PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2016, 43(8): 810~816
www.pibb.ac.cn

DRSP: a Structural Database for Single Residue Substitutions in PDB*

LIU Ji-Long[1)**], MIAO Zhi-Chao[2)**], LI Lei[1)], XIAO Zhi-Xiong[1)***], CAO Yang[1)***]

([1)] *Center for Growth, Metabolism and Aging, Key Laboratory of Biological Resources and Ecological Environment of Ministry of Education, College of Life Sciences, Sichuan University, Chengdu* 610064, China;

[2)] *Architecture and Reactivity of the RNA, University of Strasbourg, Institute of Molecular and Celluar Biology of CNRS* 67000 Strasbourg France)

Abstract Substitutions of residues introduced by gene mutations can impact protein structures and often alternate protein properties. Identifying substitution-related structural changes is crucial for understanding their effects in biological functions. With the explosive numbers of experimentally determined protein structures, more and more wild-type-mutant structures have been used in comparable studies of structural biology. In this work, we constructed a structure-pair database, named DRSP, which collected non-redundant pairs of proteins with single residue substitutions in the entire Protein Data Bank. DRSP demonstrates abundant information for mutant-structure prediction and protein design. More importantly, DRSP presents the superimposed pairs of wild-type-mutant structures for backbone-perturbation studies, which are very important for the highly accurate structure modeling. DRSP is available at http://www.labshare.cn/drsp/index.php.

Key words database, protein residue substitution, backbone flexibility, protein structure prediction, protein design
DOI: 10.16476/j.pibb.2016.0056

Mutations occur frequently in genomic DNA. A change of base in the coding sequence can lead to a miss-sense mutation, resulting in a substitution of an amino acid. A single substitution of an amino acid can lead to disruption of protein structure and severe alteration on protein folding, stability, interaction and biological functions [1-2]. A well-known case is the sickle-cell anemia, which is caused by a point mutation in the gene encoding the β-globin chain of hemoglobin. This mutation leads to the replacement of a hydrophilic glutamic acid by a hydrophobic valine at the sixth position [3]. The absence of the polar amino acid promotes the aggregation of hemoglobin, which distorts red blood cells into a sickle shape and decreases their elasticity. Another example is mutation H1047R of PI3K protein, a key enzyme involved in cancer development. The wild type histidine forms hydrogen bond with the main-chain carbonyl of Leu956, while the side chain of the mutated arginine points 90° away from the position of its wild-type counterpart histidine and greatly increases the distance to Leu956. Together with changes of point direction and p*K*a, it alters the interaction with cell membrane, and enzymatic activity, which eventually promotes tumorigenesis[4]. Thus, amino acids substitution-related structural change is a key factor that impacts protein functions.

Protein structure modeling is a very useful tool for exploring the mutation-related structure changes. Previously, mutant structure prediction is usually based on rotamer substitution at the mutation site without considering the backbone movement [5-6]. Some later researchers found that backbone perturbations such as backrub motion play important roles in positioning mutant residues[7-10]. Thus, the accumulation of diverse mutational data is very important for structure modeling. For example, a set of 717 pairs of protein

*This work was supported by grants from The National Natural Science Foundation of China (31401130) and National Laboratory of Biomacromolecules (2014kf 04).
**These authors contributed equally in this work.
***Corresponding author. Tel: 86-28-85418843
XIAO Zhi-Xiong. E-mail: bmc605@hotmail.com
CAO Yang. E-mail: cao@scu.edu.cn
Received: February 23, 2016 Accepted: June 27, 2016

structures with single point mutations were used for developing MODELLER-based methods for the modeling of point mutations[11]. Another set of 2 141 pairs of structures by Abagyan has been used as a benchmark by RosettaBackrub[8]. Such available data set of protein mutation-structures has contributed greatly to the development of new methods for structure prediction and protein design[5, 12–15].

