Interdiscip Sci Comput Life Sci (2011) 3: 161–174
DOI: 10.1007/s12539-011-0099-0

# Representation of Protein 3D Structures in Spherical ($\rho$, $\phi$, $\theta$) Coordinates and Two of Its Potential Applications

Vicente M. REYES*
(Department of Biological Sciences, School of Biological & Medical Sciences, College of Science, Rochester Institute of Technology, Rochester, NY 14623-5603, USA)

Received 12 November 2010 / Revised 16 January 2011 / Accepted 17 January 2011

**Abstract:** Three-dimensional objects can be represented using Cartesian, spherical or cylindrical coordinate systems, among many others. Currently all protein 3D structures in the PDB are in Cartesian coordinates. We wanted to explore the possibility that protein 3D structures, especially the globular type (spheroproteins), when represented in spherical coordinates might find useful novel applications. A Fortran program was written to transform protein 3D structure files in Cartesian coordinates (x,y,z) to spherical coordinates ($\rho$, $\phi$, $\theta$), with the centroid of the protein molecule as origin. We present here two applications, namely, (1) separation of the protein outer layer (OL) from the inner core (IC); and (2) identifying protrusions and invaginations on the protein surface. In the first application, $\phi$ and $\theta$ were partitioned into suitable intervals and the point with maximum $\rho$ in each such '$\phi$-$\theta$ bin' was determined. A suitable cutoff value for $\rho$ is adopted, and for each $\phi$-$\theta$ bin, all points with $\rho$ values less than the cutoff are considered part of the IC, and those with $\rho$ values equal to or greater than the cutoff are considered part of the OL. We show that this separation procedure is successful as it gives rise to an OL that is significantly more enriched in hydrophilic amino acid residues, and an IC that is significantly more enriched in hydrophobic amino acid residues, as expected. In the second application, the point with maximum $\rho$ in each $\phi$-$\theta$ bin are sequestered and their frequency distribution constructed (*i.e.*, maximum $\rho$'s sorted from lowest to highest, collected into 1.50Å-intervals, and the frequency in each interval plotted). We show in such plots that invaginations on the protein surface give rise to subpeaks or shoulders on the lagging side of the main peak, while protrusions give rise to similar subpeaks or shoulders, but on the leading side of the main peak. We used the dataset of Laskowski *et al.* (1996) to demonstrate both applications.
**Key words:** protein outer layer, protein inner core, computational epitope mapping, spherical coordinate system, protein double-centroid reduced representation, protein functional site prediction, clustering algorithm.
**Abbreviations:** OL, outer layer; IC, inner core; AAR, all-atom representation; DCRR, double-centroid reduced representation; LBS, ligand binding site; fb, fine binning (method); cb, coarse binning (method); $\phi\theta$b, phi-theta bin; $A_{\phi\theta b}$, area of a $\phi$-$\theta$ bin, referring to the area of the spherical rectangle bounded by two $\phi$ and two $\theta$ limits; FD, frequency distribution; FDMR, frequency distribution of maximum rho's; CSA, catalytic site atlas.

## 1 Introduction

Spherical coordinate representation involves the three coordinates $\rho$, $\phi$ and $\theta$, which, when analogized with earth measurements, $\phi$ and $\theta$ would correspond to latitudes and longitudes (in angular units) respectively, while $\rho$ would correspond to elevation, not with respect to sea level, but instead to the center of the earth, which is taken as the 'origin' of the system (e.g. http://www.math.montana.edu/frankw/ccp/multiworld/multipleIVP/spherical/learn.htm and http://math.rice.edu/~pcmi/sphere/). $\phi$ goes from 0° to 180° while $\theta$ goes from 0° to 360°; $\rho$ on the other hand, is nonnegative.

We shall discuss the two applications of the spherical coordinate representation of proteins separately in each section that follows. The first application, that of separating the protein OL from the IC will be called "OL-IC Separation", while the second application, that of the identification of protrusions and invaginations on the protein surface, will be termed "Surface Topography."

### 1.1 Application #1: OL-IC Separation

#### 1.1.1 Surface properties of proteins

It is widely established that proteins fold in such a way that hydrophilic residues are exposed on the surface while hydrophobic ones are buried in the interior, although with some exceptions, such as integral membrane proteins, *etc.* Thus in general we expect the protein surface to have different properties form the protein

*Corresponding author.
E-mail: vmrsbi@rit.edu

interior.

Surface features of proteins include shallow ligand binding sites and active sites (although some are deeply buried), protein-protein interaction sites, post-translational modifications sites, and epitopes, to name a few. Thus studies of the OL separated from the IC, and vice versa, might find use in drug design, protein inhibition, computational epitope mapping and vaccine design (Gershoni *et al.*, 2007; Tarnovitski *et al.*, 2006; Mumey *et al.*, 2003). As a proof of concept, we shall describe here a method for the prediction of candidate epitopes using the OL-IC separation method. It involves clustering of points/atomic coordinates in the OL, then screening for locations on it which contains points/atomic coordinates above a certain density as potential epitopes.

We shall also briefly address the hypothesis that, since they are exceptions to the rule, any hydrophobic residues found within the OL might have biological roles.

#### 1.1.2 Buried properties of proteins

Buried features of proteins include deep ligand binding sites and active sites (although some are shallow or close to the surface), prosthetic group binding sites, deep metal ion binding sites, and other features largely hitherto unidentified that contribute to the overall stability and integrity of the folded protein structure.

The protein IC is largely hydrophobic as the aqueous environment of the cell prefers to interact with the hydrophilic residues which orient themselves on the surface to attain the most energetically stable overall configuration of the protein molecule. Our hypothesis regarding the IC is that occurrence of hydrophilic residues there must have some functional significance to overcome the energy constraint. We thus screened the IC of our test proteins (Laskowski *et al.*, 1996) for hydrophilic residues and demonstrate potential applicability of this method for predicting buried functional sites in proteins.

### 1.2 Application #2: Surface Topography

The second application involves characterization of the exterior topography of proteins by the detection of invaginations and protrusions on the surface. This application is different from the first in that it does not involve separation of the OL from the IC, but merely an investigation of the "surface positions" in a protein molecule, *i.e.*, those positions which are farthest away from the protein centroid. If the protein structure is in spherical coordinate representation with its centroid as origin - as they are in our algorithm - these are the points with maximum $\rho$ in each $\phi$-$\theta$ bin. These positions with $\rho$ maxima may be thought of as objects on the surface of the earth if the earth was the protein. In this application, we analyze the frequency distribution (FD) of such $\rho$ maxima and, as we demonstrate in the next section, features emerge in the FD plot that indicate the presence of protrusions and invaginations on the protein surface.

This application may be of practical importance because protrusions and invaginations on the protein surface commonly have biological significance. For example, clefts may represent ligand binding sites, and protrusions may represent loops or small lobes that open and close onto a binding pocket.

## 2 Methods

Both applications require that we partition or bin both $\phi$ and $\theta$ in order to create "$\phi$-$\theta$ bins". This binning process can be done coarsely or finely: we define "coarse binning" as partitioning both $\phi$ and $\theta$ into 10° intervals, while we define "fine binning" as partitioning $\phi$ into 6° intervals and $\theta$ into 8° intervals. Coarse binning results in 18 x 16 = 648 $\phi$-$\theta$ bins, while fine binning results in 30 x 45 = 1,350 $\phi$-$\theta$ bins. We typically use fine binning for proteins in all-atom representation (AAR), while we typically use coarse binning for proteins in reduced representation called 'DCRR' (see below and references following). The two applications dichotomize after the $\phi$-$\theta$ binning step. All constructions and calculations were accomplished by writing and executing Fortran 77 or 90 programs in a UNIX environment.

### 2.1 Application #1: OL-IC Separation

With a computational tool to virtually separate the protein OL from its IC, several novel protein structure analytical investigations become more tractable. For instance, the OL can be searched for potential epitopes or protein-protein interactions sites, while the IC can be screened for potential ligand binding sites (LBS) or catalytic sites.

#### 2.1.1 Conversion of AAR to DCRR

In a PDB file, the protein structure is represented as an all-atom representation (AAR), where each individual atom of the protein has a Cartesian coordinate. We use a reduced structure representation called the Double Centroid Reduced Representation (DCRR); there is a web server for it and the URL is http://tortellini.bioinformatics.rit.edu / vns4483 / dcrr.php (Sheth, 2009; Reyes and Sheth, in press). In this reduced representation, each amino acid is represented as two data points: the centroid of the backbone atoms (N, Cα, C and O), and the centroid of the side-chain atoms (Cβ and beyond). The centroid of the backbone is calculated by finding the average position of the backbone atoms N, Cα, C, and O); similarly the centroid of the side-chain is calculated by finding the average position of the side-chain atoms (Cβ and beyond). One advantage of using DCRR instead of AAR for the protein structure is in reducing the noise, or false positives, that may come up during our analysis.

### 2.1.2 Conversion of Cartesian to spherical coordinates

The algorithm takes a PDB file, which contains the Cartesian coordinates of the protein, as the first input. The second input is the protein molecular centroid, which is the average of the x, y, and z coordinates of all the (non-hydrogen) atoms in the protein. The entire protein molecule is then translated so that its centroid is at the origin, (0,0,0). Then the protein Cartesian coordinates are converted to spherical coordinates, ($\rho$, $\phi$, $\theta$), using a Fortran 90 program written for the purpose.

