Running Head: LOCAL PROTEIN SURFACE CLASSIFICATION

Characterization and Classification of Local Protein Surfaces Using Self-Organizing Map

Lee Sael[1] and Daisuke Kihara[2,1,3*]

[1]Department of Computer Science, [2]Department of Biological Sciences, [3]Markey Center for Structural Biology, College of Science, Purdue University, West Lafayette, IN, 47907, USA.

[*]To whom correspondence should be addressed.

E-mail: dkihara@purdue.edu

Tel: 1-(765) 496-2284

Fax: 1-(765) 496-1189

Abstract

Annotating protein structures is an urgent task as increasing number of protein structures of unknown function is being solved. To achieve this goal, it is critical to establish computational methods for characterizing and classifying protein local structures. We analyzed the similarity of local surface patches from 609 representative proteins considering shape and the electrostatic potential, which are represented by the 3D Zernike descriptors. Classification of local patches is done with the emergent self-organizing map (ESOM). We mapped patches at ligand binding-sites to investigate how they distribute and cluster among the ESOM map. We obtained 30-50 clusters of local surfaces of different characteristics, which will be useful for annotating surface of proteins.

Keywords

protein surface, physicochemical properties, self-organizing map, clustering, 3D Zernike descriptors, spherical harmonics, affinity propagation clustering, local matching, shape analysis, protein active site.

Introduction

The importance and the impact of computational work in biology is ever increasing since large scale biological data of various types were accumulated in the past decade, which include genome sequences, protein-protein interaction networks, metabolomes, genome-scale transcriptions, and gene expression patterns. Those data contain key information for understanding orchestrated behavior of molecules in biological systems that is essential to sustain life. It is expected that bioinformatics play significant roles in analyzing such data, as computational techniques, *e.g.* clustering, feature characterization, data mining, and modeling, are indispensable in the analyses.

Of a particularly important and interesting problem is function prediction of proteins from the tertiary structures, as the structural genomics projects (Burley, 2000; Westbrook et al., 2003; Zhang & Kim, 2003) have been solving an increasing number of protein structures of unknown function. Indeed, as of May 2009, there are over 2800 proteins of unknown function in the Protein Structure Databank (PDB) (Berman et al., 2000). These proteins are remained as unknown function because so far no one has yet conducted experiments to characterize their function, and moreover, conventional sequence comparison based methods (Hawkins & Kihara, 2007), *e.g.* homology searches (Altschul et al., 1990; Altschul et al., 1997; Pearson & Lipman, 1988), functional motif (Hulo et al., 2006), and domain searches (Coggill et al., 2008), did not find significant similarity against protein sequences of known function. Ongoing efforts for better function prediction include development of sequence-based methods which are more sensitive and accurate than the conventional methods (Chitale et al., 2009a; Hawkins et al., 2008). For example, our group has recently developed two sequence-based methods, named the automated Protein Function Prediction (PFP) method (Hawkins et al., 2006; Hawkins et al., 2009) and the Extended Similarity Group (ESG) method (Chitale et al., 2009b), which efficiently and accurately mine function information from PSI-BLAST searches.

Alternatively, one can use the tertiary structure information for capturing similarity to proteins with known function that are stored in PDB (Thornton et al., 2000). Potential advantages of using structure information are two folds: Firstly, evolutionarily more distantly related proteins to a query protein could be identified because the global structure is better conserved than the primary sequence (Chothia & Lesk, 1986; Kihara & Skolnick, 2004). Secondly, physical features of functional local sites of proteins can be directly compared where interactions with ligand molecules or other proteins take place (Laskowski et al., 2005). A number of methods have been proposed which use local structure as a key feature for predicting function of proteins. Since small ligand molecules usually bind to a protein at its surface pocket regions, simply identifying pockets in the protein surface can identify active sites of enzyme in most of the cases (Li et al., 2007; Laskowski et al., 1996; Kawabata & Go, 2007). Programs which identify pockets include Visgrid (Li et al., 2007), POCKET (Levitt & Banaszak, 1992), LIGSITE (Hendlich et al., 1997; Huang & Schroeder, 2006), SURFNET (Laskowski, 1995), and PocketDepth (Kalidas & Chandra, 2008). An identified pockes can be further compared with known ligand binding pockets in a database to make prediction of the type of ligand that binds to it (Kahraman et al., 2007; Tseng et al., 2009; Kihara et al., 2009; Chikhi & Kihara, 2009; Binkowski & Joachimiak, 2008; Kalidas & Chandra, 2008; Yeturu & Chandra, 2008). In these methods, pockets are characterized by geometrical shapes, amino acid residues at pockets, and physicochemical properties such as the electrostatic potentials and hydrophobicity.

A fundamental limitation of these methods is that they only deal with pocket regions, *i.e.* prediction of ligand binding at pockets in protein surface. Although binding small ligand molecules is a very important class of protein function, it only occurs in a subset of proteins, mainly enzymes and proteins that use co-factors. Moreover, on average pocket regions in an enzyme share only about 5% of the entire surface of the protein (Li et al., 2007). Hence, many of proteins in a genome and most of surface region of proteins are left out from applicability of these methods. Ideally, methods should be more generalized so that they can describe and compare any protein surface regions and annotate local regions with its potential functions.