In recent years, with the explosive numbers of experimentally determined protein structures, current database of structure pairs of residue substitutions is limited and out of date. It is therefore urgent to establish a more complete and high-quality database for the research of structure prediction and protein design. Here we present a structure-pair based single residue substitution database, named as DRSP. To our knowledge, DRSP is currently the largest publicly available protein structure database of single residue substitutions. More importantly, DRSP offers the superimposed pairs of wild-type-mutant structures for backbone perturbation studies, which are crucial for the highly accurate structure modeling. To achieve this purpose, we first developed a computer program combining the Needleman-Wunsch algorithm[16] and Jacobi eigenvalue method for quick alignment of protein sequences and structures. Then we applied it to identify single residue substitutions in the entire Protein Data Bank (PDB)[17–18], which were carefully examined and filtered with resolution and sequence redundancy. Finally, we calculated abundant information such as secondary structures, relative solvent accessibility, number of hydrogen bonds, backbone compatibility and side-chain packing compatibility, to demonstrate the structural and physic-chemical features of the substitution region. We expect that this database can be useful for the mutant structure analyses and protein design studies.

1 Materials and methods

1.1 Measurement of structural change

The root-mean-square deviation (RMSD) is widely used to measure structural similarity. In this work, we defined two other RMSD values, A_RMSD and M_RMSD, to quantify structural changes during substitution. Before the calculation, proteins in a pair are superimposed according to the flanking regions, which are the 10 adjacent residues of the mutation site other than the 2 nearest ones at both sides (8 for each protein) in terms of primary sequence. Then A_RMSD is defined as the RMSD of the backbone atoms of 16 (8×2) flanking residues. And M_RMSD is defined to be RMSD of backbone atoms of the mutation site. A_RMSD is used to evaluate the quality of the protein superimposition, while M_RMSD measures the backbone perturbation of the mutation site. As we focus on local structure perturbation, mutations with the relatively large backbone perturbation are defined to fulfill M_RMSD $\geqslant$1.0Å and A_RMSD < 1.0Å.

1.2 Detection of point mutations

To rapidly and sensitively identify pairs of proteins with residue substitutions, we developed a sequence- structure combined alignment program, ProMut. It extracts protein sequences from the PDB files and aligned the sequences using Needleman-Wunsch dynamic programming algorithm together with a simple equivalent scoring matrix:

$$Score_{ij}=\begin{cases}1, & \text{if } a_i = a_j\\ 0, & \text{if } a_i \neq a_j\end{cases}$$

Where a_i, a_j are the amino acid types of i and j sites on a protein. If point mutation were detected, flanking residues were found accordingly and structures were superimposed according to the flanking residue with the Jacobi eigenvalue algorithm. Compared with other widely used sequence alignment programs, such as BLAST[19], it is a simple but sensitive method to detect point mutations and to obtain the superimposed 3D structures. ProMut is available to non-commercial users at our website.

1.3 Construction of the database

To construct the database of structure pairs of single point mutation, we downloaded all the protein X-ray structures with resolution < 3.0Å from Protein Data Bank (PDB) of September, 2015[17]. The protein structures were split into single chains, and renamed by the PDB code and the chain ID. For example, 1a2cA indicates the chain A of PDB 1a2c. In addition, protein subunits with backbone atom defects were removed, since the absence of backbone atoms will result in uncertain local structures. In addition, a number of proteins recur in the PDB database, such as the same protein determined in different laboratory or the same protein subunits exist in several complexes. Protein subunits with the same sequence and RMSD < 1.0Å were ranked by their resolutions. The one with the best resolution was selected as a representative. Thus, we collected a non-redundant single chain structure database of 34 274 protein subunits(Figure 1).

Then we employed ProMut, which identified

11 130 pairs of structures that differ by a single residue type. To reduce the influence by thermal oscillations or disorder, residue substitutions occur in terminal residues (<10) were removed. Besides, the A_RMSD value generated by ProMut indicates the quality of the protein superimposition. To filter the bad superimposed data, structure pairs with A_RMSD > 1.0Å were also removed, resulting in 7 419 structure pairs finally (Figure 1). Those structure pairs with A_RMSD > 1.0Å were collected in a subset for analyzing mutation related topology changes later.

The web server of the database was built with PHP and Apache. All the data was organized by MySQL database.