### 2.1.3 Binning the spherical representation of the protein

The spherical representation of the protein is partitioned, or binned, in two different modes. In the 'fine binning' mode, the protein is partitioned into 6° intervals in $\phi$, and 8° intervals in $\theta$, while in the 'coarse binning' mode, it is partitioned into 10° intervals in both $\phi$ and $\theta$. Since $\phi$ goes from 0° to 180° while $\theta$ goes from 0° to 360° the fine binning mode yields 1,350 $\phi$-$\theta$ bins, and the coarse binning mode yields 648 $\phi$-$\theta$ bins. The $\phi$-$\theta$ binning process is illustrated in Fig. 1, panels A and B.

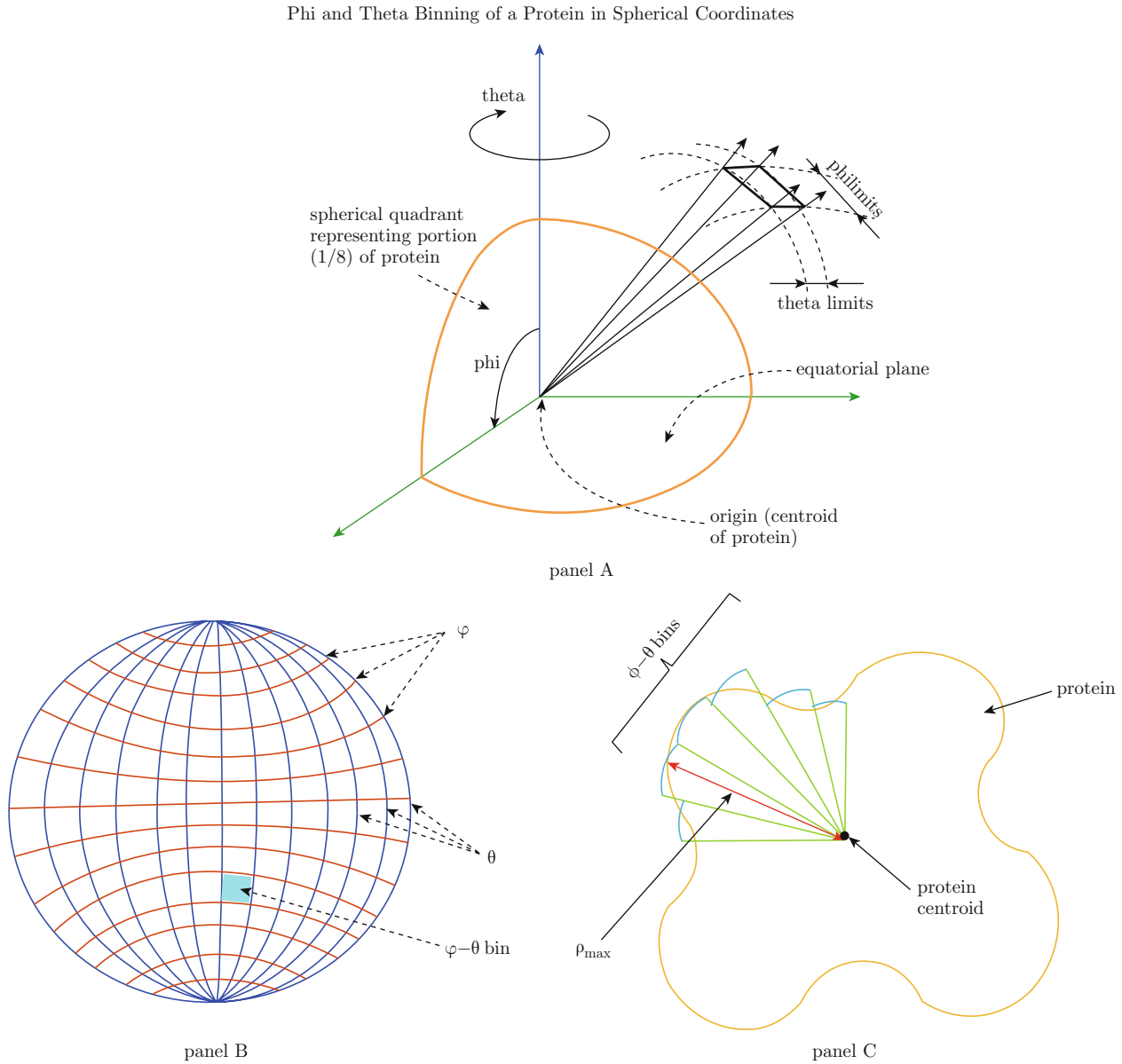



Fig. 1 The phi-theta binning process. The $\phi$-$\theta$ binning procedure is illustrated in panel A. The $\phi$ and $\theta$ limits used in the binning process are shown. Binning may be fine or coarse depending on the widths of these limits. The surface portion of a $\phi$-$\theta$ bin for a perfect sphere is shown in panel B, the area of which depends on $\phi$ (for a given $\theta$ width). Since proteins are not perfect spheres, the area of the surface of a $\phi$-$\theta$ bin in such a case depends on $\rho$ as well as on $\phi$ (panel C).

### 2.1.4 Separation of the inner core (IC) and the outer layer (OL)

The majority of proteins, especially the globular type, have an inner core (IC) analogous to a 'medulla', and an outer layer (OL) analogous to a 'cortex'. In separating the OL from the IC, the maximum $\rho$ value in each $\phi$-$\theta$ bin is first determined. Then for each $\phi$-$\theta$ bin, the protein coordinates with $\rho$ values less than an empirically determined 'cutoff' $\rho$ value (typically 95%±3% of the maximum $\rho$ in the particular $\phi$-$\theta$ bin) are assigned to the IC, while those $\rho$ values equal to or greater than the cutoff $\rho$ value are assigned to the OL. Note that if the protein is in AAR, these points are individual atomic coordinates; if the protein is in DCRR, these points are backbone or side chain centroids. This separation process using $\rho$ cutoff values is illustrated in Fig. 2.

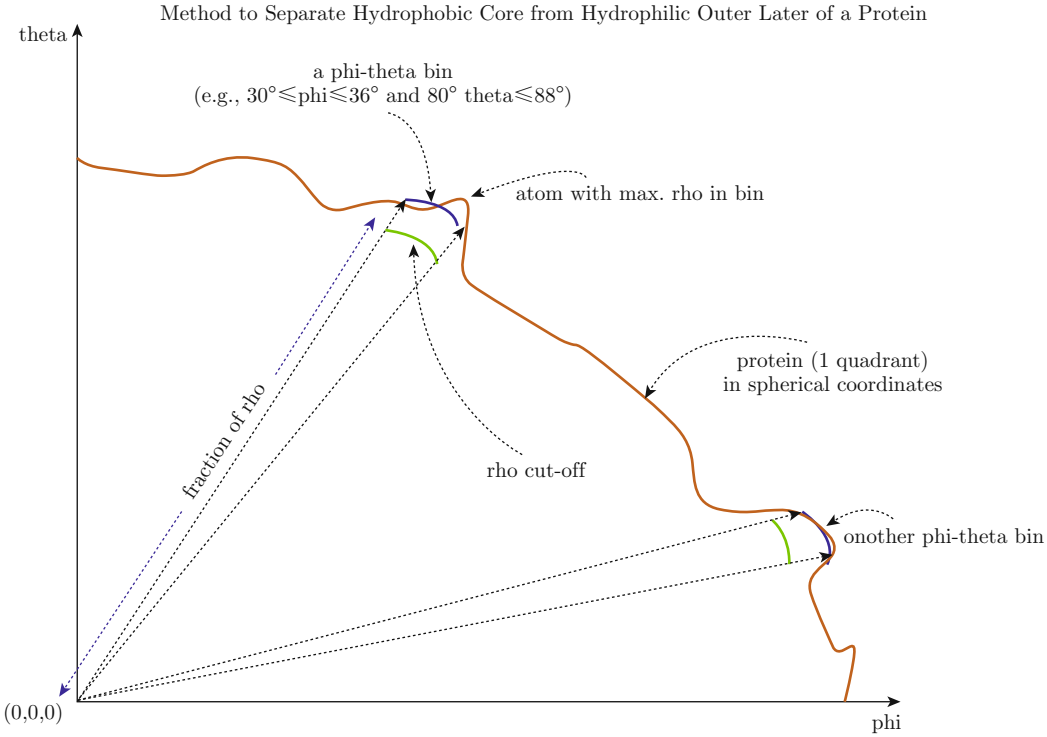



Fig. 2 Derivation construction of protein hydrophobic core. A rho cutoff based on the point (atomic or centroid coordinates) with maximum rho inside each phi-theta bin is illustrated. All other points in the bin are then classified as belonging to the IC or the OL of the protein based on whether they are less than or greater than this cutoff value, respectively.

With the protein in DCRR and its OL separated from its IC, the amino acid residues naturally fall into four classes, and we term them as follows (see Fig. 3): (a) "OL residues", in which case both backbone and side chain centroids are located within the OL; (b) "IC residues", in which case both backbone and side chain centroids are located within the IC; (c) "boundary inward residues", in which case the backbone centroid is in the OL while the side chain centroid is in the IC; and (d) "boundary outward residues", in which case the backbone centroid is in the IC while the side chain centroid is in the OL. However, a simpler classification is possible if only the side chain centroids are considered as these points will lie either on the IC or OL but only extremely rarely on the boundary. We shall adopt this last classification.