Toward this goal, here, we classified local surface regions of proteins to see how diverse local surface patches are and how many different types of surface patches exist. We characterized a surface patch by two features, geometric shape and the electrostatic potential. Among shape descriptors that have been studied (Sael & Kihara, 2009; Tangelder & Veltkamp, 2008), we have chosen the 3D Zernike descriptors, a projection-based approach, because they have already been shown to be very effective in describing the shape and other physicochemical properties of proteins (Kihara et al., 2009; Sael et al., 2008b; Sael et al., 2008a).

Classification of surface patches is done by the emergent self-organizing map (ESOM), which is a variant of self-organizing map with a large size of neurons. Application of the ESOM to 118,003 surface patches taken from 609 proteins yields 30-50 clusters of different characteristics. To demonstrate practical usefulness of this classification, we examined classes of local surface patches occurring at binding sites of six ligand molecules and found that the combination reflects chemical similarity of the ligands. The manageable size of the clusters will be convenient to label local patches of proteins, and thus will be useful for annotating surface regions of proteins and also for defining signatures of functional sites of proteins.

Materials and Methods

Figure 1 illustrates the procedure we conducted in this study. First, for a set of proteins, the surface is defined and the surface electrostatic potential is computed. Then, surface regions are segmented into local patches, using a sphere of a fixed size. Each local surface patches is characterized by the 3D Zernike descriptors of the surface shape and the electrostatic potential. Next, we use the ESOM to classify surface patches in the dataset. Three ESOM maps are constructed; one for classifying surface shape of patches, another one for classifying the electrostatic potential of patches, and the third one for classifying the combination of the surface shape and the electrostatic potential. The resulting ESOM maps are further clustered to identify distinct groups of patches with different characteristics. Also, we compared binding pockets of six different ligands in terms of types of the clusters occurring at the binding pockets. Below we describe each step more in details.

Protein Dataset

A data set of representative proteins used in this study consists of 609 protein chains, each taken from a different family defined in the SCOP database (Andreeva et al., 2008). These families belong to one among 72 SCOP superfamilies. We selected the 72 superfamilies, each of which has at least three and no more than twenty families. These structures have a crystallographic resolution of 3.0 Å or better, have no more than 10 missing residues in the structure solved, have all heavy atom positions solved, are longer than 100 residues, and the structure similarity of each pair is less than a Z-score of 3.8 by the Combinatorial Extension (CE) program (Shindyalov & Bourne, 1998). The protein classification database by CE available at ftp://ftp.sdsc.edu/pub/sdsc/biology/CE/db/ata_3_8.txt is used for comparing structures of the proteins.

The ligand binding protein data set we use later in this study is composed of proteins taken from a previous work by Kahraman *et al.* (Kahraman et al., 2007) (Table 1, Fig. 2). A binding site of a ligand is defined as a set of protein heavy atoms which are closer than 3.5Å to any ligand atoms. Conformation of ATP (adenosine-5'-triphosphate) and AMP (adenosine-5'-diphosphate) are similar in shape. FAD (flavin-adenine dinucleotide) and NAD (nicotinamide-adenine-dinucleotide) are relatively flexible and take diverse bound conformations to proteins. FMN (flavin mononucleotide) and HEM (haemoglobin) are more distinct in shapes compared to the others in the dataset.

Defining protein surface and computing electrostatic potentials

Two properties, the surface shape and the electrostatic potential are used to characterize surface of proteins. These are computed with the Adaptive Poisson-Boltzmann Solver (APBS) program (Baker et al., 2001). APBS is software for evaluating the electrostatic properties of biomolecules by solving the Poisson-Boltzmann equation. APBS first defines the solvent accessible regions of the protein and calculates the electrostatics potential for each of the voxels (3D grid points). Surface region of a protein is computed by taking the boundary of the solvent accessible and the solvent excluded region using APBS. The electrostatic potential is mapped to the computed protein surface.

Segmenting protein surface to local patches

After obtaining the global shape and the electrostatic potential of a protein surface, the surface is segmented into local patches. First, seed points, or the centroids of the local surfaces, are selected by taking every 100th surface voxel. The average number of the seed points is approximately 194 per protein. The local patches are then extracted by taking a surface region that is within 6Å of each seed point. The resulting voxels are considered as an input 3D function, *f(x)*, which is expanded into the 3D Zernike Descriptors (3DZD) as described in the next section. 118,003 patches are obtained from the 609 proteins in the dataset.

3D Zernike Descriptors

The 3DZD are series expansion of an input 3D function, which allow rotation invariant and compact representation of a 3D object that is considered as the 3D function. The mathematical foundation of the 3DZD was laid by Canterakis (Canterakis, 1999) and later Novotni and Klein (Novotni & Klein, 2003) have applied them to 3D object retrieval. Below is brief mathematical derivation of the 3DZD. For detailed derivations and discussions, refer to the two papers (Canterakis, 1999; Novotni & Klein, 2003).

