Structural bioinformatics

Measure the Relative Position of Residue or Atom in Protein Structure using Statistical Depth Function

Wei Zheng[1,2], Qiqige Wuyun[1], and XXX[1,*]

[1] School of xxx

[2] State Key Laboratory xxx

ABSTRACT

Motivation: Protein structure is the key to understand the protein function and evolution. Different folds have different kinds of shapes that implement all kinds of protein functions. Especially some particular shapes have special functions such as protein-protein or protein-ligand interaction interface. How to describe these structures is the fundamental basis of protein structure analysis. In addition, residues in different positions of protein might have different functions. For instance, most of the enzyme catalytic residues are on the surface of proteins; the protein ligand binding residues are often in the pockets of the protein surface. Description of these characters of residues will help us understand the protein function and predict the functional residues of proteins.

Results: We propose here a new protein structure descriptor, residue statistical depth (RSD), using statistical depth function. This new descriptor can measure the relative position of residues or atoms in protein. Unlike residue depth and DPX, RSD is not based on RSA, which means the RSDs of residues on protein surface are different and the residues in the surface cavity or pocket can be distinguished easily by their RSDs, which is important because most of the functional residues expose to the solvent. We compare DPX, residue depth and RSD here and the results show that all DPX, residue depth and RSD are relative with physicochemical properties of amino acid, such as hydrophobicity, flexibility, polarity and accessibility as well as the secondary structure of protein. Furthermore, the ligand protein binding residues have significantly greater RSDs than other residues on protein surface. The same results are found on residue conservation and phosphorylation site. The exposed residues with greater RSDs are more conserved and the phosphorylation sites have greater RSDs than other residues. In addition, based on RSD, we propose a protein shape index to classify proteins.

1 INTRODUCTION

Protein structure is the fundamental basis of understanding of protein functions and cell mechanism. Different protein structures lead to the different protein functions. For instance, small ligands tend to bind to proteins in the pockets of protein surface (Campbell, 2003); the proteins interacting with each other always have complement structures (Baker and Sali, 2001); and most of the enzyme catalytic residues are on the surface of proteins (Fersht, 1999). Description of these special structure or residues would help us understand how these functions work and predict the functions of proteins.

The solvent-accessible surface area (SASA) (Lee and Richards, 1971) or relative solvent accessibility (RSA) has been widely used in the analysis and description of protein structure, prediction of the functional residues and protein-protein interactions (Branden and Tooze, 1991; Jones and Thornton, 1996; Anderson 2002). RSA could be used to identify the protein surface residues but it provides little information on buried atoms and residues. Based on the RSA, Chakravarty and Varadarajan proposed the residue depth, which is the average of the entire atom depths of residue. (Chakravarty and Varadarajan, 1999), and DPX uses Connolly surface instead of RSA to measure the depth of the residues or atoms in protein structure (Pintar*et al*., 2003a). Residue depth and DPX can measure how depth the residues or atoms from the protein surface and are proven related with H/D exchange rates, phosphorylation site and secondary structure of proteins (Chakravarty and Varadarajan, 1999, Pintar*et al*., 2003a). But residue depth or DPX consider the residues or atoms with the same distance to surface equally and cannot measure the relative position of residues or atoms which are on the surface, for instance, pockets or cavities on the protein surface. Shen *et al.* (Shen *et al.*, 2007) use modified half space depth (Tukey, 1975) to measure the relative position and find that the half space depth of residues are relative with their physicochemical properties, such as hydrophobicity, polarity and charge. Half space depth can measure the relative position of both buried and exposed residues, but the computation of half space depth is time consuming. So it is difficult to apply this index to atoms or large proteins.

Measuring the relative position for one dimensional data or ordering them is relatively simply and have been widely used in statistics, such as median and order statistics. But extension median or order statistics to higher dimension is difficult because there is no clear order principle for multi-dimension data. Recent years, statistical depth functions have become increasingly pursued as a useful tool in nonparametric inference for multivariate data, particularly for ordering the multi-dimensional observations (Zuo and Serfling, 2000; Serfling 2002). Many kinds of definition have been raised, such as halfspace depth (Tukey, 1975; Shen *et al*, 2007), simplicial depth (Liu, 1988), regression depth (Rousseeuw and Hubert, 1999), L_1 depth (Vardi and Zhang, 1999) and project depth (Zuo and Serfling, 2000a, 2000b). Zuo and Serfling introduced the general notions of statistical depth function in 2000 (Zuo and Serfling, 2000). Statistics depth functions have been used in both statistics and bioinformatics. Yeh and Singh studied bootstrap confidence regions based on halfspace depth (Yeh and Singh, 1997). Liuproposed depth tools for multivariate analysis (Liu 1990). Wang *et al.* selected the negative functional sites of protein for protein ligand interaction prediction by depth functions (Wang

[*] To whom correspondence should be addressed.

et al., 2013).Shen *et al.* modified the definition of halfspace depth and applied it to protein structure (Shen *et al.*, 2007).

In this paper, we propose a new kind of descriptors to measure the relative position of residues or atoms in protein structure using statistical depth function. We analyze the relationship between physicochemical propensities and the statistical depth of amino acid. We show that our new descriptors are strongly relative with the physicochemical propensities of amino acid as well as the secondary structure of residues. Our descriptors can be used to distinguish the protein shapes and classify the protein structures. Unlike RSA or other RSA based depth, our new descriptors could measure the relative position on protein surface. For example, we show that the application of the new raised method to prediction of the phosphorylation site or ligand-binding site also shows improvements when compared with other depth. At last, we calculate the mean conservation for the whole CULLPDB's surface, show that a hole or a convex on the surface has a high conservation and the new definition L_1 depth, in particular, was seen as another significant feature.

2 DATASETS AND METHODS

Datasets: Three datasets are used in this paper.

(1) CULLPDB (http://dunbrack.fccc.edu/Home.php) contains 2082 protein chains. The sequence similarity is removed by the cutoff 25%, resolution cutoff is 1.6 Å, and the R-factor cutoff is 0.25. The dataset was generated on March 5, 2013by PISCES (Wang and Dunbrack, 2003).

(2) Phospho3D (80 single protein chains, PDB ids and phosphorylation site information are listed in S4), which is downloaded from Phospho3D database (1071 single protein chains) (Andreas*et al.*, 2007; 2011) and clustered by BLASTCLUST (-L=0.4) (Altschul*et al.*, 1997). This dataset is used for studyingthe depth of phosphorylation site.

(3) PLBD (33 Positive Ligand-Binding Datasets, 18053 proteins)(Hu*et al.*, 2012) set is used for studying L1 depth of protein-ligand binding sites. The dataset includes 33 biological relevant ligands which are not similar to each other and totally 18053 protein-ligand complexes from PDB.

RSA: Amino acid solvent accessibility (ASA)is a degree to which a residue in a protein is accessible to a solvent molecule. Relative solvent accessibility (RSA) of an amino acid residue is defined as the ratio of the solvent-accessible surface area of the residue observed in 3D structure to that observed in an extended tripeptide (Gly-X-Gly or Ala-X-Ala) conformation (Minh and Jagath, 2005).In this paper, we calculated residue RSA by NACCESS (Hubbard and Thornton, 1993). A residue is considered exposed if its RSA is greater than 0.

Conservation: The residue conservation is calculated based on Shannon en tropy (Wang and Samudrala, 2006). We define the conservation of a given residue as follows:

$$S = -\sum_{k,p_k \neq 0}^{20} p_k \log_2 p_k \qquad (1)$$

where$p_k$represents theobserved frequency of residue type k in the aligned column. A aligned column(v_1, v_2,… v_{20}) for the residue is got by PSI-BLAST (Altschul*et al.*, 1997), the aligned column can be found from the 21st-40th column by the PSSM file, and p_k=v_k/100. For a uniformamino acid frequency distribution, S can achieve its maximal value $\log_2 20$. A residue evolutionary conservation can be defined as follows:

$$conservation = 1 - S / \log_2 20 \qquad (2)$$

Secondary Structure: Secondary structure of a protein was calculated by DSSP (Joosten*et al.*, 2010; Kabsch and Sander, 1983) and classified as helix (H+G), sheet (B+E) and coil (I+T+S+other).

