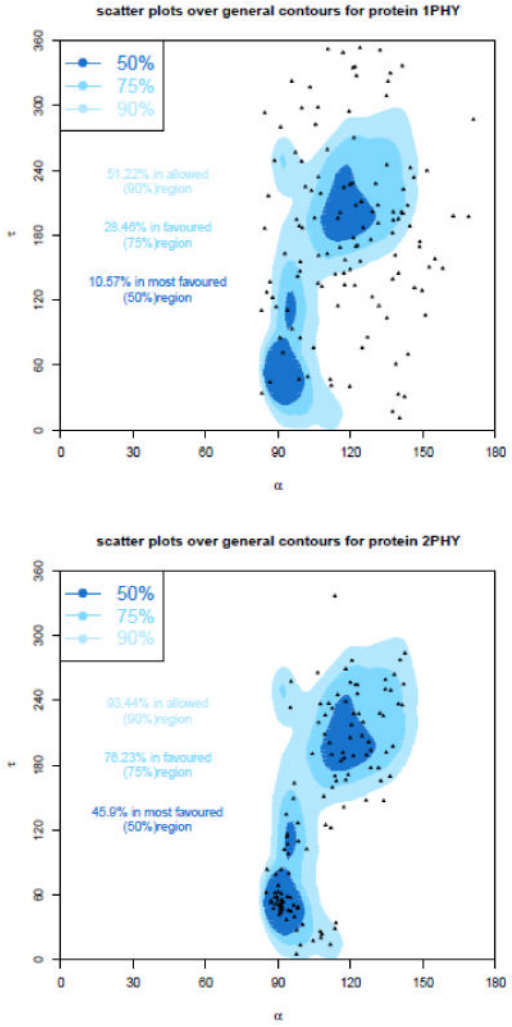

Fig. 14. The α-τ correlation plots for a protein at two different resolutions

The photoreactive yellow protein in the dark state, 1PHY (2.7Å), shown in (a) compared with the DD-peptidase, 2PHY (1.4Å), shown in (b). The background contours are generated from the general density distributions of the α-τ angle pairs in known proteins. Regions of different densities are outlined with colours in different gradients. They are defined as Most Favoured (high 50% density), Favoured (high 75% density), and Allowed (high 90% density) regions. The scattered triangles correspond to the α-τ angle pairs in the given protein structures. The lines in (a) indicate that there are 51.22% of the triangles of the α-τ angles pairs in 1PHY falling in the 90% region, 28.46% of triangles falling in the 75% region, and only 10.57% of the triangles falling in the 50% region. On the other hand, In (b), there are 93.44% of the triangles of the α-τ angles pairs in 2PHY falling in the 90% region, 76.23% of triangles falling in the 75% region, and 45.9% of the triangles falling in the 50% region.

