PRTAD: A DATABASE FOR PROTEIN RESIDUE TORSION ANGLE DISTRIBUTIONS

By

Xiaoyong Sun

Di Wu

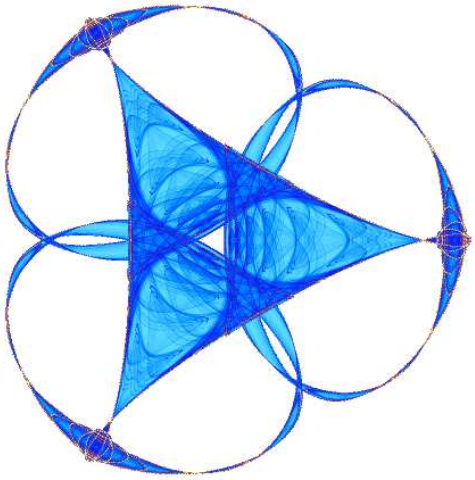
Robert Jernigan

and

Zhijun Wu

IMA Preprint Series # 2179

(November 2007)



INSTITUTE FOR MATHEMATICS AND ITS APPLICATIONS

UNIVERSITY OF MINNESOTA

400 Lind Hall

207 Church Street S.E.

Minneapolis, Minnesota 55455–0436

Phone: 612-624-6066 Fax: 612-626-7370

URL: http://www.ima.umn.edu

PRTAD: A Database for Protein Residue Torsion Angle Distributions

Xiaoyong Sun*, Di Wu†, Robert Jernigan‡, and Zhijun Wu§

*Program on Bioinformatics and Computational Biology, ‡Department of Biochemistry, Biophysics, and Molecular Biology, §Department of Mathematics
Iowa State University, Ames, Iowa 50011, U.S.A.

†Department of Mathematics
Western Kentucky University, Bowling Green, Kentucky 42101, U.S.A.

Email: sunx1@iastate.edu, zhijun@iastate.edu

Abstract

PRTAD is a dedicated database and structural bioinformatics system for protein analysis and modeling. The database is developed to host and analyze the statistical data for protein residue level "virtual" bond and torsion angles obtained from their distributions in databases of known protein structures such as in the PDB Data Bank. PRTAD is capable of generating, caching, and displaying the statistical distributions of the angles of various types. The collected information can be used to extract geometric restraints or define statistical potentials for protein structure determination. PRTAD is supported with a friendly designed web interface so that users can easily specify the angle types, and retrieve, visualize, or download the distributions of the angles as they desire. The database PRTAD is freely accessible at http://www.math.iastate.edu/prtad.

1. Introduction

Consider the residues as points located at the positions of their Cα atoms and the link between any two connected residues in sequence as a "virtual" bond. We define the angle between any two connected virtual bonds as a "virtual" bond angle and the dihedral angle around any three connected bonds as a "virtual" torsion angle (Fig. 1). A virtual bond is related to two connected residues; a virtual bond angle is related to three connected residues; and a virtual torsion angle is related to four connected residues. Different sequences of residues tend to form different residue level virtual bond lengths, bond angles, and torsion angles. The knowledge on these geometric elements of proteins is a valuable source of information for protein structural analysis and structure determination.

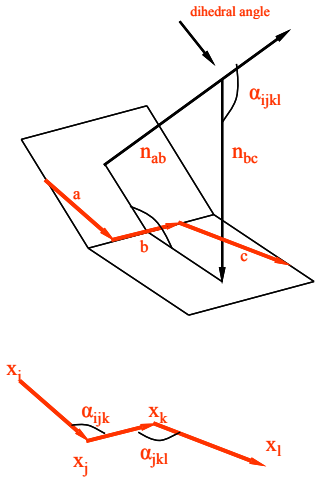


Fig 1. Virtual angles. A sequence of residues i, j, k, l are located at x_i, x_j, x_k, x_l. The vectors a, b, c are virtual bonds connecting the residues, α_{ijk} and α_{jkl} are virtual bond angles, and α_{ijkl} is a virtual torsion angle, defined as the dihedral angle between the normal vectors n_{ab} and n_{bc} of the planes formed by a, b and b, c respectively.