### 2.1.5 Using the OL to predict candidate epitopes

Our epitope prediction algorithm assumes that potential epitopes are clusters of points (atoms or centroids) on the OL. We thus assign a point density for each $\phi$-$\theta$ bin on the OL; this number is equal to the number of p oints (atoms or centroids) in the $\phi$-$\theta$ bin divided by the area of the $\phi$-$\theta$ bin, which we denote by $A_{\phi\theta b}$. Note that by $A_{\phi\theta b}$ we mean the area of the spherical rectangle bounded by two $\phi$ and two $\theta$ limits. Whereas the number of points in a $\phi$-$\theta$ bin is readily obtained from our algorithm programs, the areas of the $\phi$-$\theta$ bins vary with respect to their location on the surface, with those close to the "equator" ($\phi = 90°$) significantly larger than those lying close to the "poles" ($\phi = 0°$ and $180°$). They also vary with respect to their distances from the protein molecular centroid, with those farther from the molecular centroid (large $\rho$'s) larger than those closer to the molecular centroid (small $\rho$'s). We thus require a formula for the area of a $\phi$-$\theta$ bin that captures these two dependencies - this idea is illustrated in Fig. 1C. Such a formula

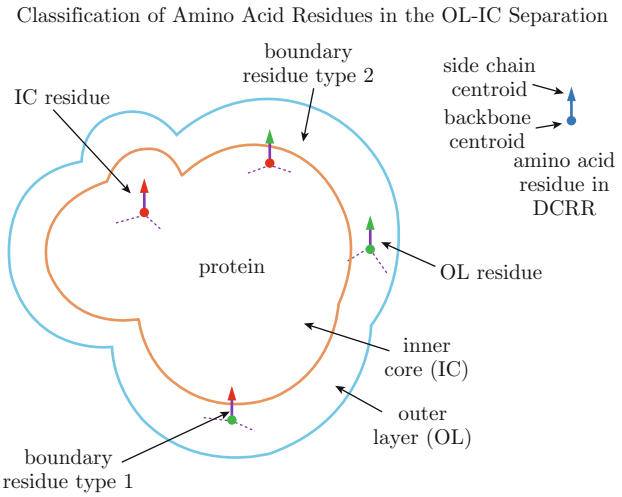



Fig. 3 Classifying amino acid residues after phi-theta binning. After applying the OL-IC separation procedure for a protein in DCRR, its amino acid residues may be classified as being located in the OL or the IC, based simply on their side chain centroids. If backbone centroids are considered, four possibilities are possible, since residues near the boundary can have backbone and side chain centroids in the OL or IC, respectively, and vice versa.

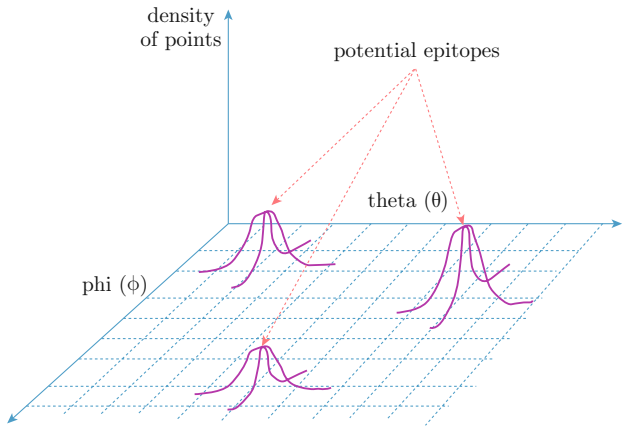



Fig. 4 Searching for point clusters in the OL. The density of points within the OL of a phi-theta bin can be used to gauge point clustering on the surface of a protein. However, a pre-processing is needed, since the phi-theta bins have unequal areas. This can be done by using a normalization factor (see text). When the point densities are then plotted against phi and theta, peaks will correspond to such point clusters, which may in turn correspond to potential epitopes.

is:

$$A_{\phi\theta b} = \left(\frac{\pi\rho_{\max}^2}{180}\right)|\theta_1 - \theta_2||\sin\phi_1 - \sin\phi_2|$$

where $A_{\phi\theta b}$ is the area of a $\phi$-$\theta$ bin that is bounded by $\phi_1$ and $\phi_2$, and by $\theta_1$ and $\theta_2$, and $\rho$ is the maximum $\rho$ in that $\phi$-$\theta$ bin. The derivation of this formula is straightforward and is left to the reader. The point density, D, for each $\phi$-$\theta$ bin is thus equal to $B/A_{\phi\theta b}$, where B is the number of coordinates in the $\phi$-$\theta$ bin. For fine binning, the factor $|\theta_1$-$\theta_2|$ is 8° while for coarse binning it is 10° Hence $A_{\phi\theta b}$ depends only on $\phi$ and $\rho$ for a given binning mode. A typical plot of point density (z-axis) as a function of $\phi$ and $\theta$ (x-y plane) is illustrated in Fig. 4. Peaks in this plot represent clusters of points on the OL of the protein.

The final step in the algorithm is in determining candidate epitopes from the point density plots. A cutoff value is determined using a distribution curve that is generated by running the density calculation against a comprehensive set of PDB files. From this distribution curve, the top 10% from the mean, is taken to represent potential epitopes. Currently we use the values 8.4 for the coarse binning mode, and 6.6 for the fine binning mode, as minimum point densities in a $\phi$-$\theta$ bin for a point cluster to be classified as a candidate epitope.

#### 2.1.6 Using the IC to find possible protein functional sites

After the protein has been separated into the OL and the IC, the IC can be analyzed to find deeply buried active or functional sites. Most proteins fold in a way such that hydrophilic residues are located in the OL while hydrophobic residues are located within the IC of the protein. The IC is thus scanned (using Perl and/or UNIX scripts) for the presence of hydrophilic resides which could be potential buried active or functional sites of the protein.

### 2.2 Application #2: Surface Topography

#### 2.2.1 An artificial protein

In order to illustrate the method in an ideal system, we constructed an "artificial protein" in the form of equally-spaced grid of points inside the scalene ellipsoid:

$$\left(\frac{x}{22}\right)^2 + \left(\frac{y}{26}\right)^2 + \left(\frac{z}{30}\right)^2 = 1.0$$

The distance between neighboring points is 1.5 units along the x, y and z directions to mimic molecular bond lengths in Å. The three major axes of this scalene ellipsoid are of non-identical lengths, namely: $X_AX_B$ where $X_A$=(22,0,0) and $X_B = (-22, 0, 0)$ along the x-axis, $Y_AY_B$ where $Y_A = (0, 26, 0, 0)$ and $Y_B = (0, -26, 0)$ along the y-axis, and $Z_AZ_B$ where $Z_A = (0, 0, 30)$ and $Z_B = (0, 0, -30)$ along the z-axis. Thus the shortest and longest dimensions of the ellipsoid are along the x- and the z-axes, respectively.

Three variants of the artificial protein were created (Figs. 5A and 5B). The first variant was one with an invagination along the shortest dimension of the ellipsoid, at $X_A = (22, 0, 0)$. Meanwhile, the second variant was one with a protrusion along the longest dimension of the ellipsoid, at $Z_A = (0, 0, 30)$. Finally, the third

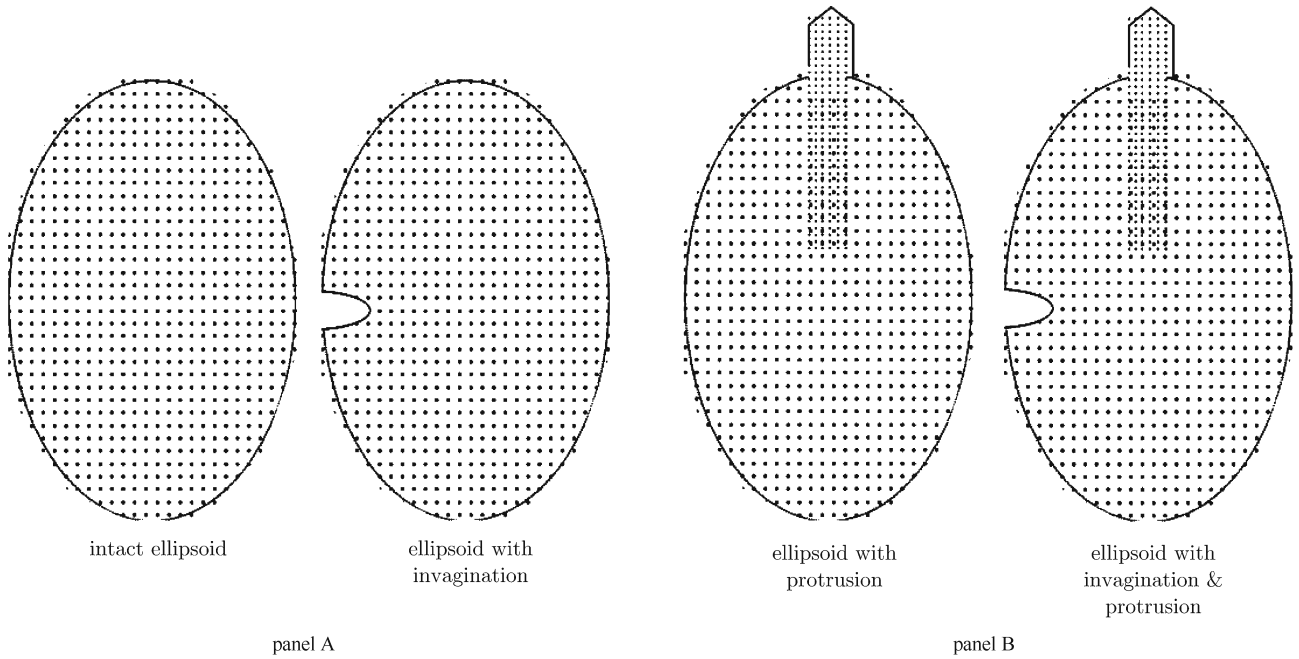



Fig. 5 The artificial protein and its variants. Panel A shows the artificial protein as a grid of points inside a scalene ellipsoid, as well as a variant of it with an invagination. Panel B shows a second variant containing a protrusion, as well as a third variant containing both an invagination and a protrusion.

variant was one with both an invagination along the shortest dimension of the ellipsoid, at $X_A = (22, 0, 0)$, and at the same time a protrusion along the longest dimension of the ellipsoid, at $Z_A = (0, 0, 30)$. The invagination at $X_A = (22, 0, 0)$ was created by "scooping out" a hemisphere points inside the sphere with center at $X_A = (22, 0, 0)$ and radius 4.0. The protrusion at $Z_A = (0, 0, 30)$ was created by translating by 2.0 units those points in the ellipsoid lying inside the cylinder with axis $z = 0$ and radius 2.0, and replacing the points vacated on the other end of the ellipsoid at $Z_B = (0, 0, -30)$ so that there would not be any invaginations there.