Residue Depth: Chakravarty and Varadarajan (Chakravarty and Varadarajan, 1999) define the depth of an atom in a protein as the distance of the atom from the nearest surface water molecule. The depth of a residue is the average of the constituent atom depths. We calculated residue depth by depth-1.0 (Kuan *et al.*, 2011; 2013).

DPX: The DPX (Pintar*et al.*, 2003a; 2003b) value of a residue is the average atom depth of all its atoms, where the atom depth was calculated as the distance between a given atom and the nearest positive Connolly surface atom. In this paper, DPX means residue DPX. We calculate DPX by software DPX (Pintar*et al.*, 2003c).

HalfSpace Depth: Halfspace depth is a statistical depth function. Tukey (Tukey, 1975) introduced the halfspace depth to order the high dimensional data. Shen *et al.* (Shen *et al.*, 2007) modified the definition of halfspace depth and applied it to analyze protein structure. For a point x in R^d with a probability measure *P* on R^d, halfspace depth (HD) is defined as the minimum probability mass carried by any closed halfspace containing x. That is:

$$HD(x;P)=\inf\{P(H):H \text{ is a closed halfspace, } x \in H, x \text{ in } R^d\} \qquad (3)$$

For a protein structure, we use the position of CA as the position of the residue, then halfspace residue statistical depth (HSRSD) can be defined simply as:

$$\mathrm{HSRSD}(\mathrm{Res_i}) = \inf\left\{\mathrm{Num}_{hs}(\mathrm{Res_i}) / \mathrm{N}(\mathrm{Res_{total}})\right\} \qquad (4)$$

Where $\mathrm{Num}_{hs}(\mathrm{Res_i})$ is the residue number of the closed halfspace which is divided by the level through $\mathrm{Res_i}$. And N ($\mathrm{Res_{total}}$) is total residue number of the protein. The time complexity of halfspace depth for 3D data is O(N^3) and it is can be reduced to O(N^2logN) by algorithm design (Rousseeuwand Struyf, 1998). The geometrical center of the residue also can be considered as the position of the residue in (4). Because the computation of halfspace depth is time consuming, we do not calculate the halfspace depth for atoms here.

L_1 Depth Global/Local(R): L_1 depth(Vardi and Zhang, 1999) is also a typical statistical depth function. The L_1 depth (L_1D) of a point x with respect to a data set S={$X_{1,}$ $X_{2\ldots}X_n$} in R^d is one minus average of the unit vectors from x to all observations in S.

$$L_1D(x;S) = 1 - \|\vec{e}(x)\|, where\ e_i = \frac{x - X_i}{\|x - X_i\|}, \vec{e}(x) = \frac{\sum_{i=1}^{n}\eta_i e_i}{\sum_{i=1}^{n}\eta_i} \qquad (5)$$

Where $\eta_i$is a weight assigned to observation X_i (and is 1 if all observations are unique), and $\|x-X_i\|$ is the Euclidean distance between x and X_i. The time complexity of L1 depth is O(N), which is much faster than halfspace depth. Therefore we calculate the L1 depth for each atom of protein structure instead of each residue.

For a protein structure, the Global L_1atom Statistical Depth (L_1ASD_g) is defined as: if we let data set S be all atom in protein, it is called global L_1 depth; otherwise if let S be the atoms whose distances from x are less than R(Å), then it is local L_1 depth (R).

Global L_1atom statistical depth (L_1ASD_g) is defined as follows:

$$L_1ASD_g(P_i) = 1 - \frac{1}{n}\left|\sum_{j \neq i}\frac{\overrightarrow{P_jP_i}}{|P_jP_i|}\right| \qquad (6)$$

Where P_i and P_j are the atoms of protein structure. And local L_1atom statistical depth (R) (L_1ASD_{lR}) is calculated as follows:

$$L_1ASD_{lR}(P_i) = 1 - \frac{1}{n}\left|\sum_{j \neq i}^{|p_jp_i|<R}\frac{\overrightarrow{P_jP_i}}{|P_jP_i|}\right| \qquad (7)$$

Global L_1 residue statistical depth (L_1RSD_g) is the average value of the$L_1ASD_g$of the atoms in the residue and local L_1 residue statistical depth

(R) (L_1RSD_{lR}) is the average value of total atoms' L_1ASD_{lR}, where R is represent local radius. For example L_1RSD_{l17} means distance cutoff of the local L_1statistical depth is 17 Å.

We call all kinds of depth calculated using statistical depth function as Residue or Atom Statistical Depth (RSD or ASD).

From the definition of the RSD, the further a residue is from protein "center", the smaller L1RSDgof this residue could be. Thus, if a residue'sL1RSDg is small enough (<0.3), we can consider this residue is in a convex area on protein surface. On the contrary, if the L1RSDgof this residue is big, this residue will locate in the center of the protein.

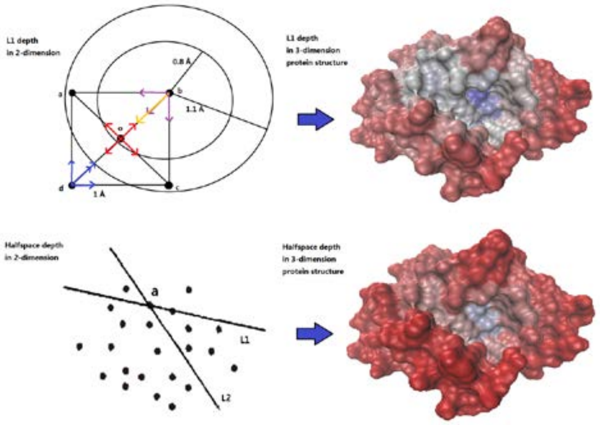


Fig.1. The left top one is L1 depth in 2-dimension, o point's global L_1 depth is 1-| $\vec{0}$ |/5=1 (vectors are shown in red color), d point's global L_1 depth is 1-|1+1+ $\sqrt{2}$ |/5=0.32 (vectors are shown in blue color), b point's $L_1RSD_{l0.8}$ is 1-|1|/2=0.5 (vectors are shown in yellow color) and its' $L_1RSD_{l1.1}$ is 1-|1+ $\sqrt{2}$ |/4=0.4 (vectors are shown in purple color). The left bottom one is Halfspace depth in 2-dimension, a point's HD(a:X)=2 (L1 line) and its' HSRSD is 2/23=0.087. The right top is the global L_1 depth of protein 1A3C chain A, and the right bottom one is half space depth.

Data Scale: To compare the consistency between different kinds of depth, we scaled DPX, residue depth and HSRSD by the following formula in order that different kinds of depth all range from 0 to 1.

$$f(\mathrm{Re}\, s_i) = \frac{depth(\mathrm{Re}\, s_i) - depth_{\min}}{depth_{\max} - depth_{\min}} \quad (8)$$

ECDF: Let $\{\text{depth}\}_n$ be depth variables, its empirical cumulative distribution function (ECDF) was defined as:

$$Fn(t) = count(depth_i \le t) / n = (\sum_{1}^{n} 1\{depth_i \le t\}) / n \quad (9)$$

where $depth_i$ is one kind of depth value and $1\{\text{depth}_i \le t\}$equals1 if $\text{depth}_i \le t$ otherwise equals 0. We present here three numeric characters of ECDF as follows:

(a) W_{ecdf}: the difference between the maximal $L_1RSD_g$and the minimal L_1RSD_g in protein. This value represents the width of L_1 depth's ECDF. It is defined as follows:

$$W_{ecdf} = depth_{\max} - depth_{\min} \quad (10)$$

(b) K_{std}: the standard deviation of slopes defined by every adjacent two points in ECDF.It is calculated as follows:

$$K_{std} = SD\{k_i \mid k_i = (Fn(X_{i+1}) - Fn(X_i)) / (X_{i+1} - X_i),$$
$$X_i \in Fn_{ecdf}^{-1}\} \quad (11)$$
$$SD\{k_i\}_n = \frac{1}{n-1}\sum_{i=1}^{n}(k_i - \overline{k})^2$$