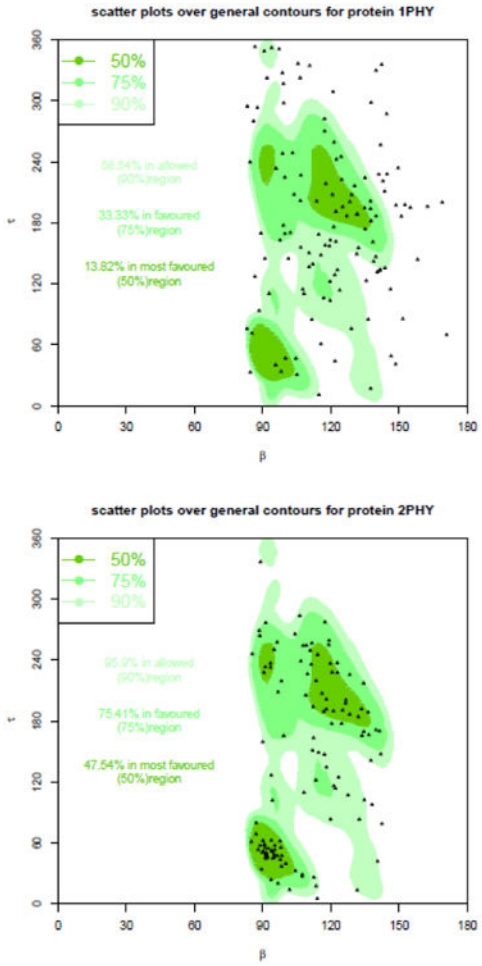

Fig. 15. The τ-β correlation plots for a protein at different resolution

The photoreactive yellow protein in the dark state, 1PHY (2.7Å), shown in (a) compared with the DD-peptidease, 2PHY (1.4Å), shown in (b). The background contours are generated from the general density distributions of the τ-β angle pairs in known proteins. Regions of different densities are outlined with colours in different gradients. They are defined as Most Favoured (high 50% density), Favoured (high 75% density), and Allowed (high 90% density) regions. The scattered triangles correspond to the τ-β angle pairs in the given protein structures. The lines in (a) indicate that there are 58.54% of the triangles of the τ-β angle pairs in 1PHY falling in the 90% region, 33.33% of the triangles falling in the 75% region, and only 13.82% of the triangles falling in the 50% region. On the other hand, in (b), there are 95.9% of the triangles of the τ-β angle pairs in 2PHY falling in the 90% region, 75.41% of the triangles falling in the 75% region, and 47.54% of the triangles falling in the 50% region.

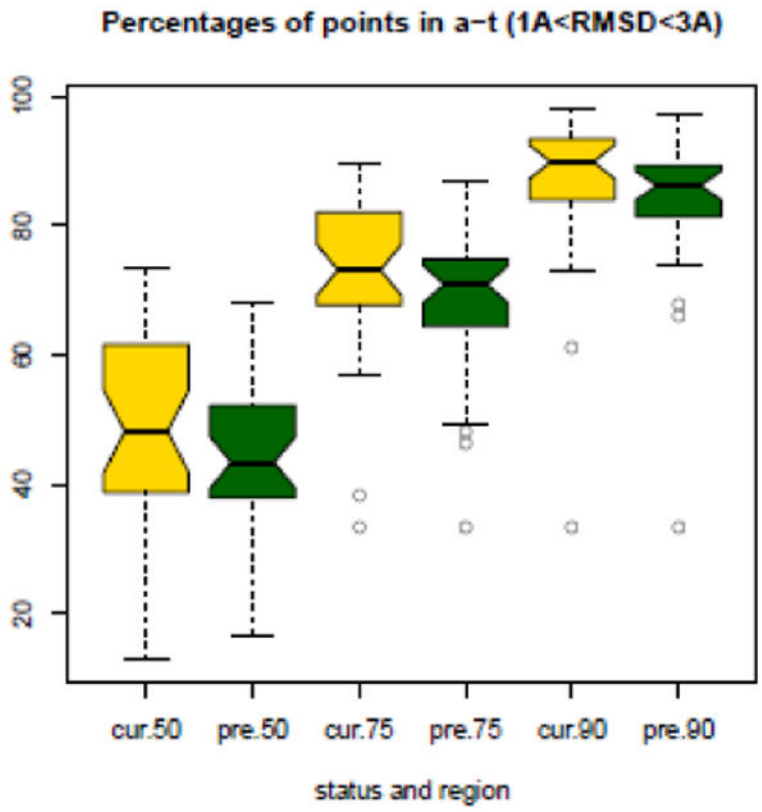

Fig. 16. The boxplots of percentages of points in α-τ correlation regions for obsoleted proteins and the successors (1A< RMSD<3A)

For the protein pairs with RMSD less than 3A but bigger than 1A, the gap between the two groups of structures can be observed from the difference in median and the difference in mean.