The first step in computing the 3DZD is derivation of the 3D Zernike moments. The 3D Zernike moments are series expansion of an input 3D function, *f(x)*, into the 3D Zernike polynomial. The input function, *f(x)*, in this work is a 3D grid where each grid cell (voxel) is assigned 1 if it is on the protein surface and 0 otherwise to represent a protein shape. In the case of the surface electrostatic potential, surface voxels are assigned with the electrostatic potential value of that position instead of 1. For the electrostatic potential, the voxels are classified into two groups, ones which have positive values and the other ones for negative values. The 3DZD are computed separately for the two groups of voxels and later combined (Sael et al., 2008a).

3D Zernike polynomials defined on degree *l*, order *n*, and repetition *m*, are given by

$$Z_{nl}^{m}(r,\vartheta,\varphi) = R_{nl}(r)Y_{l}^{m}(\vartheta,\varphi)\, , \qquad (1)$$

where $-l < m < l,\ 0 \le l \le n$, and *(n-l)* even. Spherical harmonics, $Y_l^m(\vartheta,\varphi)$, are functions of a set of a polar angle, *θ*, and an azimuthal angle, φ. They are the angular portions of the solution to the Laplace's equations. The radial function defined by Canterakis(Canterakis, 1999), $R_{nl}(r)$, directly incorporates radius information into the basis function and are constructed so that $Z_{nl}^m(r,\vartheta,\varphi)$ are polynomials when written in the Cartesian coordinates **x** = (x, y, z). Using all three spherical coordinate parameters (radius: *r*, polar angle: ϑ, and zimuthal angle: φ), the 3D Zernike polynomials are able to be formulated against any Cartesian coordinates.

The 3D Zernike moments of *f(x)* are defined as the coefficients of the expansion in this orthonormal basis:

$$\Omega_{nl}^m = \frac{3}{4\pi}\int_{|\mathbf{x}|\le 1} f(\mathbf{x})\bar{Z}_{nl}^m(\mathbf{x})d\mathbf{x}\,. \qquad (2)$$

Finally, rotation invariance is obtained by taking L_2 norm of 3D Zernike moments as the descriptor.

The parameter *n* is called the order of the 3DZD. The order determines the resolution, *i.e.* the number of terms, thus coefficients, of the descriptors. As mentioned earlier, *n* defines the range of *l*, and 3DZD are series of invariants for each pair of *n* and *l*, where n ranges from 0 to the order specified. In this work we set the order n=15, which yields 72 values for shape and 144 values for the electrostatic potential, to describe the local patches of proteins. The 3DZD of the electrostatic potential is twice longer because those for voxels with positive values and those with negative values are separately computed and later combined (Sael et al., 2008a). Since the ESOM assumes that data have a distribution that is close to the Gaussian distribution, we further performed log normalization to the resulting descriptors.

SOM and ESOM

The self-organizing map (SOM) is a data classification technique which reduces the dimensions of data through the use of self-organizing neural networks (Rojas & Feldman, 1996). The output of the SOM is a lower dimension map, usually two, of neurons which group similar data items together on the map. Each neuron on the map has a vector of weights, which are trained based on a set of input training data. The weights assigned to each neuron are composed of two parts. The first part represents the feature of the neuron, which has the same dimension as the input vectors. The second part represents its location in the 2D map. SOM tries to order the neurons such that they represent the distribution of the similarity of the input vectors. The SOM algorithm first initializes the weight vector map. Among several methods available for the initialization, we used sampling from the Gaussian distribution. Then, the feature weights of neurons and their location in the 2D map are updated iteratively. The process consists of three steps: selecting an input feature vector, finding the neuron that has a closest feature weights to the input, updating the weights of the closest neuron and its neighboring neurons, and updating the position of neighboring neurons so that similar neurons locate closer to each other (Rojas & Feldman, 1996).

Emergent SOM (ESOM) is a variation of SOM, which handles a larger number of neurons (at least 4000) and uses boundless maps (Ultsch, 2005; Ultsch, 2003). It embeds the maps to a finite boundless space such as sphere or toroid. Two visualization methods of the ESOM maps are used, namely, the P-matrix and the U-matrix. The P-matrix visualizes the density in the input data space using the Pareto density estimation. In general, it is suitable for dealing with slowly changing densities and overlapping clusters. The U-matrix visualizes neurons on an ESOM map by a color coding that represents the sum of distances to all immediate neighbors normalized by the largest value in the neighboring neurons. Generally, the

U-matrix is appropriate for handling data points which are clearly separated from each other. The ESOM program is available at http://databionic-esom.sourceforge.net/. The original paper by Ultsch [36] describes the general ESOM training procedure in details.