(c) $D_{\max}$: the maximal distance along y axis betweenthepoints ofECDF and the line across the first and last points of ECDF. It is calculated as follows:

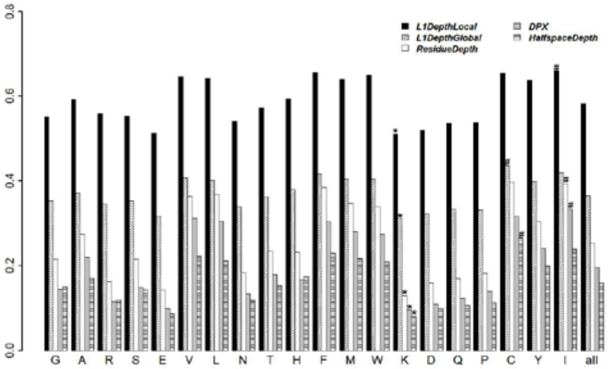
$$D_{\max} = \max\{d_i \mid d_i = | Fn(X_i) - line(X_i) |, line(X) = k*(X - X_{\min})$$
$$+ Fn(X_{\min}), k = Fn(X_{\max}) - Fn(X_{\min}) / X_{\max} - X_{\min}\} \quad (12)$$

These three indexes will be used in Results section to classify protein shapes.

3 RESULTS

3.1 Statistics information in CULLPDB

We calculated the means of five kinds of depth and four types of HSE for each of the 20 amino acid types in CULLPDB. The means of DPX range from 0.46 to 1.65Å and residues depth's means range between 4.39 and 6.84Å. The results are shown in **Fig.2 (a)**, Lys(K) has the minimal L_1RSD_g, HSRSD, DPX and residue depth. Cys(C) has the maximal L_1RSD_g and HSRSD. And Ile(I) has the maximal DPX and residue depth. (b) Ile(I) has the maximal $HSEAU_{13}$ and $HSEBD_{13}$, Thr(T) has the maximal $HSEBU_{13}$ and $HSEAD_{13}$, while Glu(E) has the minimal $HSEAU_{13}$ and $HSEBD_{13,}$ and Pro(P) has the minimal $HSEAD_{13,}$ Leu(L) has the minimal $HSEBU_{13.}$ On the other hand, Lys has highest hydrophobicity index (Fasman, 1989) and Cys and Ile have lowest and second lowest hydrophobicity indices. In fact, all these means of depth values are consistent with the hydrophobicity of amino acids. We calculate the correlation coefficients between the means of depth values of amino acids and hydrophobicity index of amino acids (**Table.1**). The results show that the means of depth values is strongly relative with hydrophobicity of amino acids.



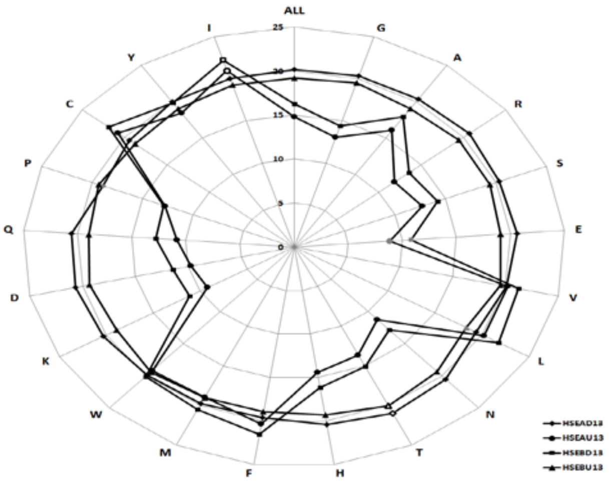


Fig.2 (a) Means of five kinds of depth in CULLPDB.X-axis is 20 amino acids and the average of all amino acids. Y axis is depth values. All these depth values and HSE values are scaled to 0-1 by (8) in each protein. Star mark means the minimal depth and sharp mark means maximal depth of 20 kinds of amino acid. (b) Means of four kinds of HSE in CULLPDB. It is strongly indicate that HSEAU has high similarity to HSEBD and HSEAD has high similarity to HSEBU.

The Pearson correlation coefficient between mean L_1RSD_g and mean HSRSD of amino acids is 0.990, and the correlation coefficient is 0.991 between mean DPX and mean residue depth of amino acid. We also calculate the correlation coefficient between L_1RSD_g with HSRSD of the residues for every protein in CULLPDB, and the average correlation coefficient in all proteins is 0.88, and the average correlation coefficient between DPX and residue depth is 0.77. The correlation coefficients between L_1RSD_g and HSRSD of "sphere" proteins are bigger than other irregular shape proteins. For instance, the correlation coefficient of influenza neuraminidase (PDB id: 1F8E) is almost 1 (0.97). Thus, we can use L_1ASD instead of HSASD when the shapes of proteins are nearly sphere, which would be much easier to calculate. And we found there is high similarity between $HSEAU_{13,}$ $HSEBD_{13}$ with depth function. For example, the Pearson correlation coefficient between mean $HSEAU_{13}$ and mean DPX, residue depth, HSRSD, $L_1RSD_{g,}$ L_1RSD_{l17} are 0.987,0.988, 0.977, 0.973, 0.984. Generally speaking, the deeper a residue it is, the more atoms number it is surrounded by. Results(Fig 2 b) also indicates that HSEAU has high similarity to HSEBD and HSEAD has high similarity to HSEBU.

Besides hydrophobicity, we also calculate the correlation coefficients between depth values, HSE values and other amino acids physicochemical propensities. The results are shown in Table1. The depth values are not only correlative with hydrophobicity, but strongly correlative with flexibility, polarity and accessibility of amino acids. Moreover, we extract 531 physicochemical propensities of amino acids from AAIndex database (Kawashima *et al.*, 1999; 2000; 2008) and calculated the correlation coefficients between depth values, HSE values and these propensities. The results are shown in **S1**. More than 140 propensities are strongly correlative with all these five kinds of depth. And $HSEAU_{13,}$ $HSEBD_{13}$ have the same propensity with depth functions.

Not only physicochemical propensities, we also study the correlation between depth values and secondary structure in the same dataset. The results are shown in **Table 2**. The secondary structures are calculated by DSSP and classified into 3 types: H(helix), E(sheet) and C(coil). From **Table 2**, sheet has highest average depth values, which means sheet tends to be buried in protein. And coil has lowest average depth values so coil tends to be on the protein surface. All these five depth values have consistent results here. The depth values of secondary structures follow the order: sheet> helix> coil. All four types of HSE have the similarity results.

Table 1 Correlation coefficients between 4 physicochemical propensity of different amino acids with means of the depth values or HSE values.

Name	Hydrophobicity	Flexibility	Polarity	Accessibility	Total Number
L1RSDg	-0.947	-0.929	-0.895	0.959	140
$L1RSDl_{17}$	-0.931	-0.915	-0.918	0.964	163
HSRSD	-0.955	-0.902	-0.908	0.962	148
DPX	-0.949	-0.884	-0.936	0.959	167
RD	-0.957	-0.889	-0.938	0.971	162
$HSEAD_{13}$	0.617	0.415	0.696	-0.632	6
$HSEAU_{13}$	-0.949	-0.884	-0.956	0.970	157
$HSEBD_{13}$	-0.945	-0.905	-0.938	0.962	160
$HSEBU_{13}$	0.345	0.350	0.288	-0.314	0

RD: residue depth. Hydrophobicity index (Fasman, 1989),Flexibility parameter for one rigid neighbor (Karplus and Schulz, 1985), Polarity (Grantham, 1974), Information value for accessibility, average fraction 35% (Biou*et al*., 1988).The total number is the number of physicochemical propensities with correlation coefficients greater than 0.7 or less than -0.7.