#### 2.2.2 Frequency distribution of maximum rhos

The set of maximum $\rho$ values collected in the last section are arranged from lowest to highest, partitioned or binned in 1.50 Å intervals, and the number of maximum $\rho$ values in each interval counted. These are then plotted as a frequency distribution, with the $\rho$ intervals or bins on the horizontal axis and the frequency in each bin along the vertical axis. We call these plots 'FDMR plots'. This step was applied to both the artificial protein and the real test proteins.

## 3 Results and discussion

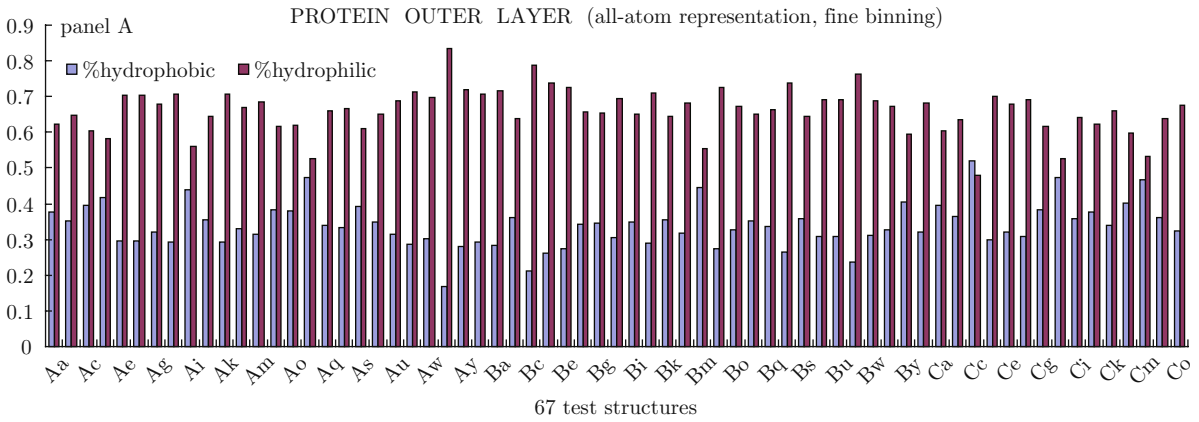
### 3.1 Application #1: OL-IC Separation

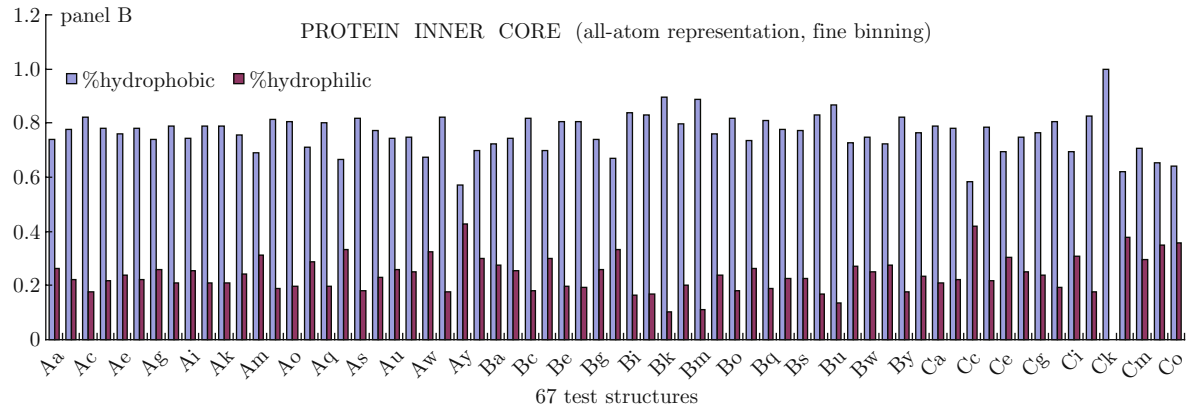
#### 3.1.1 Amino acid compositions of OL and IC

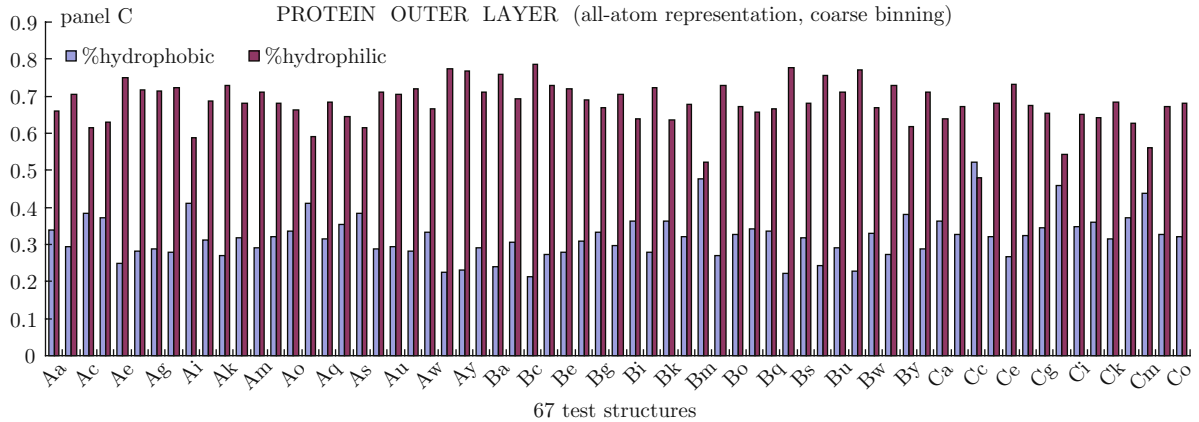
In the following discussion, we classified the amino acid residues based on the location of their side chains - whether they are located in the OL or IC; the backbone positions were not considered. Thus 'boundary outward' (as described earlier) amino acids were considered to be in the OL, while 'boundary inward' residues are considered to be in the IC.

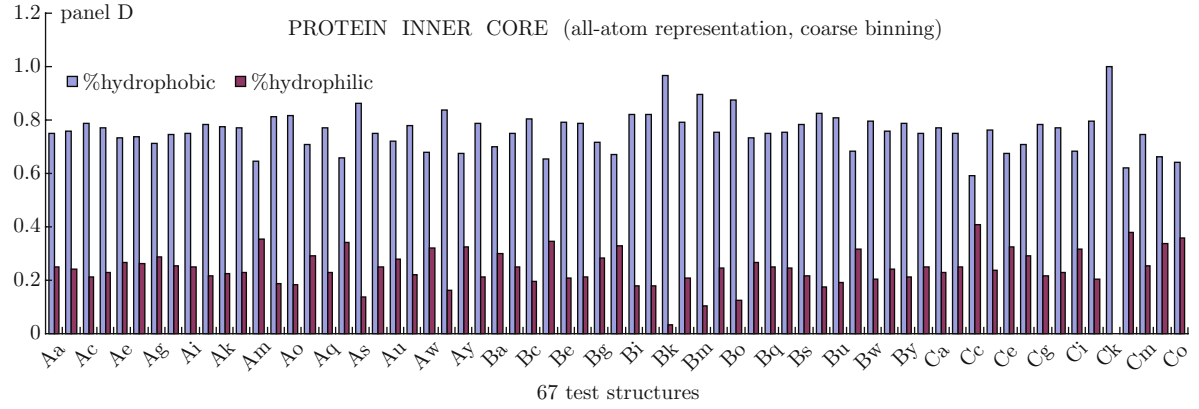
In the following discussions, refer to Fig. 6 (6 panels) and Table 1, which lists the PDB IDs of each protein in the dataset and their 2-letter abbreviations (upper case, lower case) we used. For proteins in AAR, fine binning, OLs of the structures all contain significantly higher percentages of hydrophilic amino acids than hydrophobic ones (Fig. 6A), except for structure Cc (PDB ID: 2POR, Weiss and Schultz, 1992), whose OL has slightly more hydrophobic residues. In structures Ap (2YHX, Anderson *et al.*, 1978), Bm (1HNE, Navia *et al.*, 1989), Ch (1MNS, Landro *et al.*, 1994) and Cm (2CND, Lu *et al.*, 1995), the OL is predominantly hydrophilic as expected, but to an extent that seems to be less than average for the group.