The advantage of SOM/ESOM is that it is able to provide an intuitive visualization of the similarity of input data. On the other hand, a potential drawback is that often SOM/ESOMs trained on the same data set do not converge well to a similar map. In our work, to examine the convergence of the ESOM map we obtained, we further clustered neurons in the resulting map by using the affinity propagation clustering method (see the next section). As will be shown in Results, in our case we think that the ESOM performed satisfactorily for the purpose of the classification of protein surface patches, because neurons locating close to each other are well clustered.

The training data in this work are feature vectors of the 118,003 local surface patches in the 609 proteins. A feature vector used for local surface patches is the 3DZD describing the local shape, the electrostatic potential, or the combination of the two. As explained in the previous section, the 3DZD of the local shape have 72 values, those of the electrostatic potential have 144 values, and the combination of the shape and the electrostatic potential have 216 (*i.e.* 72+144) values. The training data set is inputted to ESOM program and trained using 1000 iterations and 4100 neurons.

Affinity propagation clustering method

The affinity propagation clustering method clusters data by employing an idea of passing messages between them (Frey & Dueck, 2007). It was shown to have a low error rate and fast as compared to other common clustering methods. The number of clusters is influenced by the so-called preference parameter. Setting the preference parameter to the median of input distances results in a moderate number of clusters and setting them to the minimum of input distances results in a smaller number of clusters.

Results

Reconstruction of local surface shape from the 3D Zernike descriptors

To begin with, we examine how well the 3DZD capture the shape of local surface patches by reconstructing the original shape back from the 3DZD. In Figure 3, three local patches of protein 1kvkA (PDB ID) are reconstructed using four different resolution levels, the order n of 5, 10, 15, and 20. As seen in the first row of Figure 3, the three local patches are significantly different in shape: The local patch 1 has a wrap-like shape, the local patch 2 has a pocket shape, and the local patch 3 is saddle-like. The 3DZD capture this difference well as shown in the clear distinction in the invariant in Figure 4. Figure 4 shows invariant values of the order n=20 (121 invariants). An invariant of a smaller order is a subset of a higher order invariant. For example, an invariant that corresponds to the order n=15 is the first 72 values of the 121 invariants. Figure 3 shows that the reconstruction results of the 3DZD with the order n=15 is as good as those of n=20. In our previous works, we used the order of 20 for describing the global protein surface shape (Sael et al., 2008b; Sael et al., 2008a). On the other hand, the order of 15 would be sufficient for local surface patches because the local surface is less complex than that of global surface. Thus, we use the order n = 15 in what follows.

ESOM maps

118,003 local patches from the 609 proteins represented by the 3DZD are analyzed with the ESOM with 4100 neurons. The resulting ESOM maps are boundless toroid, *i.e.* the right edge of the map is continuously connected to the left edge and the upper edge is connected to the lower edge. Three types of input feature vectors are used to characterize the surface patches: the 3DZD of 1) the surface shape, 2) the electrostatic potential, and 3) a combination of the surface shape and the electrostatic potential (Fig. 5). The maps are shown in two color-coding schemes, the P-matrix and the U-matrix. Note that these two matrices visualize the same data, although their appearance is different. Five examples of the local patches that corresponds to the labeled positions in each ESOM map are also shown in Figure 5.

The U-matrix maps show the mutual distance of neighboring neurons. In the U-matrix maps of all three types of input feature vctors, a large blue area and a couple of smaller green areas are observed. The blue area indicates that there are a large region of neurons that have a small distance to each other. These regions are mostly composed of patches of more or less flat shape. Pocket-like shape patches locate at the regions that have a higher distance (green regions). This is reasonable because it indicates that pocket shapes have more diversity in shape compared to flat regions.

The P-matrix maps show the density of the data set that are assigned to each neuron and are colored from dark to light color as the density decreases. Compared to the U-marix, more distinct clusters are identified. Local patches of dense regions on the P-matrix (dark colored regions) are relatively flat and have mixture of the positive and negative electrostatic potential that are near 0.0. In the maps for the electrostatic potential (the middle row), clear separation of the patches is not observed except for some cases (*e.g.* a cluster with a patch from 1aquB at (e) in the U-matrix of the electrostatic potential, which is strongly positively charged). This is because the 3DZD of the electrostatic potential intrisically convolute information of the electostatic potential and the shape, as the potential is mapped on the surface. The convolution of the two information was also observed in our previous work (Sael et al., 2008a).

Figure 5 shows the overall landscape of the similarity of the surface patches. To obtain more distinct clusters, we clustered the 4100 neurons of the ESOM maps using the affinity propagation clustering method (Fig. 6). We used two ESOM maps, that of the surface shape (Fig. 6A & B) and another map of the combination of the shape and the electrostatic potential (Fig. 6C & D). Setting the preference parameter to the median of input distances results in 369 clusters for the surface ESOM map (Fig. 6A) and 215 clusters for the surface and the electrostatics combination (Fig. 6C). When the minimum of the input distances is used as the parameter, the number of clusters is reduced to 48 in the surface ESOM map (Fig. 6B) and 27 for the surface and the electrostatics combination (Fig. 6D). The small number of resulting clusters of neurons indicates that the neighboring neurons have similar weights, *i.e.* the ESOM was well trained. These distinct clusters are convenient for labeling a protein surface. As each local patch in the given surface can be assigned to one of the clusters, the surface can be represented by a set of the indices of clusters to which local patches of the surface belong.