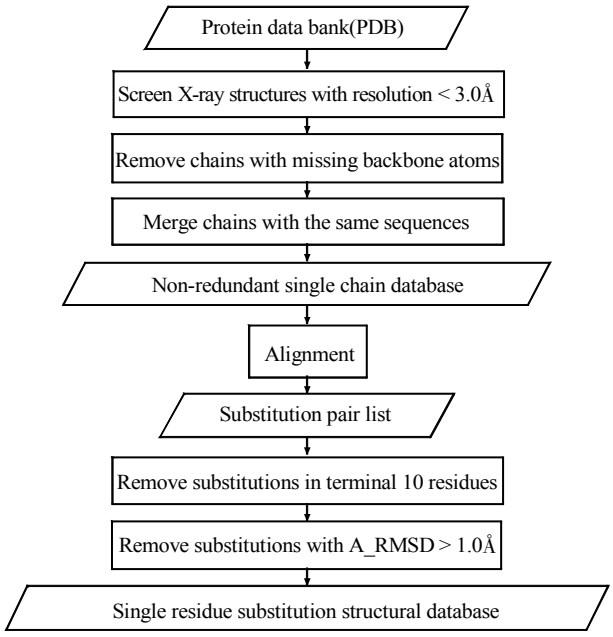


Fig. 1 The flowchart of constructing DRSP

All the protein X-ray structures were first obtained from PDB, and then were filtered with the criterion of resolution < 3.0Å. Chains with missing backbone atoms were removed, and chains with the same sequences were merged to get a non-redundant single chain database. Next, those non-redundant single chains were aligned using ProMut to identify the pairs of proteins with single residue substitution. Finally, all the protein pairs were filtered with substitution positions and A_RMSD values.

1.4 Structural features

This database offers residue substitution related features to users. They include the M_RMSD, relative solvent accessibility, secondary structure, number of hydrogen bonds, backbone compatibility and side-chain packing compatibility at the mutation site (Figure 2b). The relative solvent accessibility was calculated by NACCESS [20]. Secondary structure was calculated by DSSP [21–22]. Backbone hydrogen bonds were calculated by the criteria of distance of hydrogen-acceptor < 3.0Å and angle of donor-hydrogen-acceptor> 120° [23]. Backbone compatibilities were calculated by the difference of logarithms of the neighbor-dependent Ramachandran probabilities for protein pairs [24]. Side-chain compatibility was defined by side-chain packing score of CIS-RR[25–26]. The detailed information for backbone compatibility and side-chain packing compatibility are described in section **2.3**.

2 Results

2.1 Overview

The online version of the DRSP is available at http://www.labshare.cn/drsp/index.php (Figure 2a). In the current release this resource includes 7 419 pairs of structures, which can be browsed by:

(1)Types of amino acid in residue substitutions.

(2)Solvent accessibility of substituted residues.

(3)Number of backbone-related hydrogen bonds of substituted residues.

(4)Secondary structures of substituted residues.

(5)Backbone compatibility of substituted residues.

(6)Side-chain packing compatibility of substituted residues.

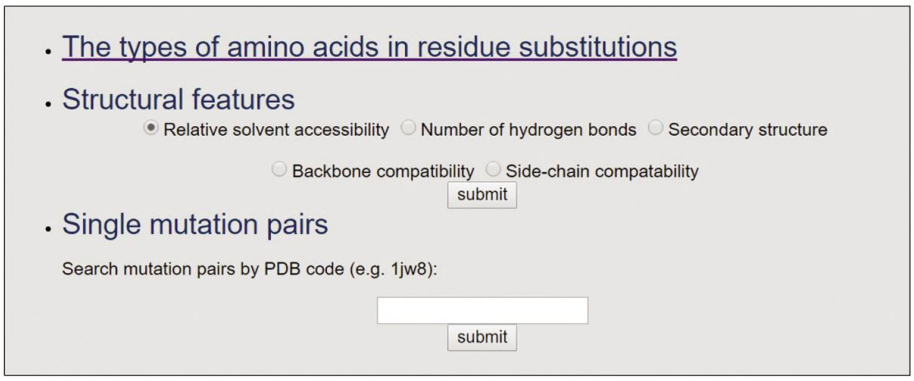
(7)PDB code.

The data hit the items are listed with abundant information such as secondary structure, relative solvent accessibility, number of hydrogen bond, backbone compatibility and side-chain packing compatibility (Figure 2b). All the data can be downloaded from the website.