Table 2. Mean and standard deviation of different kinds of depth for three secondary structures

SS	H	E	C
L_1RSD_g	0.37±0.020	0.44±0.025	0.32±0.020
L_1RSD_{l17}	0.59±0.023	0.67±0.024	0.52±0.027
HSD	0.053±0.004	0.092±0.007	0.039±0.004
DPX	1.03±1.19	1.53±1.71	0.60±0.52
RD	5.62±4.52	6.77±6.92	4.91±3.38
$HSEAU_{13}$	14.80±96.53	19.12±78.08	12.39±93.09
$HSEAD_{13}$	21.15±38.26	23.13±47.84	17.73±61.30
$HSEBU_{13}$	19.15±43.07	22.47±49.07	17.41±57.02
$HSEBD_{13}$	16.80±85.08	19.77±74.45	13.89±86.70

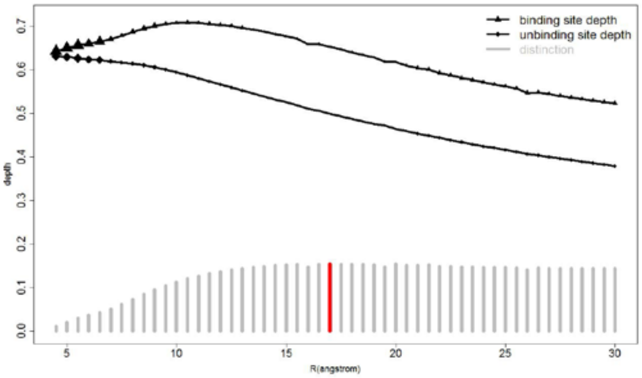
SS: secondary structure RD: residue depth.

3.2 Local L_1 depth and protein-ligand interaction

Proteins perform their biological functions by their interactions with other molecules, such as drugs, coenzymes, antigens, nucleic acids, other proteins, etc. Here we show that the local L_1 depth of protein is highly correlative with the binding residues between protein and ligands.

We calculated the local L_1 depth, residue depth, four types HSE of binding residues and unbinding residues on the protein surface in PLBD. The residues on the protein surface are considered as binding sites if the distances between the ligands and the residues are less than 4.5 angstroms (Å). The radius of local L_1 depth varied from 4.5Å to 30Å by the step 0.5 Å. And the radius of HSE varied from 7Å to 26Å by the step 1Å. The **Fig. 3 (a)** shows the difference between the average L_1local depth values of binding residues and unbinding residues over the whole PLBD for each step. The depth of binding residues are always greater than the depth of unbinding residues with any radius, which means the binding residues tend to locate at deeper surface or cavity on the surface. In other words, from our results, the region in a protein for binding ligand is usually a “pocket” on protein surface, which is consistent with the study about protein ligand interaction (Abdullah *et al.*, 2007; Deng *et al.*, 2004). With the radius increasing, the difference between binding residues and unbinding residues increases quickly until radius is greater than 10 Å. The difference (0.1536) is greatest when the radius of L_1 local depth is 17 Å. Fig 3 (b) shows the same results, the HSE of binding residues are greater than the HSE of unbinding residue after 10 Å, From 13.5Å to 20.5 Å all 33 ligands binding site’s L_1RSD is significant greater than unbinding site, while results show that none of 4 types HSE in binding site is always significantly greater than unbinding site in any radius. Result shows in radius 13Å, the total number of ligands which HSE of binding residue is not significantly greater than unbinding residue is minimal at 11 (HSEAD:2/33, HSEAU:3/33, HSEBD:1/33, HSEBU:5/33).

To study the depth difference in different ligands, we calculate the average local L_1 depth for every single ligand. Table 3 shows the average local L_1depth values and differences between binding residues and unbinding residues over each ligand. The radius shown in fifth column is the best radius by which the difference between binding and unbinding residues is greatest. The binding residues have greater depth values than unbinding residues in all 33 ligands. And their best radiuses vary from 15Å to 30Å. It is notable that best radiuses of the ligands are not relative with their size as we expect. The second column in Table 3 shows the heavy atoms of the ligands. FAD has most heavy atom number (53), but its best radius is small (18.5Å); And BTB has biggest best radius (30 Å) but this ligand is very small (18 heavy atoms). Residue depth and four types HSE are also be calculated, results show that there are five ligands which residue depth of binding residue is not significantly greater than residue depth of unbinding residue.



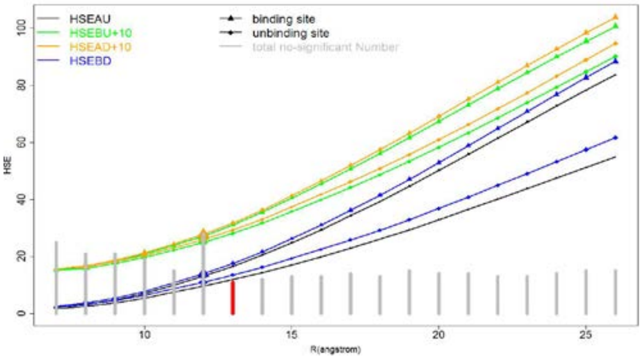


Fig. 3.(A) L1RSDl17 for binding site/unbinding site and their distinction. X axis is radius value for local L1 depth and Y axis is stand for local L1 depth. Binding site’s depth (black triangle line) is always greater than unbinding site’ depth (black diamond line) and the difference (grey line) of them get maximal when local radius is 17 Å (red line). Diamond or triangle size indicates no-significant ligands numbers. From 13.5 Å to 20.5 Å all 33 ligands binding site’s L_1RSD is significant greater than unbinding site. (B) HSE (4 types) for binding site/unbinding site and sum no-significant number of 4 types HSE in 33 ligands. Binding site’ HSE (triangle line) is always greater than unbinding site’ HSE (diamond line). For all 33 ligands, results show that none of 4 types HSE in binding site is always significantly greater than unbinding site in any radius. Diamond or triangle size indicates no-significant ligands numbers. The total no-significant numbers of 4 types HSE (grey line) get minimal 11 (HSEAU is 3, HSEBD is 1, HSEBU is 5, HSEAD is 2) when radius is 13 Å (red line).

The significant tests (T test or Wilcoxon test) suggest that the local L_1depth values of binding and unbinding residues are significantly different (p-value<0.05) in each ligand. Furthermore, even though we use the uniform radius of the local L_1 depth value 17 Å, the significant tests results of all the ligands are still positive (p-value<0.05).We also calculate the global L_1 depth for each ligand. The average global L_1 depth values of binding residues are all greater than unbinding residues as well. But the significant tests show that not all the global L_1 depth values of binding residues are greater than unbinding residues significantly. So we use L_1RSD_{l17}as the indicator to determine whether a residue is in a pocket or not. This indicator can be used to predict binding residue, too.**Fig.4**shows the difference of L_1RSD_{l17} between binding residues and unbinding residues. As comparison, DPX, HSEAU$_{13}$, HSEBU$_{13}$ and residue depth are shown in **Fig.4** as well. Although DPX and residue depth are good measures which are able to describe how far a residue or atom from the protein surface, they couldn’t give much more information about the residues on the surface, because nearly all of the residue on protein surface get the same DPX and residue depth (**Fig.4,**).However, RSD is a powerful way to measure the “hole” or “pocket” (**Fig.4**) on the protein surface. Fig. 4 also indicates HSEAU$_{13}$ is a better indicator than HSEBU$_{13}$ in pocket recognize.

Table 3. Best local L_1 depth radius value and depth for binding site/unbinding site.