The protein atomic level bond lengths, bond angles, and torsion angles can usually be determined by using physical experiments such as X-ray crystallography and nuclear magnetic resonance spectroscopy (NMR), or estimated with knowledge from stereochemistry [1]. However, similar properties at the residue level can

hardly be determined experimentally. They can only be estimated theoretically or statistically. For example, the distances between heavy atoms across different residues have been extracted based on their distributions in databases of known protein structures [2-4].

In this paper, we follow the approach in [4] to obtain a statistical estimate on residue level virtual bond and torsion angles based on the distributions of the angles in databases of known protein structures. We describe the development of a database PRTAD for calculating and storing the distributions of residue level virtual angles in databases of known protein structures, and show that the collected angle data can be very useful for extracting geometric restraints or defining statistical potentials, as studied recently in [5], for protein analysis and modeling.

In order to estimate the virtual angles for different sequences of residues, we find all the information for how the angles for different sequences of residues are distributed in known proteins or more accurately, known protein structures. Then, for each angle, we assign a probability according to the distribution of the angles of the same kind. Such probability information can be very useful for evaluating the corresponding angles in given proteins and building proper protein conformations. For example, in order to see if 100º is a proper angle for a sequence of residues Alanine, Tryptophan, Cysteine, we calculate all the angles of the same kind in the known proteins in structural databases and then group the angles according to their values. We can then obtain the distribution of this type of angles in between 0 and 180º, where the probability for the angle to be 100º can be easily identified. Figure 2 shows more examples for protein residue level virtual angle distributions calculated from databases of known protein structures.

Indeed, based on our calculations on the distributions of the residue level angles in the structures in PDB Data Bank [6], we have found that 1) The majority of the angles are non-uniformly distributed, indicating that proteins do have preferences when forming these angles; 2) as more and more protein structures are determined, good estimations on the distributions of the angles are possible, and they can be obtained with reasonable statistical significances; 3) statistical constraints or potentials can be derived from the distributions of the angles and be applied to predicting or evaluating protein structures.

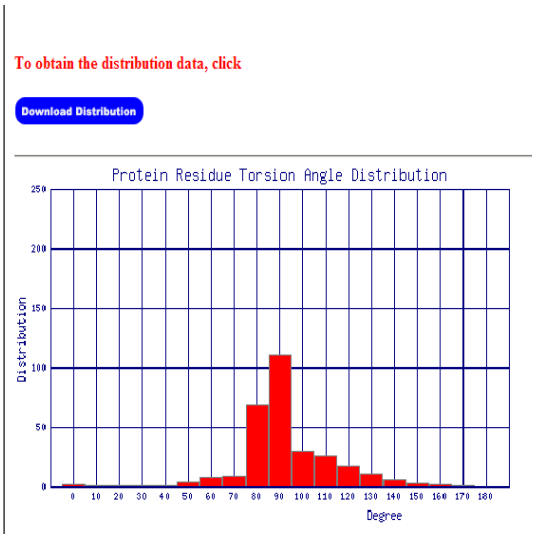


Fig 2. Example angle distribution. Shown in the graph is the distribution of the virtual bond angle formed by three connected LYS residues in sequence.

While the importance of the residue level virtual angle data is easy to justify, the calculation of the data can be daunting, requiring a complete search for the angles in structural databases for each different angle type, while there can be hundreds of thousands of different angle types, defined in terms of the types of the residues in sequence. Even just storing and managing such a large amount of data can be quite challenging. For this reason, we have developed a database system for automatically generating, storing, and analyzing all the distribution data for protein residue level virtual angles. The system consists of two coupled databases, one called the structural database for storing high-resolution structures downloaded from structural databases, and another called the angle database for storing the distribution data for the angles. The data in the angle database is calculated and collected from the structural database. The angle database can be used by the users to store, query, and analyze the distributions of any angles of interest. In any event, at the beginning, only the data for commonly used angle types are computed and stored, to avoid unnecessary space use. If the distributions for certain angles are requested, but not pre-calculated and -stored yet, they will be computed on fly from the structural database and stored into the angle database afterwards. In this way, the database can eventually be developed to contain necessary angle distributions, yet does not have to keep all the overwhelming information. The database system is developed using MySQL. Currently, it has 5475 high-resolution structures downloaded from PDB Data Bank and up to

3,293,309 angle distribution records. The system is supported with a friendly designed web interface so that users can easily specify the angle types, and retrieve, visualize, or download the distributions of the angles as they desire. The database PRTAD is accessible freely at http://www.math.iastate.edu/prtad/.