2.2 The types of amino acids in residue substitutions

We classified all the data into $(20 \times 19)/2 = 190$ groups according to the types of amino acids of the wild type and the corresponding mutant. On average, one group includes 39 pairs of substitution, with the minimum number of 2 and the maximum number of 301. The percentages of residue substitutions that cause the protein backbone perturbation with M_RMSD $\geqslant$1.0Å are shown in Figure 3. Overall, most of the residue substitutions have a minor impact on protein backbone conformation. Nevertheless, some of them perturb backbone conformation considerably. It is known that proline (P) owns special main chain structure and substitution for P may result in backbone conformational changes. Our statistic result shows the

(a)



(b) The feature you chosen is backbone hydrogen bond:

Wildtype	Mutant	Length	Site	M_RMSD	Substitution	Secondary Structure	Solvent accessibility	*Hydrogen bond*	Backbone compatibility	Side-chain compatability
1izqA	3n2jE	128	117	6.447	HIS->GLY	helix	10%	*two*	1.769	3.181
2dekB	2owfA	265	143	4.019	TRP->MET	sheet	70%	*two*	0.331	-0.558
1r3yA	2q71A	356	168	3.650	GLY->ARG	coil	10%	*two*	7.143	2.736
3qzoA	3qznC	121	166	3.287	TYR->ALA	sheet	30%	*two*	1.586	1.185
4k1jB	4k1kA	388	147	2.479	GLY->ASP	coil	30%	*two*	10.659	-1.890
4k1kA	4k1jB	388	147	2.479	ASP->GLY	helix	30%	*two*	2.302	-0.037
2j1mA	2ij3B	454	264	2.269	ALA->HIS	helix	30%	*two*	0.778	-1.994
2ij3B	2j1mA	454	264	2.269	HIS->ALA	helix	10%	*two*	0.130	0.940
2v5vB	1dx9C	168	57	2.212	GLU->ALA	sheet	30%	*two*	0.247	0.307
2q7bA	3fa7A	137	59	2.147	ARG->GLU	sheet	50%	*two*	0.810	0.228
2fbrB	2b14B	114	55	2.060	LEU->PRO	sheet	10%	*two*	9.804	5.019
3f8aA	2q7bA	137	59	2.023	TRP->ARG	coil	70%	*two*	-1.162	-2.070

Fig. 2 The interface of DRSP database website

(a) Users can survey the database *via*: (1) Types of amino acids. (2) Solvent accessibility of substituted residues. (3) Number of backbone-related hydrogen bonds of substituted residues. (4) Secondary structures of substituted residues. (5) Ramachandran possibilities of substituted residues. (6) Side-chain packing compatibility of substituted residues. (7) PDB code. (b) Table of the query results. Results include the PDB codes, length of protein, sequence number of mutation site, M_RMSD, types of amino acids, secondary structure, solvent accessibility, the number of hydrogen bonds of backbone, backbone compatibility, and side-chain compatibility of the corresponding mutation.

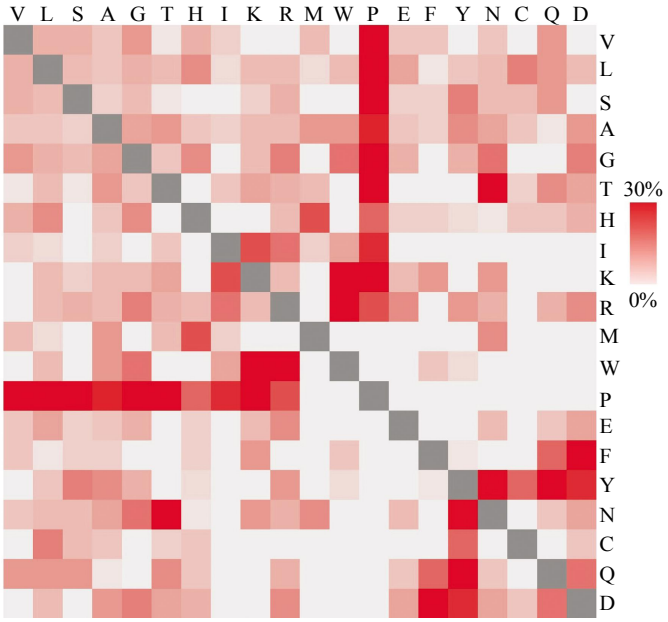


Fig. 3 Protein backbone perturbation heat map

The heat map illustrates percentages of residue substitutions that cause the protein backbone perturbation with M_RMSD≥1.0Å. The darker the color, the larger the percentages. The symmetrical grids show(20×19)/2=190 groups of substitutions, according to the types of amino acids of the wild type and the corresponding mutant. Color gray indicates the null data.