Ligand	Heavy Atom Numbers	Binding L_1RSD_{lR}	Un-binding L_1RSD_{lR}	Radius(Å)	Binding RD	Un-binding RD	HSE Radius Range(Å)
1PE	13	0.56	0.49	19.0 (*)	5.38	4.99(#)	13-26
2PE	28	0.60	0.47	18.5 (*)	5.20	4.85(*)	12-26
ACO	51	0.67	0.47	21.5 (*)	5.26	4.63(*)	8-26
ACP	31	0.66	0.44	21.5 (*)	5.78	4.74(*)	9-26
ADN	19	0.72	0.50	16.5 (*)	6.76	4.75(*)	7-26
AMP	23	0.66	0.47	19.5 (*)	6.03	4.83(*)	8-26
APC	31	0.70	0.46	21.5 (*)	5.53	4.70(*)	10-26
ARG	12	0.69	0.45	23.0 (*)	6.23	4.66(*)	7-26
BGC	12	0.71	0.47	21.0 (*)	7.20	5.09(*)	7-26
BOG	20	0.61	0.45	23.0 (*)	4.91	4.78(#)	7
BTB	14	0.55	0.40	30.0 (*)	6.31	5.00(*)	8-26
CIT	13	0.80	0.54	16.0 (*)	5.99	4.93(*)	7-26
COA	48	0.74	0.48	21.5 (*)	5.12	4.67(*)	8-26
EPE	15	0.76	0.50	19.5 (*)	5.02	4.77(#)	12-23
FAD	53	0.77	0.51	18.5 (*)	7.38	4.90(*)	7-26
FLC	13	0.59	0.42	23.0 (*)	5.87	4.86(*)	8-26
FMN	31	0.69	0.46	19.5 (*)	6.45	4.70(*)	7-26
GDP	28	0.75	0.51	16.0 (*)	5.68	4.72(*)	7-26
GTP	32	0.62	0.40	26.0 (*)	5.35	4.73(*)	8-26
HEC	43	0.71	0.48	19.5 (*)	5.76	4.71(*)	15-26
HEM	43	0.76	0.52	15.0 (*)	6.40	4.71(*)	7-26
MAN	12	0.64	0.41	26.0 (*)	5.00	4.94(#)	NA
MES	12	0.67	0.44	23.0 (*)	5.28	4.91(*)	7
NAD	44	0.70	0.46	19.5 (*)	6.26	4.77(*)	7-26
NAP	48	0.67	0.44	21.5 (*)	6.22	4.80(*)	7-26
NDP	48	0.75	0.51	16.0 (*)	6.17	4.77(*)	7-26
P6G	19	0.61	0.40	26.0 (#)	5.39	5.14(#)	NA
PG4	13	0.69	0.46	20.0 (*)	5.41	5.04(*)	11-20
SAH	26	0.77	0.52	15.0 (*)	6.83	4.72(*)	7-26
SAM	27	0.74	0.52	16.0 (*)	6.37	4.67(*)	7-26
SUC	23	0.61	0.41	26.0 (*)	5.48	4.85(*)	10-26
TRP	15	0.69	0.47	20.0 (*)	5.45	4.33(*)	7-10,22-26
UDP	25	0.74	0.51	16.5 (*)	6.46	4.86(*)	7-26

RD: Residue depth, * means depth (L_1RSD or RD) of binding site is significantly greater than unbinding site. # means binding site's depth is no-significantly greater than unbinding site. NA means there is no radius value for all 4 types HSE in binding site is significantly greater than unbinding site. Further, we found the second best radius value for P6G is 16 Å. In this radius, P6G's binding site's depth is significantly greater than unbinding site.

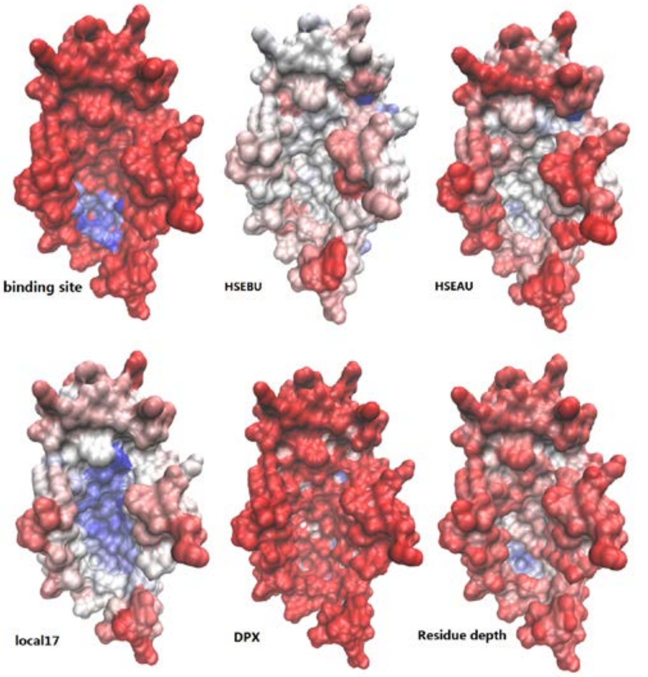


Fig.4. L1RSD117 (bottom lest), DPX (bottom middle), residue depth (bottom right), HSEBU (top middle), HSEAU (top right) and PENTAETHYLENE GLYCOL (1PE) binding site (top left, site=blue) in protein NEURONAL CALCIUM SENSOR 1 (1G8I chain A) (Bourne et al, 2001).Depth (red) <depth (white) <depth (blue).HSE (red) <HSE (white) <HSE (blue). HSEBU, Residue depth and DPX nearly give the same depth to protein surface residue, but L_1RSDl_{17} and $HSEAU_{13}$ make protein surface more layering (we found HSEBD is similar to HSEAU and HSEAD is similar to HSEBU).

3.3 Global L_1 depth and protein shape

By far, proteins are the most structurally complex and functionally sophisticated molecules known (Albertset *al.*, 2002).To knowing more about protein functions, protein structure has been widely studied and classified (Sillitoeet *al*., 2013). Different protein structures or protein shapes implement different protein functions. Here, we propose a protein structure descriptor to describe the shape of protein and classify the protein structures using statistical depth function. Specifically, we use the empirical cumulative distribution function (ECDF) of the protein L_1RSD_g as the protein structure characters.

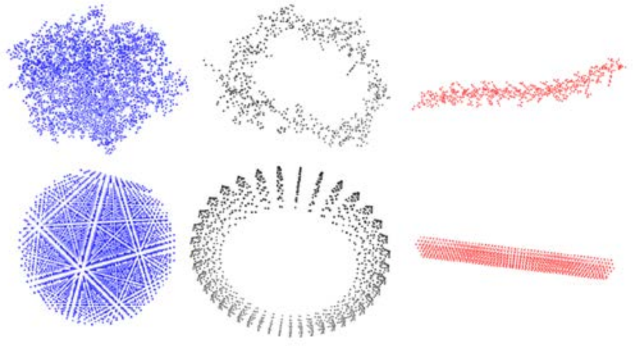


Fig. 5. Three kinds of protein shape (Top: 1HDHA, 1HFES, 1N7SB. Bottom: simulated points sets, sphere, ring, chain) we simulate some points sets which have the same shape with the given proteins, the points set have the cube grids.