On the other hand, the ICs of all the 67 test structures all contain considerably higher percentages of hydrophobic amino acid residues than hydrophilic ones (Fig. 6B); in one case, that of Ck (2ABK, a DNA endonuclease III; Thayer *et al.*, 1995), the inner core is 100% hydrophobic. In proteins Ay (1ROB, Lisgarten *et al.*, 1993), and Cc (2POR), the IC is predominantly hydrophobic as expected, but to an extent that seems to be less than average for the group. We also performed coarse binning for proteins in AAR, and the results are very similar to the above (Figs. 6C and 6D).


panel A
PROTEIN OUTER LAYER (all-atom representation, fine binning)
%hydrophobic
%hydrophilic
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Aa Ac Ae Ag Ai Ak Am Ao Aq As Au Aw Ay Ba Bc Be Bg Bi Bk Bm Bo Bq Bs Bu Bw By Ca Cc Ce Cg Ci Ck Cm Co
67 test structures


panel B
PROTEIN INNER CORE (all-atom representation, fine binning)
%hydrophobic
%hydrophilic
1.2
1.0
0.8
0.6
0.4
0.2
0
Aa Ac Ae Ag Ai Ak Am Ao Aq As Au Aw Ay Ba Bc Be Bg Bi Bk Bm Bo Bq Bs Bu Bw By Ca Cc Ce Cg Ci Ck Cm Co
67 test structures


panel C
PROTEIN OUTER LAYER (all-atom representation, coarse binning)
%hydrophobic
%hydrophilic
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Aa Ac Ae Ag Ai Ak Am Ao Aq As Au Aw Ay Ba Bc Be Bg Bi Bk Bm Bo Bq Bs Bu Bw By Ca Cc Ce Cg Ci Ck Cm Co
67 test structures


panel D
PROTEIN INNER CORE (all-atom representation, coarse binning)
%hydrophobic
%hydrophilic
1.2
1.0
0.8
0.6
0.4
0.2
0
Aa Ac Ae Ag Ai Ak Am Ao Aq As Au Aw Ay Ba Bc Be Bg Bi Bk Bm Bo Bq Bs Bu Bw By Ca Cc Ce Cg Ci Ck Cm Co
67 test structures

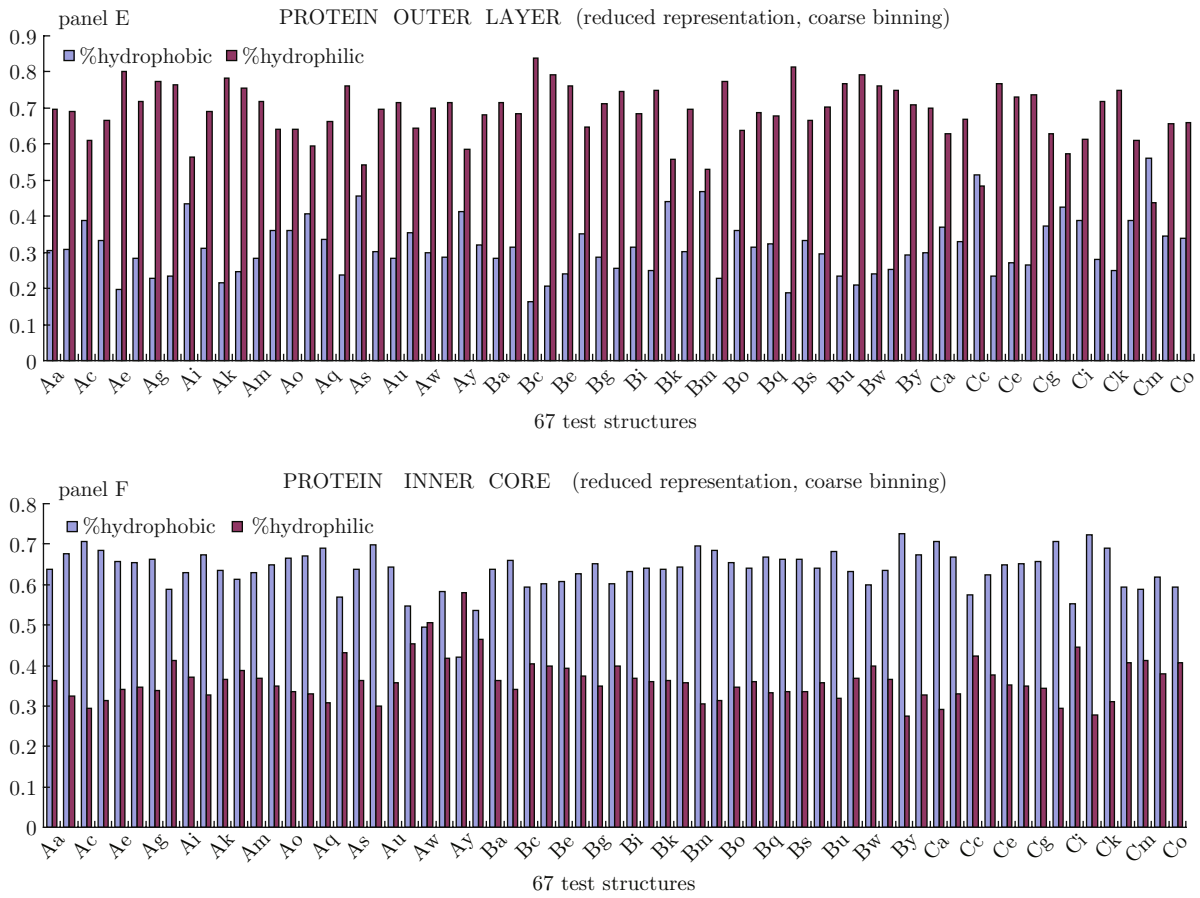



Fig. 6 Hydrophobicity/hydrophilicity ratios of the protein OLs versus ICs. The number of amino acid residues that are hydrophilic and hydrophobic in the OL and the IC are plotted for each of the 67 proteins in the Laskowski data set. Three methods were used: (a) all-atom representation with fine binning: panel A shows the ratios in the OL, panel B, those in the IC; (b) all-atom representation with coarse binning: panel C shows the ratios in the OL, panel D, those in the IC; and (c) double-centroid reduced representation with coarse binning: panel E shows the ratios in the OL, panel F, those in the IC.

For proteins in DCRR, coarse binning, the OLs of the structures have all have significantly higher percentages of hydrophilic amino acid residues than hydrophobic residues (Fig. 6E), except for structures Cc (2POR) and Cm (2CND), which have a higher percentage of hydrophobic residues. In structures As (2CUT, Martinez *et al.*, 1994), Bk (2ALP, Fujinaga *et al.*, 1985) and Bm (1HNE) the OL is predominantly hydrophilic as expected, but to an extent that seems to be less so than the average for the group.

On the other hand the ICs of all the structures have higher percentages of hydrophobic amino acid residues than hydrophilic ones (Fig. 6F), except in Aw (1ONC, Mosimann *et al.*, 1994) and Ay (1ROB) which are predominantly hydrophilic. In Av (1RNH, Yang *et al.*, 1990), Az (1SNC, Loll and Lattman, 1989) and Ci (3PGM, Winn *et al.*, 1981), the inner core is still predominantly hydrophobic but to an extent that seems to be less than the average for the group.

In summary, the major exception cases for the OL (*i.e.*, higher percentage of hydrophobic than hydrophilic amino acid residues) are 2POR and 2CND, while the major exception cases for the IC (*i.e.*, higher percentage of hydrophilic than hydrophobic amino acids) are 1ONC and 1ROB. Protein 2POR is porin, while 2CND is nitrate reductase from corn (*Zea mays*); both are integral membrane proteins (Trieschmann *et al.*, 1996; Ward *et al.*, 1989) that must have hydrophobic OLs. On the other hand, protein 1ONC is P-30, an amphibian ribonuclease, while 1ROB is bovine ribonuclease A. Type A ribonucleases such as 1ONC and 1ROB are known to be composed of two flaps or flattened lobes, at the interface of which lie positively charged residues which together bind the negatively charged RNA substrate, thus explaining their hydrophilic ICs.

Taken together, the above results clearly demonstrate that our spherical coordinate system-dependent OL-IC separation algorithm does its job with reasonable accuracy.

**Table 1 Tabulation of the 67 test structures from the Laskowski dataset and their abbreviations**

| PDB ID | abbrev. | PDB ID | abbrev. | PDB ID | abbrev. |
|---|---|---|---|---|---|
| 1ADS | Aa | 1FUT | Ax | 1EPM | Bt |
| 9ICD | Ab | 1ROB | Ay | 1MPP | Bu |
| 1IPD | Ac | 1SNC | Az | 1HYT | Bv |
| 1GOX | Ad | 1CDG | Ba | 1IAG | Bw |
| 1TDE | Ae | 1BYB | Bb | 1ADD | Bx |
| 1OYB | Af | 1XNB | Bc | 2DHC | By |
| 2NPX | Ag | 2SIM | Bd | 1PII | Bz |
| 1CCA | Ah | 1BYH | Be | 2DKB | Ca |
| 1ARP | Ai | 1FMP | Bf | 1CSH | Cb |
| 1PBE | Aj | 1BLL | Bg | 2POR | Cc |
| 1HMY | Ak | 2CTC | Bh | 1CIL | Cd |
| 3CLA | Al | 2GMT | Bi | 8ACN | Ce |
| 1GPB | Am | 1PPC | Bi | 5ENL | Cf |
| 1ULA | An | 2ALP | Bk | 1PDA | Cg |
| 1STO | Ao | 1ELA | Bl | 1MNS | Ch |
| 2YHX | Ap | 1HNE | Bm | 3PGM | Ci |
| 1PHP | Aq | 1PEK | Bn | 1BIB | Cj |
| 1GKY | Ar | 3SGA | Bo | 2ABK | Ck |
| 2CUT | As | 1PIP | Bp | 2ACK | Cl |
| 1THG | At | 1SMR | Bq | 2CND | Cm |
| 1RPA | Au | 1PPI | Br | 2PGD | Cn |
| 1RNH | Av | 3APR | Bs | 4BCL | Co |
| 1ONC | Aw | | | | |

### 3.1.2 Finding candidate epitopes

We tested our epitope prediction method using the dataset of 67 proteins of Laskowski *et al.*, (1996). Table 2 shows the top 10% proteins with the most number of predicted epitopes, with results from both the fine binning and coarse binning modes (MacCreary, M., personal commun.). The coarse binning mode consistently predicted more candidate epitopes than the fine binning mode. This illustrates the ability for the algorithm to produce either more sensitive or more selective results depending on which binning method is chosen. When these proteins were cross-referenced against the Immune Epitope Database (IEDB; Vita *et al.*, 2010), it was found that all of these proteins contained true epitopes and were antigenic in some capacity. Discrepancies between the numbers of predicted epitopes and true epitopes is due to several factors, including: overly sensitive candidate epitope cutoff criteria, the incomplete and ever-changing nature of the IEDB and PDB databases, and protein-protein interaction sites that appear to be candidate epitope according to the algorithm. These parameters are currently being refined.