residue substitutions associated with P are apt to perturb the backbone conformation when it substitutes with V, L, S, A, G, T, H, R and K. However, it has a minor impact in substitution with M, W, F, Y, N, C, Q, E, and D. Especially, for groups of P substituted with E, F, D and N, which owns more than 10 pairs in each group, none of them causes the protein backbone perturbation with M_RMSD ⩾1.0Å. This result implies that aromatic or negative charged amino acids are not prone to perturbed by proline substitution. Besides P, the result also shows that, substitutions of tyrosine (Y) and tryptophan (W) have comparatively high possibilities(>20%) in perturbing protein backbone, especially for the substitution of Y with N, Q, and W with K, R. This result suggests that aromatic amino acids induced backbone perturbation should be carefully considered in structural studies. More detailed information could be found on our web site.

2.3 The backbone and side-chain compatibility upon residue substitution

As the twenty types of amino acids have almost the same backbones but different side-chains, a direct assumption is that backbone perturbation is the result of the compatibility of side-chain packing rather than backbone. To verify this assumption, we employed a well-established side-chain packing software CIS-RR [25–28]. Given the backbone of a protein, it can predict the most suitable side chains by enumerating the discrete side-chain rotamers and calculating van der Waals interaction energy plus the rotamer probability. Its side-chain packing score is the indicator of fitness of a particular amino acid packing into the given protein environment. Thus we defined side-chain compatibility (*SP*) as the average changes of side-chain packing score of both mutations from wild type protein and from mutant protein:

$$SP=\frac{|SP_{wt}-SP_{vmut}|+|SP_{vwt}-SP_{mut}|}{2}$$

Side-chain packing compatibility (*SP*) represents the preference of a particular side chain of amino acid fitting the original conformation. The higher the values of *SP* indicates the larger the difference of side chain packing. The footnote wt and mut indicate experimental structures of the mutation pair. And the vwt and vmut indicate the predicted mutant structures using wt and mut structures with software CIS-RR in mutation mode.

Ramachandran plots are widely used to measure the variation of backbone conformations from the objective of finding stable ones[24]. Here we defined the backbone compatibility as the average changes of Ramachandran probabilities for both mutations from wild type protein and from mutant protein:

$$RP=\frac{|RP_{wt}-RP_{vmut}|+|RP_{vwt}-RP_{mut}|}{2}$$

Backbone compatibility (*RP*) indicates the preference of a particular type of amino acid with the backbone of dihedral angles ψ and φ. The above footnotes are the same as that in side-chain compatibility. In this way, we calculated side-chain and backbone compatibility values all over the mutation pairs. Then we classified the data into mutations with small (M_RMSD<1.0Å) and large (M_RMSD ⩾1.0Å) backbone perturbations. It is surprising that the average values of side-chain compatibilities between the two groups show only a few differences, which are 1.527 *vs* 1.724, while the average values of backbone compatibilities are statistically significantly different, which are 1.000 *vs* 1.508 (Figure 4). Detailed analysis shows that glycine, alanine, serine and threonine have the largest differences (～2 times) of *RP* between the two groups.

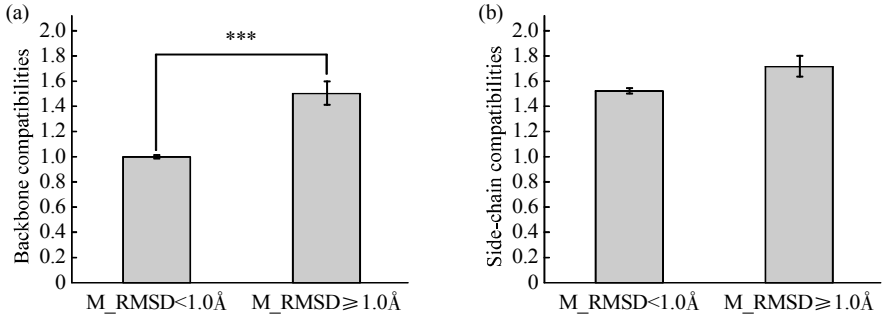


Fig. 4 The backbone (a) and side-chain (b) compatibilities in two groups of the data, which are classified by M_RMSD

Backbone compatibilities in the group of M_RMSD<1.0Å show significantly difference from those in the group of M_RMSD⩾1.0Å. *** $P < 0.005$.