2. Systems and Methods

2.1. Data Source

When downloading the known protein structures from the PDB Data Bank, we have considered only those containing the chains of amino acids rather than protein complexes such as protein-DNA, protein-RNA, and protein-protein complexes. To obtain more accurate and reliable results, we only downloaded structures determined by X-ray crystallography with resolution higher than 2.0 Å. To reduce the redundancy in homologous structures, only proteins with sequence similarities less than 70% were used. Based on these criteria, total 5475 qualified protein structures were selected from the PDB Data Bank as of November 1, 2006.

2.2. Data Structure

PRTAD has two levels of databases, one called the structural database and another called the angle database. Both databases are implemented using MySQL. The structural database stores the sequence and structure information for a large set of high-resolution protein structures, with a similar data structure as the structural data represented in the PDB Data Bank. Each record in the structural database is similar to an atom record in the PDB file, but contains a smaller number of fields. It has the PDB name of the protein, the residue name, the index for the atom, the atom name, and the x, y, z coordinates of the atom (see Fig. 3). All the PDB files of the downloaded protein structures are converted into this format and stored in the structural database as MySQL database files. By using the MySQL database management system, the structure files can be processed much more efficiently and directly. The angle database stores the distributions of the angles in known proteins calculated for every different type of angles. The calculations were based on the distributions of the angles in the downloaded structures in the structural database. There are in fact two angle databases, one for the virtual bond angles and another for the virtual torsion angles (see Fig. 3).

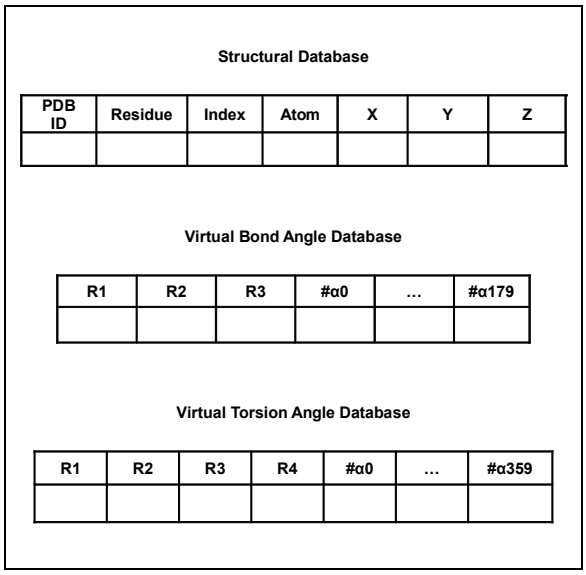


Fig 3. Data structures. The above: the record of the atom in the structural database: PDB ID – ID of protein in PDB Databank; Residue – the name of the residue containing the atom; Index – the index for the atom; Atom – the name of the atom; X, Y, Z – x, y, z coordinates of the atom. The bottom: The records for the distribution of the virtual angles, one for each different type: The first one for virtual bond angles, and the second for virtual torsion angles. R1, R2, R3, R4 – residues; #αi – the number of angles in [αi, αi+1], i = 0, …, 179 (or 359).

2.3. Definition of Angles

In order to obtain the distribution data for the angles of various types, we specify the virtual bond angles by the types of the three residues they are associated in the sequence and the virtual torsion angles by the types of the four related residues. Let d_{ij} be the virtual bond length between two residues i and j, α_{ijk} the bond angle associated with residues i, j, and k, and α_{ijkl} the torsion angle related to residues i, j, k, l. Let the coordinate vectors for residues i, j, k, l be represented by x_i, x_j, x_k, x_l (Fig. 1). Then, we can use the following formulas to calculate d_{ij}, α_{ijk}, α_{ijkl}.