**Table 2 Predicted epitopes from the 67 test structures from the Laskowski dataset (1996).**

| PDB protein ID | Number of bins containing candidate epitopes | | Positive epitope in IEDB? |
|---|---|---|---|
| | Coarse binning | Fine binning | |
| 1ONC | 95 | 34 | Yes |
| 1FUT | 88 | 38 | Yes |
| 1TDE | 76 | 40 | Yes |
| 1ROB | 73 | 41 | Yes |
| 1SNC | 64 | 34 | Yes |
| 1RNH | 63 | 31 | Yes |
| 2ALP | 54 | 43 | Yes |

### 3.1.3 Results for prediction of buried active sites

The analysis of finding possible deeply buried active sites was run against the same set of 67 proteins (Laskowski *et al.*, 1996) referred to above and compared to the set of catalytic sites curated in the Catalytic Site Atlas (CSA; Porter *et al.*, 2004). The number of predicted active sites for each protein using the coarse and the fine binning modes is shown in Table 3 (Kim, D.J., personal commun.). The coarse binning produced a larger number of predicted active sites than the fine binning mode throughout each protein analyzed. Again, this illustrates the ability for the algorithm to produce either more sensitive or more selective results depending on which binning method is chosen. Refinement of the details of the current binning process remains one of our major goals for this project.

### 3.1.4 Refinement tests for epitope prediction: fine vs. coarse binning and AAR vs. DCRR

The fine and coarse binning methods (fb and cb, respectively) were compared with each other with the protein in AAR as well as in DCRR; we designate these combinations as AAR/fb, AAR.cb, DCRR/fb and DCRR/cb. The algorithm outputted a reduced number of candidate epitopes with AAR/fb and AAR/cb compared to the other two since the $\phi$-$\theta$ bin size is too small and thus atoms in amino acid residues are being split between neighboring bins, or amino acid residues themselves are being precluded from becoming clustered in a single bin, thus preventing the algorithm form "seeing" (*i.e.*, detecting) it. This complication is further compounded when atoms in an amino acid residue are split between the OL and the IC of a $\phi$-$\theta$ bin.

We found that using proteins in DCRR greatly reduced the above ambiguities in the algorithm output (data not shown). When proteins are in DCRR, the amino acid residues may be classified as either in the OL or IC depending on where their side chain centroids

**Table 3 Predicted active sites within the IC of the 67 test structures from the Laskowski dataset (1996)**

| PDB code | Protein | Number of predicted residues to be active sites | |
|---|---|---|---|
| | | Coarse binning | Fine binning |
| **Oxidoreductases** | | | |
| 1ADS | Aldose reductase | 30 | 17 |
| 9ICD | Isocitrate dehydrogenase | 72 | 42 |
| 2PGD | 6-Phosphogluconate dehydrogenase | 87 | 57 |
| 1IPD | 3-Isopropylmalate dehydrogenase | 40 | 22 |
| 1GOX | Glycolate oxidase | 34 | 19 |
| 1TDE | Thioredoxin reductase | 44 | 33 |
| 2CND | Nitrate reductase | 33 | 21 |
| 1OYB | Old yellow enzyme | 64 | 38 |
| 2NPX | NADH peroxidase | 72 | 50 |
| 1CCA | Cytochrome c peroxidase | 33 | 19 |
| 1ARP | Peroxidase | 25 | 14 |
| 1PBE | P-hydroxybenzoate hydroxylase | 52 | 28 |
| **Transferases** | | | |
| 1HMY | HHAL DNA methyltransferase | 40 | 24 |
| 3CLA | Chloramphenicol acetyltransferase | 22 | 13 |
| 1GPB | Glycogen phosphorylase b | 181 | 138 |
| 1ULA | Purine nucleoside phosphorylase | 28 | 9 |
| 1STO | Orotate phosphoribosyltransferase | 24 | 10 |
| 2YHX | Yeast hexokinase B | 48 | 35 |
| 1PHP | 3-Phosphoglycerate kinase | 67 | 38 |
| 1GKY | Guanylate kinase | 17 | 10 |
| **Hydrolases** | | | |
| 2CUT | Cutinase | 9 | 8 |
| 1THG | Lipase triacylglycerol hydrolase | 65 | 39 |
| 2ACK | Acetylcholinesterase | 77 | 42 |
| 1RPA | Prostatic acid phosphatase | 53 | 36 |
| 1RNH | Ribonuclease H | 14 | 9 |
| 1ONC | P-30 protein | 8 | 1 |
| 1FUT | Ribonuclease F1 | 3 | 2 |
| 1ROB | Ribonuclease A | 6 | 3 |
| 1SNC | Staphylococcal nuclease | 7 | 6 |
| 1CDG | Cyclodextrin glycosyltransferase | 89 | 64 |
| 1BYB | Beta-amylase | 58 | 40 |
| 3EOJ | Bacteriochlorophyll-A protein | 41 | 28 |
| 1XNB | Xylanase | 5 | 2 |
| 2SIM | Sialidase | 45 | 27 |
| 1BYH | Glucanohydrolase H | 12 | 7 |
| 1FMP | Ricin | 22 | 16 |
| 1BLL | Leucine aminopeptidase | 77 | 63 |
| 2CTC | Carboxypeptidase A | 29 | 20 |
| 2GMT | Gamma chymotrypsin | 7 | 5 |
| 1PPC | Trypsin | 8 | 6 |
| 2ALP | Alpha-lytic protease | 5 | 1 |
| 1ELA | Elastase | 14 | 3 |
| 1HNE | Human neutrophil elastase | 8 | 4 |
| 1PEK | Proteinase K | 19 | 6 |
| 3SGA | Proteinase A | 7 | 1 |
| 1PIP | Papain | 17 | 13 |
| 1SMR | Renin | 28 | 18 |
| 1PPL | Penicillopepsin | 19 | 11 |
| 3APR | Rhizopuspepsin | 27 | 18 |
| 1EPM | Endothiapepsin | 22 | 12 |
| 1MPP | Renin | 34 | 23 |
| 1HYT | Thermolysin | 35 | 23 |
| 1IAG | Adamalysin II | 14 | 6 |
| 1ADD | Adenosine deaminase | 49 | 29 |
| 2DHC | Haloalkane dehalogenase | 31 | 16 |
| **Lyases** | | | |
| 1PII | Anthranilate isomerase | 65 | 46 |
| 2DKB | Decarboxylase | 57 | 39 |
| 1CSH | Citrate synthase | 63 | 31 |
| 2POR | Porin | 48 | 37 |
| 1CIL | Carbonic anhydrase ll | 27 | 17 |
| 8ACN | Aconitase | 130 | 92 |
| 5ENL | Enolase | 56 | 32 |
| 2ABK | Endonuclease III | 24 | 16 |
| 1PDA | Porphobilinogen deaminase | 41 | 21 |
| **Isomerases** | | | |
| 1MNS | Mandelate racemase | 39 | 27 |
| 3PGM | Phosphoglycerate mutase | 28 | 19 |
| **Ligases** | | | |
| 1BIB | Bira bifunctional protein | 32 | 20 |

are located, which is almost never exactly on the OL-IC boundary. But when the backbone centroids are taken into account, the amino acid residues may be of four types: (a) both backbone and side chain centroids are located within the OL; (b) both backbone and side chain centroids are located within the IC; (c) the backbone centroid is in the OL while the side chain centroid is in the IC; and (d) the backbone centroid is in the IC while the side chain centroid is in the OL (Fig. 3). We term the first case "OL residues", the second case "IC residues", the third case "boundary inward residues", and the fourth case "boundary outward residues". To sum up, we conclude that the best combination method is DCRR/cb, although use of the other three combinations in some special cases (e.g., small proteins or

structured peptides) is not ruled out.

### 3.2 Application #2: Surface Topography

#### 3.2.1 Use of artificial protein to test method of identifying invaginations and protrusions on protein surface

Before applying our procedure to the 67 proteins in our test set, we applied it to the artificial protein we have created and its three variants (Figs. 5A and 5B).

The resulting plots are shown in Figs. 7A-7D. The plot for the intact ellipsoid in panel A shows a lone main peak without any shoulder or subpeak anywhere in the plot. The plot for the first variant with the invagination in panel B clearly shows a subpeak on the lagging side of the main peak. A subpeak on the lagging side of the main peak in the FDMR plot is therefore diagnostic of an invagination, especially if this invagination is along the shortest dimension of the ellipsoid. If the invagination does not lie along the shortest dimension, it may be masked in the plot by some points along the shortest dimension, and a subpeak on the lagging side of the main peak will be difficult to see. In such case, the invagination will manifest itself as a non-coincidence between the superimposed FDMR plots of the intact ellipsoid and that of the invagination-containing variant elsewhere along the plot.