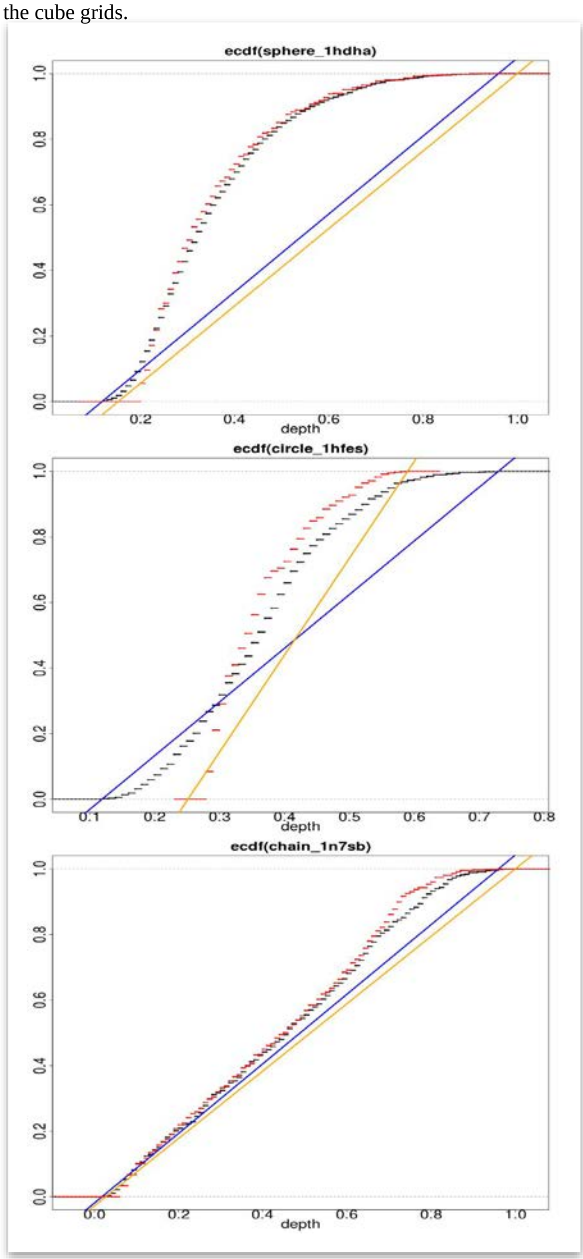


Fig. 6.L_1RSD_g's ECDF for three kinds of protein shapes (Black: sphere protein 1HDHA, ring protein 1HFES and chain protein 1N7SB in order. Red: sphere points set, ring points set and chain points set in order. Blue lines were generated by (12) for each protein. Orange lines were generated by (12) for each simulated points set.) X axis is L_1RSD_g. Three shapes protein have different character with each other in ECDF of L_1RSD_g and they are well simulated with the point sets. Ring protein's ECDF has the narrowest width. Chain protein's ECDF is the most similar to a line. Sphere protein's ECDF and ring protein's ECDF are convex.

In **Fig.5,** we give three basic protein shapes: sphere, ring and chain (**Fig.5, Top**). To study the character of different shape proteins, the same kind of shape point datasets were made (**Fig.5, bottom**). Their L_1RSD_g's ECDFs (**Fig.6**) show each shape has its own character. Sphere and ring points sets' ECDFs are convex while chain points set's ECDF is flat, and the width of ring points set's ECDF is narrower than the others'. The ECDF of L_1RSD_g is not convenient to describe or classify the protein shapes, so we use three numeric characters of ECDF which are introduced in (10) (11) (12) as the protein structure descriptor. By simulating spheres and rings which have different radius, and chains which having different length, we found the W_{max} in ring points set is around 0.4, and in sphere points set or chain points set it is about 0.9. Thus, W_{max} could be used as an indicator to distinguish rings from spheres and chains. In sphere dataset, D_{max} is around 0.4, while the value is less than 0.1 in chain points set. And K_{std}in sphere points set and ring points set are much greater than it in chain points set.

We classified the proteins in CULLPDB by using these three indexes of ECDF. The results are shown in **Table 4**.**S2** shows the detailed results which contain each protein's three indexes and protein shape type.

Table 4. Classification of protein chains in CULLPDB

Thresholds	Type	Number	Example
W_{max}<=0.5	Ring	34	3JRVC
K_{std}<=1.5 or D_{max}<=0.1	Chain	842	1JCDA
W_{max}>=0.75 and D_{max}>=0.4	Sphere	61	1LU4A
Others	Ellipsoid	1555	1BRTA

3.4 L_1 Depth and phosphorylation site

Phosphorylation is the addition of a phosphate (PO_4^{3-}) group to a protein. Protein phosphorylation is involved in a wide range of cellular processes. For example, phosphorylation turns protein enzymes on and off, thereby altering their function and activity. Phosphorylation only happens on Serine (s), Threonine (T) and Tyrosine (Y). Our research suggests that phosphorylation residues or phosphorylation oxygen atoms tend to locate in a convex on protein surface.**Fig.7** shows phosphorylation site of HUMAN TRANSTHYRETIN COMPLEXED (Meinkeand Sigler, 1999). Two phosphorylation site are both Ser (Residue 306, 316), and their L_1RSD_g are 0.15, 0.14 which is much less than other negative site. Residue 316 has the least OatomL_1ASD_{lR}value whose radius is from 7 Å to13 Å.

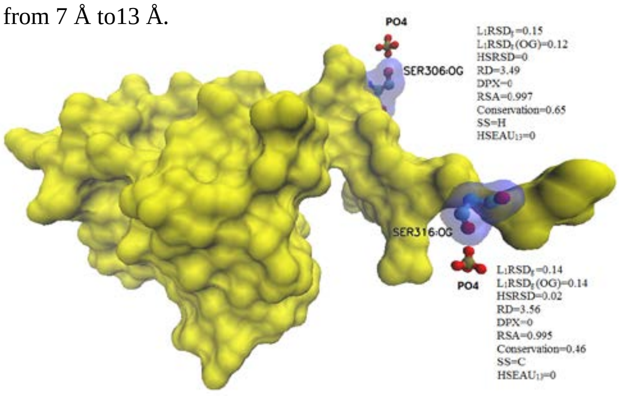


Fig. 7. Phosphorylation site of HUMAN TRANSTHYRETIN COMPLEXED (1CIT A chain), Blue: phosphorylation site (left: Residue306, right: Residue316). The L_1RSD_g of two residues in convex areas on protein surface are below 0.15.

We calculate the RSA, residue conservation, DPX, residue depth, HSRSD, L_1RSD_g, L_1RSD_{lR} and four types HSE for all Serines, Threonines, Tyrosines and oxygen atoms in hydroxyl groups, whose local L_1 depth radius is range from 4 Å to 30 Å by every 1Å and HSE radius is range from 7 Å to 25 Å by every 1 Å in Phospho3D. Table 5 shows average depth values on phosphorylation sites (P-site) and on non- phosphorylation sites (N-site).P-value of the statistical tests (t-tests or wilcoxon tests) of the difference between P-site and N-site are listed in table 5, too (column 2).Only the best top 10 p-values depth and DPX are selected to be shown in **Table 5**. All the results containing HSE are shown in **S3**.

Table 5. Depth values of phosphorylation site in phospho3D

Indexes	p-value	P-site	N-site
OatomL_1ASD_{l7}	0.002032	0.5588	0.6144
OatomL_1ASD_{l9}	0.002132	0.5648	0.6206
OatomL_1ASD_{l11}	0.002686	0.5546	0.6070
OatomL_1ASD_{l10}	0.002970	0.5630	0.6158
OatomL_1ASD_{l8}	0.003086	0.5658	0.6205
OatomL_1ASD_{l12}	0.003780	0.5465	0.5966
OatomL_1ASD_{l13}	0.006569	0.5388	0.5848
Residue Depth (Å)	0.008405	4.6959	5.2199
Oatom Residue Depth	0.008405	4.6959	5.2199
OatomL_1ASD_{l6}	0.009146	0.5568	0.6034
DPX(Å)	0.018932	0.543	0.6752
$HSEAD_{13}$	0	15.8308	19.5661
$HSEAU_{13}$	0.007348	10.7231	12.7916
$HSEBD_{13}$	0.010499	12.3	14.2564
$HSEBU_{13}$	0	14.5923	18.3657

P-site: mean of phosphorylation sites, N-site: mean of non-phosphorylation sites. Oatom represents oxygen atoms in hydroxyl groups

All the indexes in **Table 5** are significantly different between phosphorylation sites and non- phosphorylation sites. From Table 5, L_1RSD_{l7} has lowest p-value, which means the difference of L_1RSD_{l7} between P-site and N-site are most significant. If a residue is in the convex area, for example in **Fig.7**, its L_1RSD_g should be less than residues which are in the flat areas on surface. So we believe that phosphorylation site have the high probability to be located in a convex areas. The oxygen atom's local L_1 depth with radius from 7 Å to 13 Å has lower p-value, too. As we expect, the phosphorylation sites have lower depth values than non- phosphorylation sites, which suggest that the phosphorylation sites might locate in the convex area. This conclusion is consistent with the residue depth results. At last, residue depth and DPX of phosphorylation sites also have significant difference from non-phosphorylation sites. HSE also have the same result, the four types HSE of phosphorylation sites are always less than non-phosphorylation sites in any radius. It also can indicate that phosphorylation sites might prefer to locate in the convex area of protein surface.