These results suggest that the backbone dihedral angle preference could be an indicator for protein backbone perturbation. As for side-chain compatibility, our data illustrate that side-chain packing effects have minor impacts on backbone perturbation, which does not agree with the former assumption. It can attribute to the current definition of side-chain packing compatibility which is not closely correlated with backbone displacement. Nevertheless, it implies that the van der Waals interaction and side-chain rotamer preference is not the sensitive indicator of backbone perturbation on average. We have also classified the data using different cutoffs of M_RMSD and obtained the similar conclusions. Taken together, we suggest that Ramachandran probability should be carefully considered in structure modeling which allows backbone perturbation.

3 Conclusions

In order to aid the studies of mutant structure prediction and protein design, we constructed the structural database of single residue substitutions by checking the entire PDB database. The mutation-structure pairs were obtained by our program ProMut using both sequence and 3D structure alignments. To improve the quality, we filtered the data by structure resolution and various checking procedures. Moreover, we calculated the backbone RMSD, relative solvent accessibility, secondary structure, number of backbone hydrogen bonds, backbone compatibility and side-chain compatibility at the mutation site to illustrate abundant information related to point mutation. Finally, we propose a web server of DRSP freely available to academic users: http://www.labshare.cn/drsp/index.php.

Acknowledgements We would like to thank Drs. JIANG Tai-Jiao, HANG Hai-Ying and ZHANG Yu-Jun for the stimulating discussion, and Dr. CHEN Shuang for critical review of the manuscript.

References

[1] Gong S, Blundell T L. Structural and functional restraints on the occurrence of single amino acid variations in human proteins. PLoS One, 2010, **5**(2): e9186

[2] Wang Z, Moult J. SNPs, protein structure, and disease. Human Mutation, 2001, **17**(4): 263−270

[3] Schmaier A H, Lazarus H M. Concise guide to hematology. New York: John Wiley & Sons, 2011

[4] Janku F, Wheler J J, Naing A, *et al*. PIK3CA mutation H1047R is associated with response to PI3K/AKT/mTOR signaling pathway inhibitors in early-phase clinical trials. Cancer Research, 2013, **73**(1): 276−284

[5] Bordner A J, Abagyan R A. Large-scale prediction of protein geometry and stability changes for arbitrary single point mutations. Proteins: Structure, Function, and Bioinformatics, 2004, **57** (2): 400−413

[6] De Filippis V, Sander C, Vriend G. Predicting local structural changes that result from point mutations. Protein Engineering, 1994, **7**(10): 1203−1208

[7] Keedy D A, Georgiev I, Triplett E B, *et al*. The role of local backrub motions in evolved and designed mutations. PLoS Comput Biol, 2012, **8**(8): e1002629

[8] Smith C A, Kortemme T. Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. Journal of Molecular Biology, 2008, **380**(4): 742−756

[9] Davis I W, Arendall W B, Richardson D C, *et al*. The backrub motion: how protein backbone shrugs when a sidechain dances. Structure, 2006, **14**(2): 265−274

[10] Mandell D J, Kortemme T. Backbone flexibility in computational protein design. Current Opinion in Biotechnology, 2009, **20** (4): 420−428

[11] Feyfant E, Sali A, Fiser A. Modeling mutations in protein structures. Protein Science, 2007, **16**(9): 2030−2041

[12] Schaefer C, Rost B. Predict impact of single amino acid change upon protein structure. BMC Genomics, 2012, **13**(Suppl 4): S4

[13] Masso M, Vaisman I I. Accurate prediction of stability changes in protein mutants by combining machine learning with structure based computational mutagenesis. Bioinformatics, 2008, **24** (18): 2002−2009

[14] Wainreb G, Wolf L, Ashkenazy H, *et al*. Protein stability: a single recorded mutation aids in predicting the effects of other mutations in the same amino acid site. Bioinformatics, 2011, **27** (23): 3286−3292