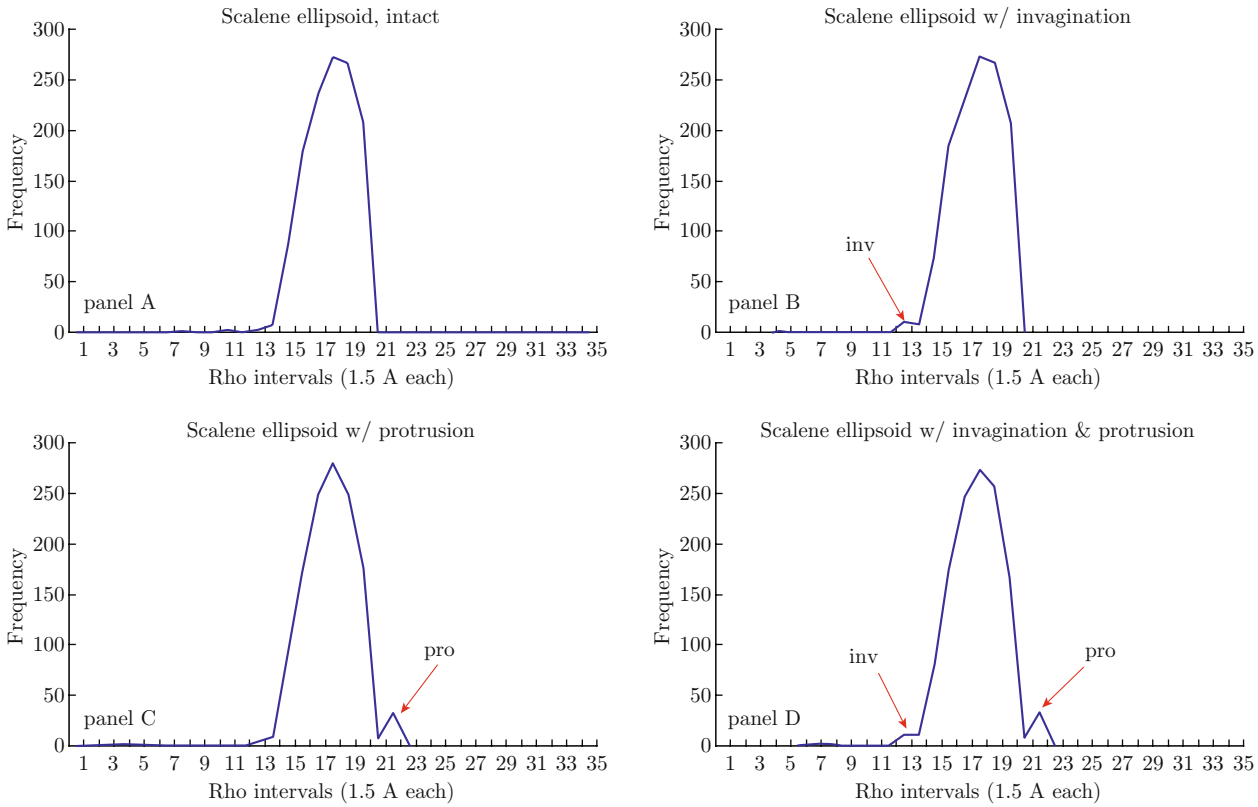



Fig. 7 FDMR plots of the artificial protein and its variants. The frequency distribution of maximum rhos are plotted for the intact artificial protein (panel A), its variant with an invagination (panel B), its variant with a protrusion (panel C), and a variant containing both an invagination and a protrusion (panel D). A subpeak or "shoulder" on the lagging side of the main peak is seen corresponding to the invagination, and a subpeak on the leading side is seen as corresponding to the protrusion.

The plot for the second variant with the protrusion in panel C, on the other hand, clearly shows a subpeak on the leading side of the main peak. A subpeak on the leading side of the main peak in the FDMR plot is therefore diagnostic of a protrusion, especially if the protrusion is along the longest dimension of the ellipsoid. If the protrusion does not lie along the longest dimension, it may be masked in the plot by some points along the longest dimension, and a subpeak on the lagging side of the main peak will be difficult to see. In that case, the protrusion will manifest itself as a non-coincidence between the superimposed FDMR plots of the intact ellipsoid and that of the protrusion-containing variant elsewhere along the plot.

Finally, the plot for the third variant with both the invagination and protrusion in panel D shows two subpeaks, one on the lagging and another on the leading sides of the main peak of the FDMR plot. These results are expected from the results described in the previous two paragraphs.

From the above results it may be concluded that an invagination on or close to the shortest dimension of the ellipsoid will result in a clear subpeak on the lagging side of the main peak FDMR plot, while a protrusion on or close to the longest dimension of the ellipsoid will result in a clear subpeak in the leading side of the main peak in the FDMR plot. In the general case where invaginations and protrusions lie on random locations on the ellipsoid surface, the superimposed FDMR plots of the smooth ellipsoid and those of its protrusion- or invagination-containing variants will display segments of non-coincidences on specific corresponding parts of the plots.

### 3.2.2 Application to real proteins in the data set of Laskowski

Due to space limitations, we can only show four sets of data here corresponding to four proteins in the data set of Laskowski *et al.* (1996) (Table 1). In these results, the FDMR plots of the liganded protein are superimposed with its unliganded form (*i.e.*, ligands deleted from the PDB file before algorithm implementation).

Fig. 8A shows the superimposed FDMR plots for

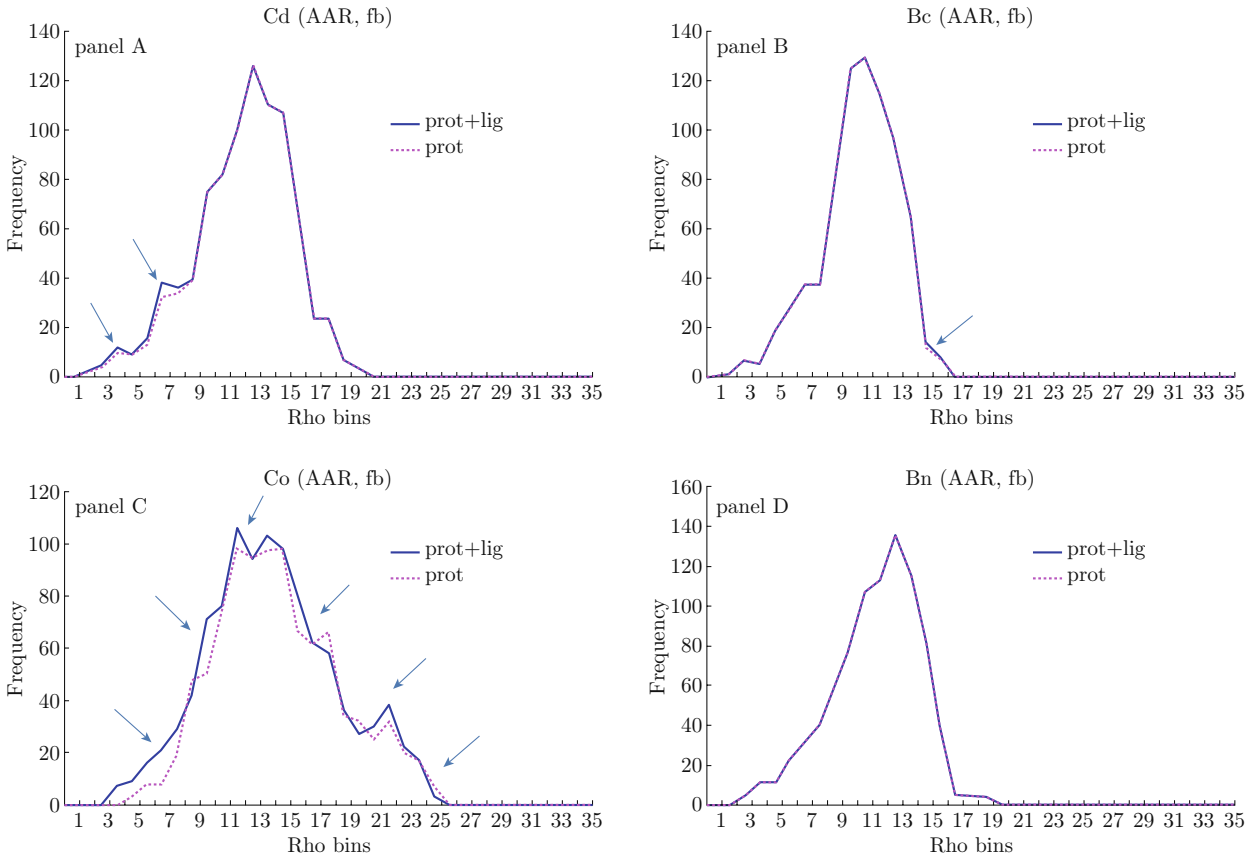



Fig. 8 FDMR plots of a sample of four proteins in the Laskowski dataset. The frequency distribution of maximum rhos is plotted for four cases, and the plots corresponding to liganded and unliganded forms were superimposed. In the first case (panel A), non-coincidence of the superimposed plots are seen in the lagging side of the main peak, while in the second case (panel B), non-coincidence of the superimposed plots are seen in the leading side of the main peak. In the third case (panel C), general non-coincidence all throughout the plot is evident, while in the fourth case, an almost perfect coincidence is seen.

proteins Cd (1CIL). A subpeak appears on the lagging side of the main peak of the unliganded relative to the liganded form. Fig. 8B shows the plots for protein Bc (1XNB). A subpeak appears on the leading side of the main peak of the unliganded relative to the liganded form. We interpret these results as local changes in the protein structure, giving rise to invaginations and protrusions, respectively, upon ligand binding. Structure 1CIL is that of human carbonic anhydrase II (Smith *et al.*, 1994), while 1XNB is that of a bacillus xylanase with a bound sulfate ion (Campbell *et al.*, 2011).