$$d_{ij} = || x_j - x_i || ,$$

where $||\cdot||$ represents the Euclidean norm, and for any vector $x = (u, v, w)$, $||x|| = \text{sqrt}\,(u^2 + v^2 + w^2)$.

$$\cos\alpha_{ijk} = \frac{(x_j - x_i)\cdot(x_k - x_j)}{d_{ij}\, d_{jk}} \ .$$

where $d_{ij} = \|x_j - x_i\|$ and $d_{jk} = \|x_k - x_j\|$.

$$\cos\alpha_{ijkl} = \frac{(a\cdot b)(b\cdot c) - (a\cdot c)(b\cdot b)}{d_{ij}\, d_{jk}^2\, d_{kl} \sin\alpha_{ijk} \sin\alpha_{jkl}}$$

where $a = x_j - x_i$, $b = x_j - x_i$, $c = x_j - x_i$, $\cos\alpha_{ijk} = a\cdot b \,/\, d_{ij}d_{jk}$, and $\cos\alpha_{jkl} = b\cdot c \,/\, d_{jk}d_{kl}$.

2.4. Calculation of Distributions

Let R_i, R_j, R_k be the types of the three residues in sequence related to the virtual bond angle α formed by residues i, j, and k. Then, the probability distribution of the angle can be represented by a function $P[R_1,R_2,R_3](\alpha)$ and defined for any α in $[\alpha_i, \alpha_{i+1}]$, where $\alpha_i = i$, $i = 0, 1, \ldots, 179$, to be the number of collected angles of this particular type in $[\alpha_i, \alpha_{i+1}]$, normalized by the total number of collected angles of the same type in all $[\alpha_i, \alpha_{i+1}]$, $i = 0, 1, \ldots, 179$.

$$P[R_1, R_2, R_3](\alpha): \quad \frac{\text{Number of angles of this type in } [\alpha_i,\ \alpha_{i+1}] \ni \alpha}{\text{Number of angles of this type in } [\alpha_0,\ \alpha_{180}]}$$

Each record in the database for virtual bond angles therefore contains the distribution data for a particular type of bond angles, and it has the types of residues, R_1, R_2, R_3, that define the type of the angles followed by the number of angles of this type found in each of the angle intervals $[\alpha_i, \alpha_{i+1}]$, $i = 0, 1, \ldots, 179$.

Similarly, for the virtual torsion angles, let R_i, R_j, R_k, R_l be the types of the four residues in sequence related to the virtual torsion angle α formed by residues i, j, k, and l. Then, the probability distribution of the angle can be represented by a function $P[R_i,R_j,R_k,R_l](\alpha)$ and defined for any α in $[\alpha_i, \alpha_{i+1}]$, where $\alpha_i = i$, $i = 0, 1, \ldots, 359$, to be the number of collected angles of this particular type in $[\alpha_i, \alpha_{i+1}]$, normalized by the total number of collected angles of the same type in all $[\alpha_i, \alpha_{i+1}]$, $i = 0, 1, \ldots, 359$.

$$P[R_i, R_j, R_k, R_l](\alpha): \quad \frac{\text{Number of angles of this type in } [\alpha_i,\ \alpha_{i+1}] \ni \alpha}{\text{Number of angles of this type in } [\alpha_0,\ \alpha_{360}]}$$

Each record in the database for virtual torsion angles therefore contains the distribution data for a particular type of torsion angles, and it has the types of residues, R_i, R_j, R_k, R_l, that define the type of the angles followed by the number of angles of this type found in each of the angle intervals $[\alpha_i, \alpha_{i+1}]$, $i = 0, 1, \ldots, 359$.