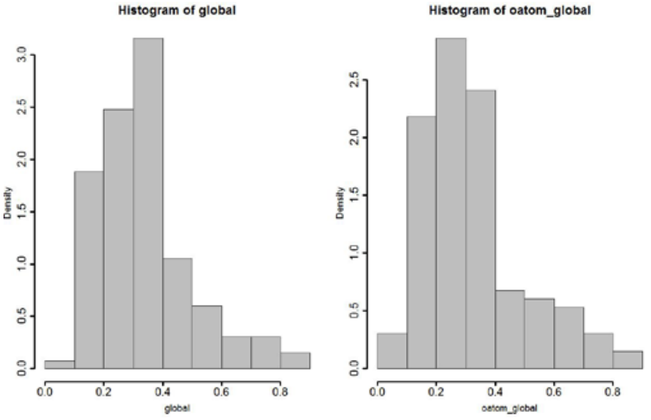


Fig. 8. Left**:** histogram of $L_1RSD_g$for phosphorylation residue. Right: histogram of $L_1ASD_g$for phosphorylation oxygen atom. X axis is global L_1 depth and Y axis is their frequency. Over 70% phosphorylation residue and phosphorylation oxygen atom L_1ASD_g is below 0.4, and 50% is below 0.3 which can be believe shallower than normal protein surface atoms' depth.

We computed the histogram of L_1RSD_g for phosphorylation residue and histogram of L_1ASD_g for phosphorylation oxygen atom (**Fig.8**). It shows there are above 50% oxygen atoms' L_1ASD_g less than 0.3. In other word, phosphorylation oxygen atom tends to locate in a convex area which is more exposed on protein surface.

3.5 L_1 depth and surface residue conservation

The conservation of residues varies a lot on the protein surface. For example, functional residues, such as protein ligand binding residue, phosphorylation site and protein-protein interaction interface, are more conserved than other residues on protein surface (Capra *et al* 2009, Ma *et al* 2003). To study the relationship between L_1 depth and the conservation of residues on protein surface, we select all the residues on protein surface in CULLPDB and calculate their L_1RSD_g and L_1RSD_{l17}. The **Fig. 9** shows the conservation values of residues vary with the changes of all kinds of depth. The conservation of reside is calculated by the output of PSI-Blast using Shannon entropy (see Method section for details).

From **Fig.9**, the residues with low depth values are most conserved on the protein surface, which suggests that the residues in the convex areas are more conserved than others. This is reasonable because lots of functional residues, such as phosphorylation sites and protein-protein interaction interface locate in the convex areas on protein surface. These residues are more conserved than others. Thus, the conservation curves have highest values at the beginning and go down very quickly. At the depth value 0.3, the conservation values get their lowest values. But after that, the curves go up slowly and conservation values achieve around 0.7eventually. The residues with high depth values are conserved, too. Similar to low depth residues, this might be because some functional sites, such as ligand binding residues, locate in the pockets or cavities of protein surface. The results in section **3.2** and **3.4** also give us the evidences to support above conclusions.

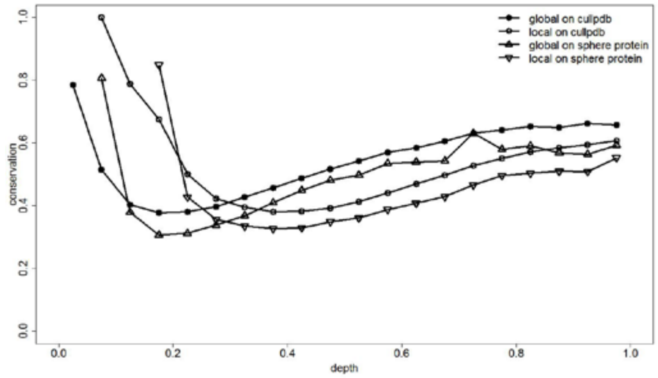


Fig.9. Mean conservation curve of surface residues. X axis is depth values and Y axis is conservation values. Sphere lines are L_1 depth calculated in CULLPDB and triangle lines are L_1 depth calculated in sphere protein dataset which was shown in **Table 4** and **S2**.All curves show the same information that with the depth values increasing the conservation values go down first and then rise to around 0.7.In other word, convex area residues and concave area (hole or pocket) residue have higher conservation.

In some proteins, such as “chain shape” protein, the L_1 depth of residues at the ends of chain could be much higher than the residues in the middle even though there is no any pocket on the surface. So we remove these kinds of proteins and calculate the depth values only on the sphere proteins. The sphere proteins are selected according to the results in section**3.3**. The results are similar with the whole CULLPDB and the curves are shown in **Fig.9**.

On the other hand, we also select the conserved residues (conservation score > 0.8) on protein surface to analyze their depth values. More than half of them (443/787) have higher L_1RSD_{l17} (>0.55), and their RSAs are less than 30% at meantime. This indicates that these residues are in the pockets. About 25% conserved residues (191/787) have low L_1RSD_{l17} (>0.3), which means they locate at convex areas on the protein surface. More than 80% conserved residues locate in the pockets or at the convex areas. **Fig.10** shows an example of relationship between depth and conservation.

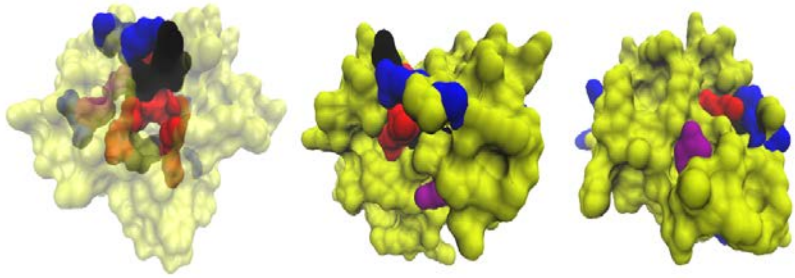


Fig. 10.High conservation (>0.8) residues in molybdenum carrier protein from SYNTROPHUS ACIDITROPHICUS SB (3IMK A chain, blue: L_1RSD_g<0.3, black: L_1RSD_{l17}>0.57, red: ligand-binding site and L_1RSD_{l17}>0.57, purple: others). Most of the high conservation residues are locate in some pockets or convex areas.

4 DISCUSSION AND CONCLUSION

Protein structure descriptor would be helpful for classifying the protein and studying the protein structure and functions. We propose here a new structure descriptor using statistical depth function, especially L_1 depth to measure the relative position of residues or atoms in protein structure or protein surface. We show that these statistical depth indexes are strongly relative with the amino acid physicochemical propensities, such as hydrophobicity, flexibility and polar. We also show that the secondary structures of residues are relative with their L_1 depth values. The residues in sheets are deeper than helix and coil.

We select more than 30 ligands and their protein complexes in PDB and show that the binding residues have significantly higher depth values than other residues on protein surface, which suggests that the ligands tend to bind to the proteins in the pockets or cavity on the protein surface. Actually, we could define the pockets or cavities using depth function and predict the protein ligand binding residues.

As structure descriptors, depth function also could describe the protein structure by their ECDFs and help us classify the protein according to the protein’s shapes. We propose 3 indexes as the shape descriptors and classify the CULLPDB using these indexes. Unlike the protein ligands binding residues, we show here that phosphorylation sites prefer to locate at the convex areas on protein surface. The L_1 depth values of phosphorylation sites are significantly lower than non-phosphorylation sites.

At last, we show that the L_1 depth of residues is relative with the conservation of the residues. As we expect, the residues with low depth values are most conserved on protein surface. But interestingly, the conservation does not decreasing with increasing depth value monotonically. The residues with high depth values also are more conserved than other residues with medium depth values.

Statistical Depth function is a novel index to describe the protein structures. For residues buried in protein, DPX, residue depth can tell us how far the residues are from protein surface or solvent. While residues are exposed on the protein surface, HSRSD and L_1 depth can give us the information about their relatively position on protein surface which DPX and residue depth cannot. DPX, residue depth and statistical depth function together can measure the relatively position of residues and atoms either buried or exposed. Halfspace depth or L_1 depth also could be used to detect the pockets or cavities on protein surface.

L_1RSD_g has high correlation with HSRSD, especially for “sphere” proteins. L_1 depth’s algorithm complexity is O(n) for a residue, while the algorithm complexity of HSRSD is $O(n^2 logn)$. For this reason, L_1RSD_g can be used instead of HSRSD if the shapes of proteins are nearly “sphere”. We also can use Genetic Algorithm

(GA) to evaluate residue's HSRSD if we do not need the exact values of HSRSD.