Fig. 8C shows the plots for protein Co (4BCL). We notice a general widespread noncoincidence between the two superimposed plots. In Fig. 8D for protein Bg (1BLL), we observe a clean coincidence of the two superimposed plots. Protein 4BCL is that of FMO antennae protein from green sulfur bacteria (Tronrud *et al.*, 2009), while 1BLL is that of bovine lens leucine

aminopeptidase (Kim and Lipscomb, 1993). We are continuing to analyze the precise structural meanings of these last two cases as we carefully refine our methods for spherical coordinate protein structure representation. A more comprehensive analysis and refinement of this application will be published in a future submission.

## 4 Conclusion and future direction

We have utilized the spherical coordinate system as an alternative to the Cartesian system to represent protein 3D structures. Using this representation, we have developed a way to separate the protein outer layer (OL) from the protein inner core (IC). Being able to separate the OL from the IC allowed us to investigate surface properties (from the OL) and buried properties (from the IC) of the protein systematically and independently. For example, we were able to predict potential epitopes and protein-protein interaction sites from the OL, and buried functional sites such as catalytic residues in the IC. We are convinced that our separation method works properly and is valid because when applied to a test set of 67 protein structures in Laskowski *et al.* (1996), we found that all but a few have OLs that are significantly enriched with hydrophilic amino acids and ICs that are significantly enriched with hydrophobic amino acids. To the best of our knowledge, this is the first time that spherical coordinate representation has been utilized in protein 3D structural analysis. Future directions for this project will include use of cylindrical coordinates to represent rod-shaped proteins and viruses and its applications.

Two web servers, namely, the "ProtMedCor Web Server" and "ProtSurfTop Web Server" are being set up to implement the OL-IC separation algorithm and the protrusion/invagination detection algorithm, respectively, for public access (MacCreary, M. and Kim, D.J., personal commun.).

**Acknowledgements** The author thanks Kyle Dewey of R.I.T. Bioinformatics Division for computing and web assistance, and R.I.T. bioinformatics M. S. students Mark MacCreary and Dong Jin Kim for their unpublished results. The first part of this project was conceived and undertaken at the University of California-San Diego, La Jolla, CA, under an IRACDA fellowship to the author, funded by NIGMS/NIH grant number GM 68524. The author acknowledges the UCSD Biomedical Library, the UCSD Academic Computing Services and the San Diego Supercomputer Center and for the help and support of their staff.

## References

[1] Anderson, C.M., Stenkamp, R.E., Steitz, T.A. 1978. Sequencing a protein by x-ray crystallography. II. Refinement of yeast hexokinase B co-ordinates and sequence at 2.1 Å resolution. J Mol Biol 25, 15–33.

[2] Campbell, R.L., Rose, D.R., Wakarchuk, W.W., To, R.J., Sung, W., Yaguchi, M. 1994. High-resolution structures of xylanases from B. Circulans and T. Harzianum identify a new folding pattern and implications for the atomic basis of the catalysis. http://www.rcsb.org/pdb/explore/explore.do?structureId=1XNB.

[3] Fujinaga, M., Delbaere, L.T., Brayer, G.D., James, M.N. 1985. Refined structure of alpha-lytic protease at 1.7 Å resolution. Analysis of hydrogen bonding and solvent structure. J Mol Biol 184, 479–502.

[4] Gershoni, J.M., Roitburd-Berman, A., Siman-Tov, D.D., Tarnovitski, N., Freund N., Weiss, Y. 2007. Epitope mapping: The first step in developing epitope-based vaccines. BioDrugs 21, 145–156.

[5] Kim, H., Lipscomb, W.N. 1993. X-ray crystallographic determination of the structure of bovine lens leucine aminopeptidase complexed with amastatin: Formulation of a catalytic mechanism featuring a gem-diolate transition state. Biochemistry 32, 8465–8478.

[6] Landro, J.A., Gerlt, J.A., Kozarich, J.W., Koo, C.W., Shah, V.J., Kenyon, G.L., Neidhart, D.J., Fujita, S., Petsko, G.A. 1994. The role of lysine 166 in the mechanism of mandelate racemase from Pseudomonas putida: Mechanistic and crystallographic evidence for stereospecific alkylation by (R)-alpha-phenylglycidate. Biochemistry 33, 635–643.

[7] Laskowski, R.A., Luscombe, N.M., Swindells, M.B., Thornton, J.M. 1996. Protein clefts in recognition and function. Prot Sci 5, 2438–2452.

[8] Lisgarten, J.N., Gupta, V., Maes, D., Wyns, L., Zegers, I., Palmer, R.A., Dealwis, C.G., Aguilar, C.F., Hemmings, A.M. 1993. Structure of the crystalline complex of cytidylic acid (2'-CMP) with ribonuclease at 1.6 Å resolution. Conservation of solvent sites in RNase-A high-resolution structures. Acta Crystallogr D Biol Crystallogr 49, 541–547.

[9] Loll, P.J., Lattman, E.E. 1989. The crystal structure of the ternary complex of staphylococcal nuclease, $Ca^{2+}$, and the inhibitor pdTp, refined at 1.65 Å. Proteins 5, 183–201.

[10] Lu, G., Lindqvist, Y., Schneider, G., Dwivedi, U., Campbell, W. 1995. Structural studies on corn nitrate reductase: Refined structure of the cytochrome b reductase fragment at 2.5 A, its ADP complex and an active-site mutant and modeling of the cytochrome b domain. J Mol Biol 248, 931–948.

[11] Martinez, C., Nicolas, A., van Tilbeurgh, H., Egloff, M.P., Cudrey, C., Verger, R., Cambillau, C. 1994. Cutinase, a lipolytic enzyme with a preformed oxyanion hole. Biochemistry 33, 83–89.

[12] Mosimann, S.C., Ardelt, W., James, M.N. 1994. Refined 1.7 Å X-ray crystallographic structure of P-30 protein, an amphibian ribonuclease with anti-tumor activity. J Mol Biol 236, 1141–1153.

[13] Mumey, B., Angel, T., Kirkpatrick, B., Bailey, B., Hargrave, P., Jesaitis, A., Dratz, E. 2003. Mapping discontinuous antibody epitopes to reveal protein structure and changes in structure related to function. IEEE Computer Society Bioinformatics Conference (CSB'03), Stanford, CA, USA, 585–586.

[14] Navia, M.A., McKeever, B.M., Springer, J.P., Lin, T.Y., Williams, H.R., Fluder, E.M., Dorn, C.P., Hoogsteen, K. 1989. Structure of human neutrophil elastase in complex with a peptide chloromethyl ketone inhibitor at 1.84-A resolution. Proc Natl Acad Sci USA 86, 7–11.

[15] Porter, C.T., Bartlett, G.J., Thornton, J.M. 2004. The Catalytic Site Atlas: A resource of catalytic sites and residues identified in enzymes using structural data. Nucl Acids Res 32, D129–D133.

[16] Reyes, V.M., Sheth V.N. 2011. Visualization of protein 3D structures in 'double-centroid' reduced representation: Application to ligand binding site modeling and screening. In: Liu, L.A., Wei, D., Li, Y. (Eds.) Handbook of Research on Computational and Systems Biology: Interdisciplinary Applications, Springer, Shanghai, 583–598.

[17] Sheth, V.N. 2009. Visualization of Protein 3D Structures in Reduced Representation with Simultaneous Display of Intra- and Intermolecular Interactions. Thesis, Master of Science in Bioinformatics, School of Biological and Medical Sciences, College of Science, Rochester Institute of Technology, Rochester , NY 14623-5603, USA.

[18] Smith, G.M., Alexander, R.S., Christianson, D.W., McKeever, B.M., Ponticello, G.S., Springer, J.P., Randall, W.C., Baldwin, J.J., Habecker, C.N. 1994. Positions of His-64 and a bound water in human carbonic anhydrase II upon binding three structurally related inhibitors. Protein Sci 3, 118–1125.

[19] Tarnovitski, N., Matthews, L.J., Sui, J., Gershoni, J.M., Marasco, W.A. 2006. Mapping a neutralizing epitope on the SARS coronavirus spike protein: Computational prediction based on affinity-selected peptides. J Mol Biol 359, 190–201.

[20] Thayer, M.M., Ahern, H., Xing, D., Cunningham, R.P., Tainer, J.A. 1995. Novel DNA binding motifs in the DNA repair enzyme endonuclease III crystal structure. EMBO J 14, 4108–4120.

[21] Trieschmann, M.D., Pattus, F., Tadros, M.H. 1996. Molecular characterization and organization of porin from Rhodobacter capsulatus strain 37B4. Gene 12, 61–68.

[22] Tronrud, D.E., Wen, J., Gay, L., Blankenship, R.E. 2009. The structural basis for the difference in absorbance spectra for the FMO antenna protein from various green sulfur bacteria. Photosynth. Res 100, 79–87.

[23] Vita, R., Zarebski, L., Greenbaum, J.A., Emami, H., Hoof, I., Salimi, N., Damle, R., Sette, A., Peters, B. 2010. The immune epitope database 2.0. Nucleic Acids Res 38(Database issue), D854–D862.

[24] Ward, M.R., Grimes, H.D., Huffaker, R.C. 1989. Latent nitrate reductase activity is associated with the plasma membrane of corn roots. Planta 177, 470–475.

[25] Weiss, M.S., Schulz, G.E. 1992. Structure of porin refined at 1.8 Å resolution. J Mol Biol 227, 493–509.

[26] Winn, S.I., Watson, H.C., Harkins, R.N., Fothergill, L.A. 1981. Structure and activity of phosphoglycerate mutase. Philos Trans R Soc Lond B Biol Sci 293, 121–130.

[27] Yang, W., Hendrickson, W.A., Crouch, R.J., Satow, Y. 1990. Structure of ribonuclease H phased at 2 Å resolution by MAD analysis of the selenomethionyl protein. Science 249, 1398–1405.