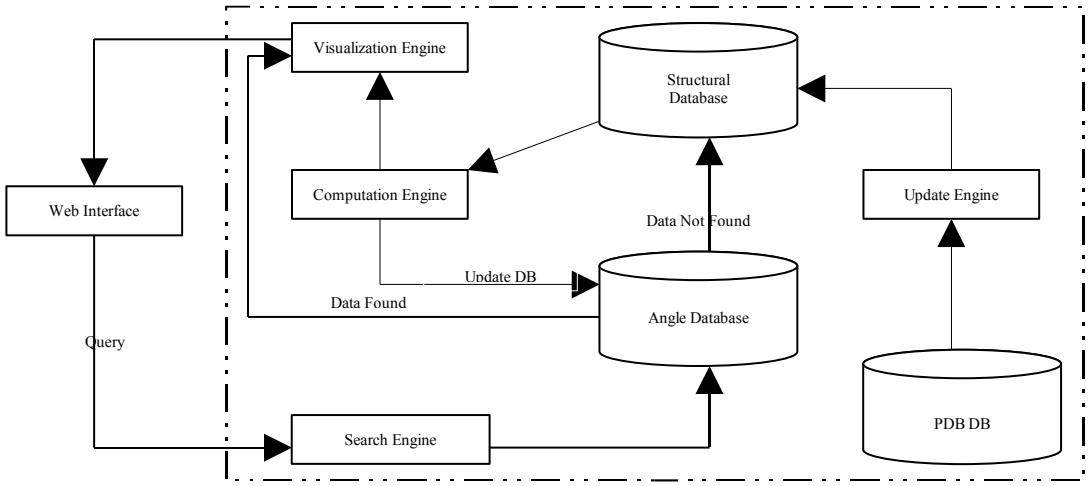


Fig 4. PRTAD organization. This automated system could generate and process the data dynamically. The system is implemented in MySQL and perl. The user could access freely the database at http://www.math.iastate.edu/prtad. It requires specifying and inputting the angle type and then the user could choose to view the graph of distribution function as well as download the related results.

2.5. System Architecture

PRTAD is implemented with MySQL. It consists of two databases, structural database and angle database, and three computational engines, search engine, distribution engine, and visualization engine (Fig. 4). In addition, there is a program written in Perl for automatically downloading the structures from PDB Data Bank and updating the structural database, and a web interface written in HTML for users to get online access to the system.

The structural database stores the sequence and structure information for a large set of high-resolution protein structures. The angle database stores the distribution data for the angles, with one record for one angle type. Since the angle type is defined in terms of types of the residues, there can be a large number of angle types: 8,000 for bond angles and 160,000 for torsion angles. For this reason, we purposely design the system to have both structural and angle databases so that the angle database can actually be built dynamically from the structural database. More specifically, at the beginning, we only compute and store the distribution data for some commonly used angle types, which can certainly be queried or processed directly in the angle database. However, if the distributions for certain angles that are not pre-calculated and -stored are requested, they will be computed on fly from the structural database and stored into the angle database afterwards. In this way, the database can eventually be developed to contain all necessary angle distributions, yet does not have to be overwhelmed by the possible combinatorial growth of data, saving both storage space and search time.

The computational engines work together as follows. The search engine takes the query from a user and searches for the distribution of the specified type of angles in the angle database. If the requested distribution has been pre-calculated and -stored in the angle database, the search engine returns with it directly. Otherwise the angles of the specified type will be computed and collected from the structural database and passed to the distribution engine. Based on the collected angles, the distribution engine calculates the distributions of the angles over discrete angle intervals, and saves them in the angle database. The visualization engine is responsible for displaying the requested distribution function through a graphics interface. Figure 4 shows the architecture of PRTAD graphically. Note that the structural database can be updated whenever new proteins are deposited into the PDB Data Bank, and the access to PRTAD can be done conveniently through a well-designed web interface.

2.6. Additional Features

A web user interface is designed so users can get access to PRTAD anywhere online. It also provides various visualization tools and functions for researchers to display and analyze requested data. The users can obtain helps from the tutorial, references, or related publications available at the website. The tutorial is well written and provides many examples.

The front page of the interface describes the PRTAD system, its design purpose, and the user guideline. Description about research on statistical constraints and potentials on the residue level virtual angles is given in the research page. The links to tutorial, references, and publications are also provided. Currently, the PRTAD front page can be reached with its internet address http://www.math.iastate.edu/prtad/.

Several pages are directed from the PRTAD front page. One of them as shown in Fig. 5 allows the users to choose the angle type to be searched for via simple menu selections. Typically, the users follow three selection steps: (i) specify the angle database to be searched: bond or torsion angle database; (ii) specify the residue sequence; (iii) submit the query. The system returns with the distribution of the specified type of angles and displays it in a graph as shown in Fig. 6.