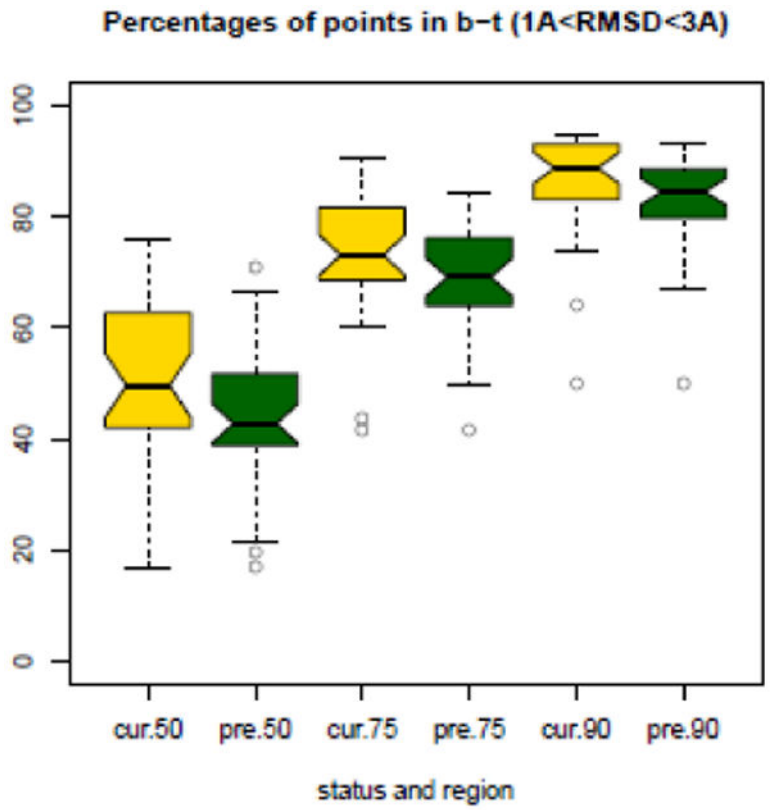

Fig. 17. The boxplots of percentages of points in β-τ correlation regions for obsoleted proteins and the successors (1A< RMSD<3A)

For the protein pairs with RMSD less than 3A but bigger than 1A, the gap between the two groups of structures can be observed from the difference in median and the difference in mean.

Table 1

Sturcture analysis for previously obsoleted structures and the current ones: compare average of the percentages for the points fall into the α-τ correlation plots

The size column gives the number of paired structures – previously obsoleted and current ones. The rest columns show the average percentages of points that fall into the allowed region (90%), the average percentage of points fall into the favoured region (75%), and the average percentage of points fall into the most favoured region (50%) for the previously obsoleted structures and the corresponding current ones in each group of paired structures with different RMSD values.

RMSD	size	pre90%	cur90%	pre75%	cur75%	pre50%	cur50%
NA	922	86.41	88.42	71.42	73.75	47.19	49.38
(0, 1A)	542	89.56	89.69	75.36	75.51	51.00	51.16
(1A, 3A)	37	83.04	86.39	68.66	72.11	44.06	48.52
(3A, 5A)	17	86.42	86.95	69.26	71.20	43.47	43.23
(5A, ∞)	136	84.09	86.31	67.69	69.84	42.62	45.03

Table 2

Sturcture analysis for previously obsoleted structures and the current ones: compare average of the percentages for the points fall into the β-τ correlation plots

The size column gives the number of paired structures – previously obsoleted and current ones. The rest columns show the average percentages of points that fall into the allowed region (90%), the average percentage of points fall into the favoured region (75%), and the average percentage of points fall into the most favoured region (50%) for the previously obsoleted structures and the corresponding current ones in each group of paired structures with different RMSD values.

RMSD	size	pre90%	cur90%	pre75%	cur75%	pre50%	cur50%
NA	922	86.44	88.27	71.83	74.01	48.03	50.22
(0,1A)	542	89.60	89.76	75.35	75.57	51.86	52.00
(1A, 3A)	37	82.95	86.14	68.65	72.68	44.42	49.53
(3A, 5A)	17	85.54	86.15	68.81	71.12	45.12	46.08
(5A, ∞)	136	83.86	84.94	67.60	69.70	43.22	45.72