[15] Eyal E, Najmanovich R, Sobolev V, *et al*. MutaProt: a web interface for structural analysis of point mutations. Bioinformatics, 2001, **17**(4): 381−382

[16] Needleman S B, Wunsch C D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology, 1970, **48**(3): 443−453

[17] Rose P W, Bi C, Bluhm W F, *et al*. The RCSB Protein Data Bank: new resources for research and education. Nucleic Acids Research, 2013, **41**(D1): D475−D482

[18] Berman H M, Kleywegt G J, Nakamura H, *et al*. The Protein Data Bank archive as an open data resource. Journal of Computer-aided Molecular Design, 2014, **28**(10): 1009−1014

[19] Altschul S F, Gish W, Miller W, *et al*. Basic local alignment search tool. Journal of Molecular Biology, 1990, **215**(3): 403−410

[20] Hubbard S J, Thornton J M. NACCESS. Computer Program, Department of Biochemistry and Molecular Biology, University

College London, 1993, **2**(1)
[21] Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen〉bonded and geometrical features. Biopolymers, 1983, **22**(12): 2577-2637
[22] Joosten R P, Te Beek T A H, Krieger E, *et al*. A series of PDB related databases for everyday needs. Nucleic Acids Research, 2011, **39**(suppl 1): D411-D419
[23] McDonald I K, Thornton J M. Satisfying hydrogen bonding potential in proteins. Journal of Molecular Biology, 1994, **238**(5): 777-793
[24] Ting D, Wang G, Shapovalov M, *et al*. Neighbor-dependent Ramachandran probability distributions of amino acids developed from a hierarchical Dirichlet process model. PLoS Comput Biol, 2010, **6**(4): e1000763
[25] Cao Y, Song L, Miao Z, *et al*. Improved side-chain modeling by coupling clash-detection guided iterative search with rotamer relaxation. Bioinformatics, 2011, **27**(6): 785-790
[26] Miao Z, Cao Y, Jiang T. Modeling of Protein Side-Chain Conformations with RASP. Methods in Molecular Biology (Clifton, NJ), 2014, **1137**: 43-53
[27] Miao Z, Cao Y, Jiang T. RASP: rapid modeling of protein side chain conformations. Bioinformatics, 2011, **27**(22): 3117-3122
[28] Dunbrack R L, Cohen F E. Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Science, 1997, **6** (8): 1661-1681

DRSP：蛋白质单残基替换结构数据库 *

刘继龙[1)**] 苗智超[2)**] 李 雷[1)] 肖智雄[1)***] 曹 洋[1)***]

([1)] 四川大学生命科学学院，生物资源与生态环境教育部重点实验室，生长代谢衰老研究中心, 成都, 610064；
[2)] *Architecture and Reactivity of the RNA, University of Strasbourg, Institute of Molecular and Celluar Biology of CNRS* 67000 Strasbourg France)

摘要 蛋白质残基替换是基因突变的产物之一，它可能改变蛋白质三维结构，对其生物学功能产生重大影响，因此研究蛋白质残基替换与结构改变的关系具有重要意义．随着实验解析蛋白质结构的数量迅猛增长，越来越多的野生型 - 突变体被应用于结构生物学的比较研究中．本研究从蛋白质三维结构数据库(PDB)出发，收集和计算了大量结构特征数据，构建了一个目前已知最大的野生型 - 突变体(单残基差异)的结构对数据库 DRSP，展示出氨基酸类型和主链偏好性对结构保守性的相关性．DRSP 的开放使用可为高精度的蛋白质结构分析预测提供有用信息，它的数据库网址是 http://www.labshare.cn/drsp/index.php．

关键词 数据库，蛋白质残基替换，主链柔性，蛋白质结构预测，蛋白质设计
学科分类号 Q811 **DOI**: 10.16476/j.pibb.2016.0056

* 国家自然科学基金(31401130)和生物大分子国家重点实验室开放研究课题(2014kf04).
** 共同第一作者.
*** 通讯联系人. Tel: 028-85418843
肖智雄. E-mail: bmc605@hotmail.com
曹 洋. E-mail: cao@scu.edu.cn
收稿日期：2016-02-23，接受日期：2016-06-27