Fig 5. PRTAD input selections. A user needs to first decide the database to be searched, bond or torsion angle database, and then specify the types of the residues (20 possibilities) in the sequence.

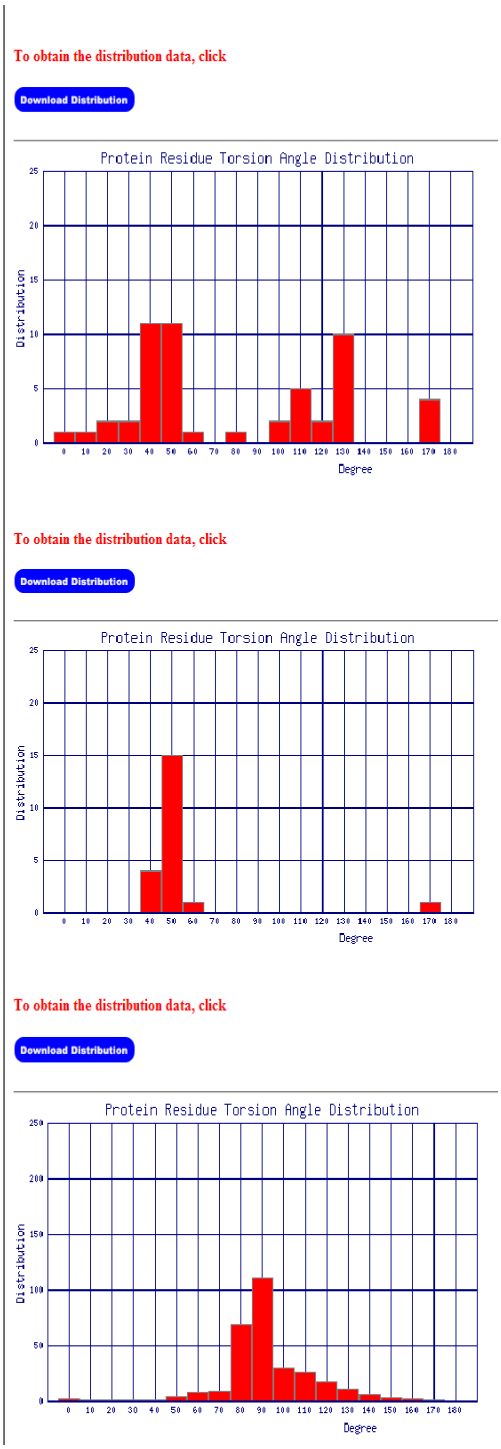


Fig 6. Graphics display. The distribution of the angles of the specified type is displayed in a graph. The angle range is up to 360º, and the size of each angle interval (bin) is 1º. (a) Torsion angle for AQHF. (b) Torsion angle for RQGL. (c) Bond angle for KKK.

3. Sample Applications

The statistical distributions of the residue level virtual angles can be an important source of information for protein structure prediction and refinement as done in [2][4] with the distributions of the inter-atomic distances. They can also be used to build residue level statistical potentials as in [5]. Here we give two examples for how the distributions of the virtual angles can also be examined through PRTAD for structural analysis.

Example 1: We have studied the prion fragment 1OEH (Fig. 7). The sequence for this fragment is HGGGWGQP. It is inside the human prion protein with residue numbers from 61 to 68, and is considered as a critical site for prion folding [7]. We have in particular examined the conformation of GGG based on the distribution of the virtual bond angle of the three residues. We have found from PRTAD that the angle is distributed in a range from 80 to 140 with mean = 90.6 and std = 13.5 (see Fig. 8 (a)). The structure for 1OEH has been determined by NMR and is available from PDB. We have therefore also calculated the angles for GGG in the 20 NMR structures for 1OEH. Figure 8 (b) shows the distribution of these angles: They are in a range from 105 to 115, and within two standard deviations from the mean value of the distribution calculated by PRTAD.