Local patch types of ligand binding surfaces

Next, using the distinct clusters of surface patches (Fig. 6), we classified local patches of the ligand binding sites into the clusters. Figure 7 shows histograms of surface patch clusters observed at ligand binding surfaces of the proteins listed in Table 1. The cluster compositions of all the ligand binding surfaces share some similarities that comes from the fact the binding sites have many pocket-like local shapes. In all the binding sites, the surface patch cluster 16 and 34 and the surface electrostatic potential cluster 6 and 22 are observed. The molecular shape and the binding sites of FMN are more different from the others, which are reflected to the relatively distinct histogram.

Figure 8 shows concrete examples of local patches of ligand binding sites. They are mapped on the ESOM map of the combination of the shape and the electrostatic potential. The ATP binding site of 1dv2A consists of surface patch clusters 6(c), 13, 16(d), 21(b), and 24(a). In the parentheses, labels shown in Figure 8 are indicated. The AMP binding site of 1c0aA is composed of the clusters 6(a), 12(c), and 23(b). The FMN binding site of 1mvlA is composed of the clusters 1(a), 5(c), 10, and 21(b). The heme binding site of 1np4A consists of the clusters 2(c), 6(d), 7(b), 10, and 17(a). The FAD binding site of 1hskA consists of the clusters 6(c), 14(d), 16(b), 20, 23(a), 24, and 27. The NAD binding site of 2npxA is composed of the clusters 1, 2, 7(b), 12(a), 16(d), 22(c), and 23. The patches found at ligand binding sites of the other proteins listed in Table 1 are also included in Figure 8 with the colored dots. It turned out that binding sites of similar ligands do not always have the same set of surface patches. For example, the adenosine binding site is assigned to the patch 24 in 1dv2A, an ATP binding protein, and to the patch 12 in 1c0aA, an AMP binding protein. This is partly due to the conformational changes of the ligands (Kahraman et al., 2007) and also due to the different distribution of the seed points on a protein surface. However, of course there are cases where binding sites of the same chemical group of ligands are assigned to the same surface patch cluster, such as the phosphate binding region in 1dv2A and 1c0aA, both of which are assigned to the cluster 6. This overall similarity in consisting surface patch clusters can also be found in Figure 7, where both ATP and AMP have high peaks at the clusters 6, 7, 23, and 24 in their histograms.

Computation Time

The time taken to generate the descriptors for each protein depends on the size of the proteins. It takes around two minutes to generate the surface geometry and the surface electrostatic information for each protein and on an average of 37.12 seconds to compute local 3D Zernike descriptors for patches in a whole protein. The training of an ESOM using the surface patch data with 1000 iterations took approximately two weeks. However, note that the training is needed only once. Once the ESOM is trained, assignment of patches in a protein to neurons can be done in less than a second. The analysis was performed on Linux machine equipped with Intel Core TM2 CPU 6400 at 2.13GHz.

Discussion

We have classified local surface patches of proteins, which are characterized by the shape and the electrostatic potential. The use of the 3DZD provided a convenient compact representation of the two characteristics of the patches as feature vectors. We used the ESOM for classifying the local patches since it is an effective method for observing landscape of similarity of data points, whose similarity is rather continuous but not very distinct. Moreover, in order to obtain distinct groups of local surface patches, we have clustered neurons of the ESOM map trained with the local patches. The procedure yielded a manageable size of clusters of local patches; 48 clusters for the surface shape, and 27 for the combination of the shape and the electrostatic potentials.

The resulting clusters have several interesting applications. The clusters can be used as surface “alphabet”, with which protein surface can be labeled. For example, the previous example of the AMP binding site of 1dv2A is described as a set of surface classes (6, 13, 16, 21, 24). This description of protein surface with a set of letters (numbers) would enable a variety of protein surface analyses, such as classification, function prediction, and database searches, in analogous ways to protein sequence analyses. In particular, a strong advantage of the surface alphabet is that it allows partial matching of protein surfaces, which may be necessary for considering flexibility of protein structure and a certain degree of difference in features in protein surface.

The idea of the surface alphabet proposed in this paper certainly needs more investigations and testing. For example, different parameters and alternative methods need to be tested, such as the definition of the ligand binding sites, the way to select the seed points in a protein surface, and methods for clustering similar surface patches. We are also planning to incorporate different features for describing surface patches, including sequence conservation and hydrophobicity. To conclude, as annotation of protein structures has become an urgent task in bioinformatics, we believe that the idea of the surface alphabet will pave the way for developing many approaches for handling protein global and local surfaces.

Acknowledgment

This work is supported by the the National Institutes of Health (GM075004 and GM077905) and National Science Foundation (DMS0604776, DMS800568).

Reference List

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol, 215,* 403-410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res, 25,* 3389-3402.