L_1RSD_g is a global scale to describe protein while L_1RSD_{lR} is a local scale. L_1RSD_{lR} will turn to be L_1RSD_g if the radius R big enough. For L_1RSD_{lR}, different radius can describe different protein characters, for example, residue in a hole or a convex. In structure-based site prediction using machine leaning, depth could be useful features. We will develop a pocket detection method and predict functional sites using statistical depth function in the future work.

REFERENCES

Abdullah,K. *et al*. (2007) ShapeVariation in Protein Binding Pockets and their Ligands. *Journal of Molecular Biology*, **368(1)**, 283–301.

Alberts,B.*et al*. (2002) *Molecular Biology of the Cell*. Garland Science, New York.

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Andersen,C.A.*et al*. (2002) Continuum secondary structure captures protein flexibility. *Structure***10(2)**: 175–184.

Andreas,Z. *et.al*. (2007) Phospho3D: a database of three-dimensionalstructures of protein phosphorylation sites. *Nucleic Acids Res.*, **35**(Database issue), D229–D231.

Andreas,Z. *et al*. (2011) Phospho3D 2.0: an enhanced database ofthree-dimensional structures of phosphorylationsites. *Nucleic Acids Res.*, **39** (Database issue), D268–D271.

Baker,D. and Sali,A.(2001).Protein structure prediction and structural genomics. *Science*, **294(5540)**, 93-96.

Biou,V. *et al.* (1988) Secondary structure prediction: combination of three different methods. *Protein Engineering*,**2**, 185-191.

Bourne,Y. *et al*. (2001) Immunocytochemical localization and crystal structure of human frequenin (neuronal calcium sensor 1).*J. Biol. Chem*., **276**, 11949-11955.

Branden,C. and Tooze,J. (1991). *Introduction to protein structure* (Vol. 2).Garland Science, New York.

Campbell,S.J. *et al.* (2003). Ligand binding: functional site location, similarity and docking. *Current opinion in structural biology*, **13(3)**, 389-395.

Capra,J.A., *et al* (2009) Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure.*PLoS Comput. Biol.*, **5(12)**, e1000585.

Chakravarty,S. andVaradarajan,R. (1999) Residue depth: a novel parameter for the analysis of protein structure and stability.*Structure*,**7**,723-732.

Deng,Z. *et al*. (2004) Structural Interaction Fingerprint (SIFt): A Novel Method for Analyzing Three-Dimensional Protein-Ligand Binding Interactions. *J.Med. Chem*., **47**,337-344.

Fasman,G.D. (1989) *Prediction of Protein Structure and the Principles of Protein Conformation*. Plenum, New York, NY, USA.

Fersht,A. (1999). *Structure and mechanism in protein science: a guide to enzyme catalysis and protein folding*. Macmillan.

Grantham,R. (1974) Amino acid difference formula to help explain protein evolution. *Science*,**185**, 862-864.

Hu,G. *et al.* (2012) Finding Protein Targets for Small Biologically Relevant Ligands across Fold Space Using Inverse Ligand Binding Predictions. *Structure*, **20**, 1815–1822.

Hubbard,S.J. and Thornton,J.M. (1993) 'NACCESS', Computer Program, Department ofBiochemistry and Molecular Biology, University College London.

Jones,S. and Thornton,J.M.(1996) Principles of protein-protein interactions.*Proceedings of the National Academy of Sciences*, **93**(1), 13-20.

Joosten,R.P. *et al*. (2010) A series of PDB related databases for everyday needs.*Nucleic Acids Res.*,**39**(Database issue), D411–D419.

Kabsch,W. and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.*Biopolymers*, **22**, 2577-2637.

Karplus,P.A. and Schulz, G.E. (1985) Prediction of chain flexibility in proteins. *Naturwiss*,**72**, 212-213.

Kuan,P.T. *et al.* (2011) DEPTH: a web server to compute depth and predict small-molecule binding cavities in proteins. *Nucleic Acids Res*., **39** (Web Server), W242−W248.

Kuan,P.T. *etal*. (2013) Depth: a web server to compute depth, cavity sizes, detectpotential small-molecule ligand-binding cavities and predict the pKa of ionizable residues in proteins. *Nucleic Acids Res*., **41**(Web Server), W314-W321.

Kawashima,S.*et al.* (1999) AAindex: amino acid index database. *Nucleic Acids Res.* **27**: 368-369.

Kawashima,S. and Kanehisa, M. (2000)AAindex: amino acid index database. *Nucleic Acids Res*. **28**:374-374.

Kawashima,S.*et al.* (2008) AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res.***36** (Database issue), D202-D205.

Lee,B. and Richards,F.M. (1971) The interpretation of protein structures: estimation of static of accessibility. *J. Mol. Biol.***55**, 379-400.

Liu,R.Y. (1990). On a notion of data depth based on random simplices. *The Annals of Statistics*, ***18*(1)**, 405-414.

Liu,R.Y.(1988).On a notion of simplicial depth. *Proceedings of the National Academy of Sciences*, ***85*(6)**, 1732-1734.

Ma,B.,*et al.* (2003) Protein–protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proc. Natl Acad. Sci. USA*, **100**:5772–5777.

Meinke,G. andSigler,P.B. (1999) DNA-binding mechanism of the monomeric orphan nuclear receptor NGFI-B.*Nat. Struct. Biol*.,**6**, 471-477.

Minh,N.N. and Jagath,C.R. (2005) Prediction of Protein Relative Solvent Accessibility With aTwo-Stage SVM Approach. *Proteins*, **59**, 30–37.

Pintar,A. *et al*. (2003a) Atom depth as a descriptor of the protein interior. *Biophysical Journal*,**84**, 2553–2561.

Pintar,A. *et al*.(2003b) Atom depth in protein structure and function. *Trends in Biochemical Sciences*,**28**, 593–597.

Pintar,A. *et al*. (2003c) DPX: for the analysis of the protein core. *Bioinformatics*, **19**, 313-314.

Rebecka J. (2004) Clustering and classification basedon the L_1 data depth. *Journal of Multivariate Analysis*,**90**, 67–89.

Rousseeuw,P.J. and Struyf,A. (1998) Computing location depth and regression depth in higher dimensions. *Statist. and Comput.*, **8**, 193–203.

Shen,S.Y. *et al.* (2007) Analysis of Protein Three-Dimension Structure Using AminoAcids Depths. *The Protein Journal*, **26**, 183-192.

Sillitoe,I. *et al*. (2013) New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. *Nucleic Acids Res*., **41**, D490–D498.

Thomason,P. andKay,R. (2000) Eukaryotic signal transduction via histidine-aspartate phosphorelay. *J. Cell. Sci*.,**113**, 3141–50.

Tukey, J.W. (1975). Mathematics and the picturing of data. *Proceedings of the International Congress of Mathematicians*,**2**, 523 – 531.

Wang,G. and Dunbrack,R.L. (2003) Jr. PISCES: a protein sequence culling server. *Bioinformatics*,**19**, 1589-1591.

Wang,K *et al* (2013). An Accurate Method for Prediction of Protein-Ligand Binding Site on Protein Surface Using SVM and Statistical Depth Function.*BioMed Research International*.**2013**:409658.

Wang,K. and Samudrala, R. (2006) Incorporating background frequency improves entropy-based residue conservation measures. *BMC Bioinformatics*, **7(1)**, 385.

Yeh,A.B. and Singh,K. (1997) Balanced confidence regions based on Tukey's depth and the bootstrap. *J. R. Stat. Soc. B*, **59**, 639-652.

Yu,Y. *et al*. (2010) Roll: A new algorithm for the detection of protein pockets and cavities with a rolling probe sphere. *Bioinformatics*, **26**, 46-52.

Zuo,Y.J. and Robert,S. (2000a) General notions of statistical depth function. *The Annals of Statistics*, **28**, *461–482.*

Zuo,Y.J. and Robert,S. (2000b) Structural properties and convergence results for contours of sample statistical depth functions. *The Annals of Statistics*, **28**, 483-499.