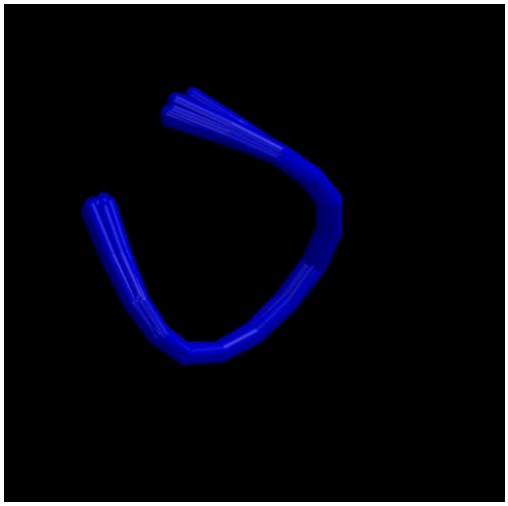


Fig 7. NMR structure of prion fragment 1OEH. The fragment extends from residue 61 to 68 in prion protein, with a sequence HGGGWGQP. The structure was determined by NMR with an ensemble of 20 structures available in PDB.

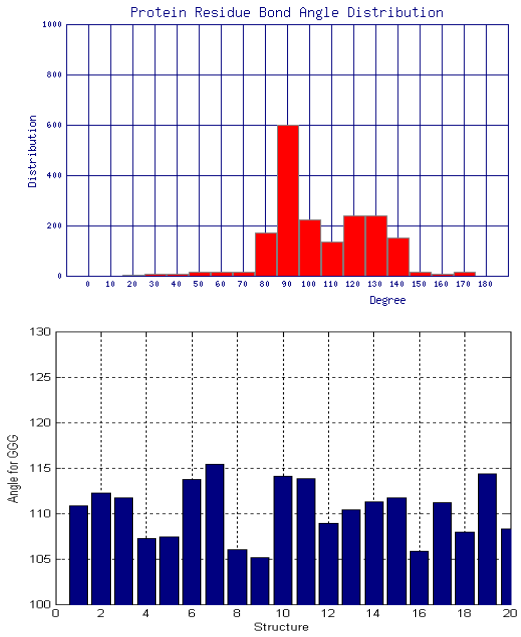


Fig 8. Angle for GGG in 1OEH. (a) The distribution of the angle calculated by PRTAD. (b) The angles in the 20 NMR structures.

Example 2: Another sequence we have examined in particular is a sequence of residues in the regulatory region of the tumor repressor p53 (Fig. 9) -- a protein important for cancer study. Recent work [8] showed that p53 is subject to lysine methylation at two sites of the regulatory region, K370 and K372, which makes the sequence KSKK in between the two sites particularly interesting. The protein p53 has been studied extensively, but the structural data for the regulatory region beyond residue 300 has not been made available. We have searched PRTAD database and investigated the possible conformations for KSKK. Interestingly, we found that the torsion angle for the sequence has a distribution with two separate high-probability regions, one from 50 to 80 and another from 110 to 150 (see Fig. 10). Lysine methylation at K370 and K372 plays an important role for the control of the DNA binding ability of p53. When methylated at K370, the DNA binding ability of p53 is reduced. However, the methylation at K370 can be inhibited by the methylation at K372. One may wonder how the structure of KSKK may affect the methylation at either K370 or K372 and the regulation between them. The two possible structural states for KSKK we found from PRTAD have certainly provided insightful information for the investigation of the structural basis for KSKK methylation in p53.

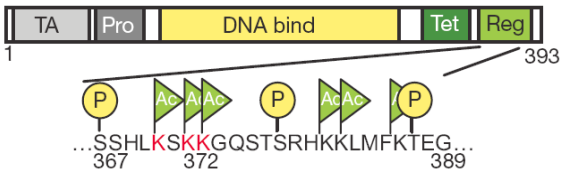


Fig 9. Structure of p53. Methylation may occur at either K370 or K372. The sequence between the two sites is KSKK (picture obtained from [8]).