Andreeva, A., Howorth, D., Chandonia, J. M., Brenner, S. E., Hubbard, T. J., Chothia, C. et al. (2008). Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res, 36,* D419-D425.

Baker, N. A., Sept, D., Joseph, S., Holst, M. J., & McCammon, J. A. (2001). Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc.Natl.Acad.Sci U.S.A, 98,* 10037-10041.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. et al. (2000). The Protein Data Bank. *Nucleic Acids Res, 28,* 235-242.

Binkowski, T. A. & Joachimiak, A. (2008). Protein functional surfaces: global shape matching and local spatial alignments of ligand binding sites. *BMC.Struct.Biol., 8,* 45.

Burley, S. K. (2000). An overview of structural genomics. *Nat Struct Biol, 7 Suppl,* 932-4.

Canterakis, N. (1999). 3D Zernike moments and Zernike affine invariants for 3D image analysis and recognition. *Proc.11th Scandinavian Conference on Image Analysis,* 85-93.

Chikhi, R. & Kihara, D. (2009). Protein binidng ligand prediction using local surface descriptors. *Submitted.*.

Chitale, M., Hawkins, T., & Kihara, D. (2009a). Automated prediction of protein function from sequence. In J.Bujnicki (Ed.), *Prediction of Protein Strucutre, Functions, and Interactions* (pp. 63-86). John Wiley & Sons Ltd.

Chitale, M., Hawkins, T., Park, C., & Kihara, D. (2009b). ESG: Extended similarity group method for automated protein function prediction. *Bioinformatics, 25,* 1739-1745.

Chothia, C. & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J., 5,* 823-826.

Coggill, P., Finn, R. D., & Bateman, A. (2008). Identifying protein domains with the Pfam database. *Curr.Protoc.Bioinformatics, Chapter 2,* Unit.

Frey, B. J. & Dueck, D. (2007). Clustering by passing messages between data points. *Science, 315,* 972-976.

Hawkins, T., Chitale, M., & Kihara, D. (2008). New paradigm in protein function prediction for large scale omics analysis. *Mol.Biosyst., 4,* 223-231.

Hawkins, T., Chitale, M., Luban, S., & Kihara, D. (2009). PFP: Automated prediction of gene ontology functional annotations with confidence scores using protein sequence data. *Proteins, 74,* 566-582.

Hawkins, T. & Kihara, D. (2007). Function prediction of uncharacterized proteins. *J.Bioinform.Comput.Biol., 5,* 1-30.

Hawkins, T., Luban, S., & Kihara, D. (2006). Enhanced automated function prediction using distantly related sequences and contextual association by PFP. *Protein Sci, 15,* 1550-1556.

Hendlich, M., Rippmann, F., & Barnickel, G. (1997). LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J.Mol.Graph.Model., 15,* 359-63, 389.

Huang, B. & Schroeder, M. (2006). LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC.Struct.Biol., 6,* 19.

Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., De, C. E., Langendijk-Genevaux, P. S. et al. (2006). The PROSITE database. *Nucleic Acids Res., 34,* D227-D230.

Kahraman, A., Morris, R. J., Laskowski, R. A., & Thornton, J. M. (2007). Shape variation in protein binding pockets and their ligands. *J.Mol.Biol., 368,* 283-301.

Kalidas, Y. & Chandra, N. (2008). PocketDepth: a new depth based algorithm for identification of ligand binding sites in proteins. *J.Struct.Biol., 161,* 31-42.

Kawabata, T. & Go, N. (2007). Detection of pockets on protein surfaces using small and large probe spheres to find putative ligand binding sites. *Proteins, 68,* 516-529.

Kihara, D., Sael, L., & Chikhi, R. (2009). Local surface shape-based protein function prediction using Zernike descriptors. *Biophys J, 96,* 650a.

Kihara, D. & Skolnick, J. (2004). Microbial Genomes have over 72% structure assignment by the threading algorithm PROSPECTOR_Q. *Proteins, 55,* 464-473.

Laskowski, R. A. (1995). SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J.Mol.Graph., 13,* 323-328.

Laskowski, R. A., Luscombe, N. M., Swindells, M. B., & Thornton, J. M. (1996). Protein clefts in molecular recognition and function. *Protein Sci., 5,* 2438-2452.

Laskowski, R. A., Watson, J. D., & Thornton, J. M. (2005). Protein function prediction using local 3D templates. *J.Mol.Biol., 351,* 614-626.

Levitt, D. G. & Banaszak, L. J. (1992). POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *J.Mol.Graph., 10,* 229-234.

Li, B., Turuvekere, S., Agrawal, M., La, D., Ramani, K., & Kihara, D. (2007). Characterization of local geometry of protein surfaces with the visibility criterion. *Proteins, 71,* 670-683.

Novotni, M. & Klein, R. (2003). 3D Zernike descriptors for content based shape retrieval. *ACM Symposium on Solid and Physical Modeling, Proceedings of the eighth ACM symposium on Solid modeling and applications,* 216-225.

Pearson, W. R. & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A, 85,* 2444-2448.

Rojas, R. & Feldman, J. (1996). *Neural networks - A systematic introduction*. Berlin, New York: Springer-Verlag.

Sael, L. & Kihara, D. (2009). Protein surface representation and comparison: New approaches in structural proteomics. In J.Chen & S. Lonardi (Eds.), *Biological Data Mining* (Boca Raton, Florida, USA: Chapman & Hall/CRC Press.

Sael, L., La, D., Li, B., Rustamov, R., & Kihara, D. (2008a). Rapid comparison of properties on protein surface. *Proteins, 73,* 1-10.

Sael, L., Li, B., La, D., Fang, Y., Ramani, K., Rustamov, R. et al. (2008b). Fast protein tertiary structure retrieval based on global surface shape similarity. *Proteins, 72,* 1259-1273.

Shindyalov, I. N. & Bourne, P. E. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng, 11,* 739-47.

Tangelder, J. H. & Veltkamp, R. C. (2008). A survey of content based 3D shape retrieval methods. *Multimedia Tools and Applications, 39,* 441-471.

Thornton, J. M., Todd, A. E., Milburn, D., Borkakoti, N., & Orengo, C. A. (2000). From structure to function: approaches and limitations. *Nat.Struct.Biol., 7 Suppl,* 991-994.

Tseng, Y. Y., Dundas, J., & Liang, J. (2009). Predicting Protein Function and Binding Profile via Matching of Local Evolutionary and Geometric Surface Patterns. *J.Mol.Biol.*.

Ultsch, A. (2003). Maps for the visualization of high-dimensional data spaces. *Proc.WSOM'03,* 225-230.

Ultsch, A. (2005). ESOM-Maps: tools for clustering, visualization, and classification with Emergent SOM. *Technical report.Dept.of Mathematics and Computer Sciece, Univeristy of Marburg..*

Westbrook, J., Feng, Z., Chen, L., Yang, H., & Berman, H. M. (2003). The Protein Data Bank and structural genomics. *Nucleic Acids Res, 31,* 489-491.

Yeturu, K. & Chandra, N. (2008). PocketMatch: a new algorithm to compare binding sites in protein structures. *BMC.Bioinformatics, 9,* 543.

Zhang, C. & Kim, S. H. (2003). Overview of structural genomics: from structure to function. *Curr.Opin.Chem.Biol., 7,* 28-32.

Table 1.

Ligands and ligand binding proteins used in this study

Binding ligand	Number of proteins	Ligand binding proteins (Pdb ID and chain ID)
ATP	14	1a0i- 1a49A 1aylA 1b8aA 1dv2A 1dy3A 1e2qA 1e8xA 1esqA 1gn8B 1kvkA 1o9tA 1rdqE 1tidA
AMP	9	12asA 1amuA 1c0aA 1ct9A 1jp4A 1khtB 1qb8A 1tb7B 8gpbA
FMN	6	1dnlA 1f5vA 1ja1A 1mvlA 1p4cA 1p4mA
HEM	16	1d0cA 1d7cA 1dk0A 1eqgA 1ew0A 1gweA 1iqcA 1nazA 1np4A 1po5A 1pp9C 1qhuA 1qlaC 1qpaB 1soxA 2cpo-
FAD	10	1cqxA 1e8gB 1eviB 1h69A 1hskA 1jqiA 1jr8B 1k87A 1poxA

Figure Captions

Figure 1. Flowchart of the methods.

Figure 2. 3D structures of the six ligands used in this study. ATP, AMP, FMN, heme, FAD, and NAD.

Figure 3. Local surface shape reconstruction from the 3D Zernike descriptors. Three local patches from protein 1kvkA are reconstructed using four different resolutions, the order of 5, 10, 15, and 20.

Figure 4. Invariants of local 3D Zernike descriptors of the three local patches of protein 1kvkA. The x axis is the indices of the invariants and the y axis is the value of each invariants.

Figure 5. ESOM maps of protein local patches. Three types of feature vectors of the patches are used as input data: the 3DZD of surface shape, those of the electrostatic potential, and the combination of the two. Resulting ESOM maps are visualized in two ways, the P-matrix and the U-matrix. The color code of the P-matrix shows the density of the assigned input data to neurons, while that of the U-matrix indicates the mutual distance (dissimilarity) of neighboring neurons. For each ESOM map, five examples of the local patches are sampled at the labeled positions. In the surface ESOM map, local patches from 1a2kC (a, b, c), 1b7eA (d), and 1af6C (e) are sampled. In the ESOM map of the electrostatic potential, patches from 1a2kA (a, b, c) and 1aquB (e, f) are shown. In the ESOM map of the combination of the shape and the electrostatic potential, local patches from 1b9mB (a), 1aurB (b, c), and 1a8uB (d, e) are shown.

Figure 6. Affinity propagation clustering of ESOM neurons. A and B are clustering results of ESOM neurons using the surface shape, while C and D are results using the surface shape and

the electrostatic potential combination as the feature of local surface patches. In A and C, the median of distance is used as the preference parameter p. B and D use the minimum distance as the preference parameter. There are 369 clusters in A, 48 clusters in B, 215 clusters in C, and 27 clusters in D. Note that the colors are used just to separate neighboring clusters; clusters in the same color at different locations do not mean similarity.