4. Future Developments

The current version of PRTAD has provided the basic functions for processing the data for protein residue level virtual angle distributions. More tools will be developed to facilitate various purposes of structural analysis including the tools for computing the distributions of these angles under more structural conditions, such as the distributions of the angles of certain types when they are in alpha helices or beta sheets. The development of the database PRTAD for protein residue level virtual angle distributions is a direct extension from our previous work on the development of the protein inter-atomic distance distribution database PIDD. With the increasing number of high-resolution structures being determined, many structural properties including the inter-atomic distances, residue level virtual angles, residue volumes, side-chain orientations can all be analyzed from their statistical distributions in known proteins [9]. Therefore, in future, we will further extend our work to the development of a general protein geometry database that includes the statistical distribution data for many other protein geometric properties besides the distances and angles. Such a system will be able to provide more complete information on protein conformations and have an even greater potential as a bioinformatics tool for protein structural analysis and structural modeling [10].

5. Acknowledgement

The work is partially supported by the Department of Mathematics, the Program on Bioinformatics and Computational Biology, and the Baker Center for Bioinformatics and Biological Statistics at Iowa State University. The support from the Ogden College of Science and Engineering of Western Kentucky University, the Institute of Mathematics and Its Applications at the University of Minnesota, and the

NIH/NIGMS grant R01GM081680 are also acknowledged.

6. References

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

[2] Cui, F., Jernigan, R., Wu, Zj., Refinement of NMR-determined protein structures with database derived distance constraints, *J Bioinformatics and Computational Biology* **3**, 2005, 1315-1329.

[3] Wu, D., Cui, F., Jernigan, R., Wu, Z., PIDD: A database for protein inter-atomic distance distributions, *Nucleic Acids Res.* (published online, DOI: 10.1093/nar/gkl802), 2006.

[4] Wu, D., Jernigan, R., Wu, Z., Refinement of NMR-determined protein structures with database derived mean force potentials, *Proteins: Struct. Funct. Bioinf.*, (published online, DOI: 10.1002/prot.21358) 2007.

[5] Feng, Y., Kloczkowski, A., Jernigan, R., Four-body contact potentials derived from two protein databases to discriminate native structures from decoys, *Proteins: Struct. Funct. Bioinf.* (published online, DOI: 10.1002/prot.21362), 2007.

[6] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., The Protein Data Bank, *Nucleic Acids Research* **28**, 2000, 235-242.

[7] Zahn, R. The Octapeptide repeats in mammalian prion protein constitute a ph-dependent folding and aggregation site, *J. Mol. Biol.* **334**, 2003, 477-488.

[8] Huang J., et al. Repression of p53 activity by Smyd2-mediate methylation, *Nature* **444**, 2006, 629-632.

[9] Gerstein, M., Richards, F. M., Protein geometry: volumes, areas, and distances, in *International Tables for Crystallography*, Rossmann, M., Arnold, E., editors, Kluwer, 2001, 531-539.

[10] Bourne, P. E. and Weissig, H., *Structural Bioinformatics*, John Wiley & Sons, Inc., 2003.

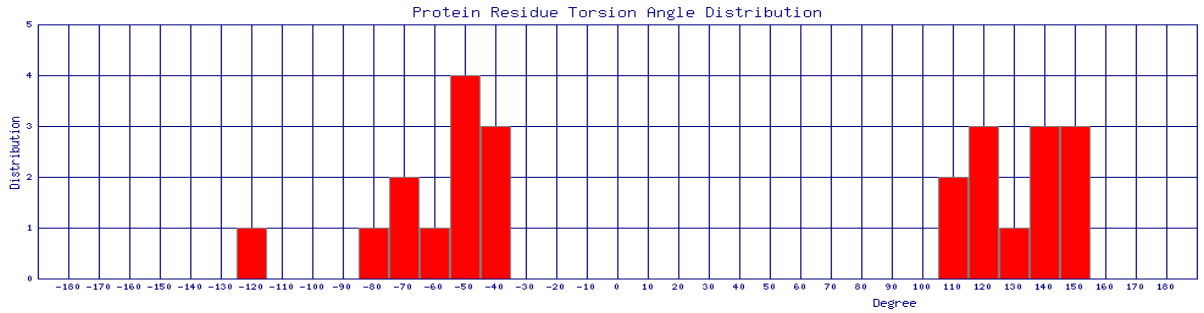


Fig 10. Angle for KSKK in p53. The distribution of the angle calculated has two separate peaks corresponding to two possible conformations of the fragment.