*Figure 7***.** Histograms of ESOM neuron clusters to which surface patches of binding sites of the six ligand molecules (Table 1) belong. The clusters are obtained by the affinity propagation clustering method using the minimum distance as the preference parameter, *i.e.* the clusters for the surface shape is shown in Fig 6B and those for when the combination of the shape and the electrostatic potential are shown in Figure 6D. The x-axis is the index of the cluster and the y-axis is the density of each cluster.

Figure 8. Distribution of ligand binding surface on the ESOM map. The ESOM map of the combination of the shape and the electrostatic potential is used (i.e. Fig. 5, bottom row). The colored dots specify the most similar neurons to the ligand binding local patches. The color of the dots indicates the source protein. For each type of ligand, one example of the binding sites is shown on the right side of each ESOM map. The proteins used are as follows: 1dv2A for ATP binding; 1c0aA for AMP; 1mvlA for FMN; 1np4A for HEM; 1hskA for FAD; and 2npxA for NAD. The labels (a, b, c, d, e) indicate the position of the corresponding local patches on the ESOM map.

Figure 1

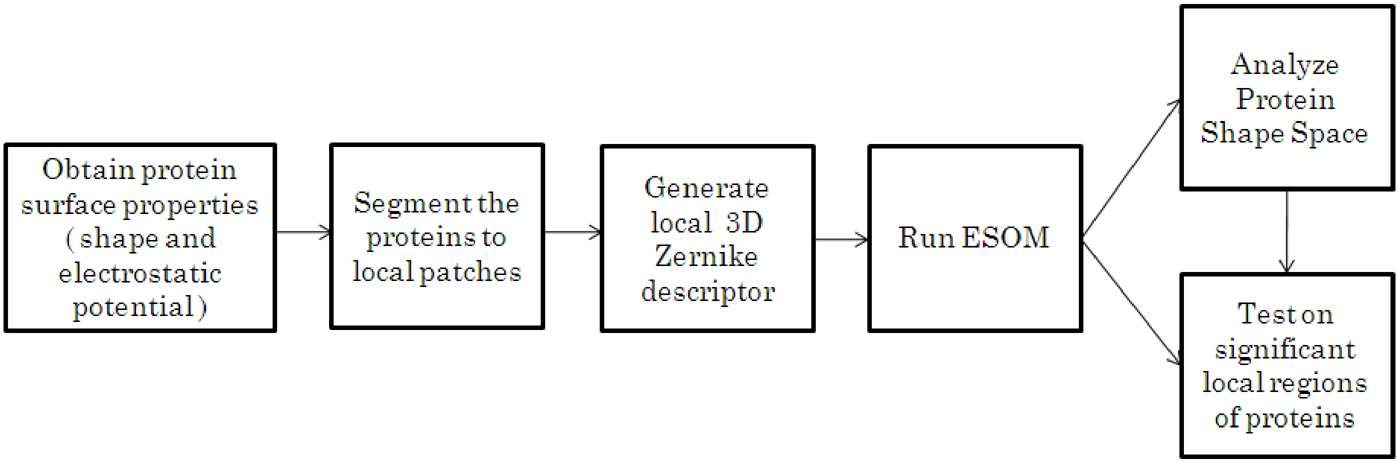

Obtain protein surface properties (shape and electrostatic potential)
Segment the proteins to local patches
Generate local 3D Zernike descriptor
Run ESOM
Analyze Protein Shape Space
Test on significant local regions of proteins

Figure 2

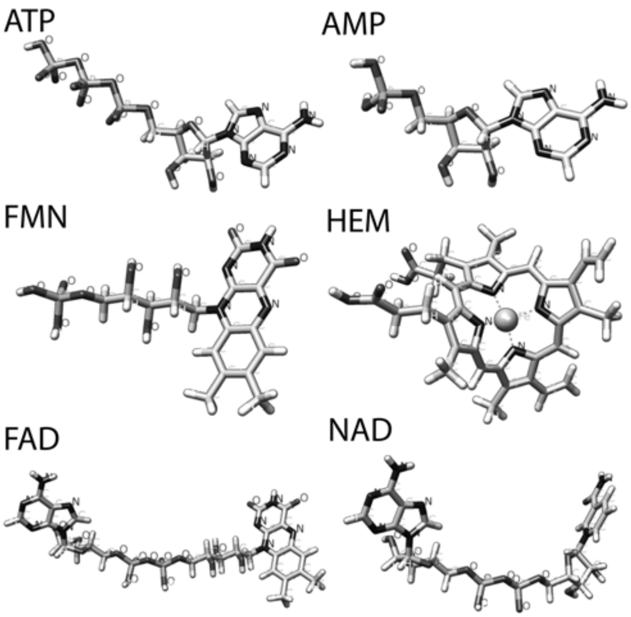


Figure 3

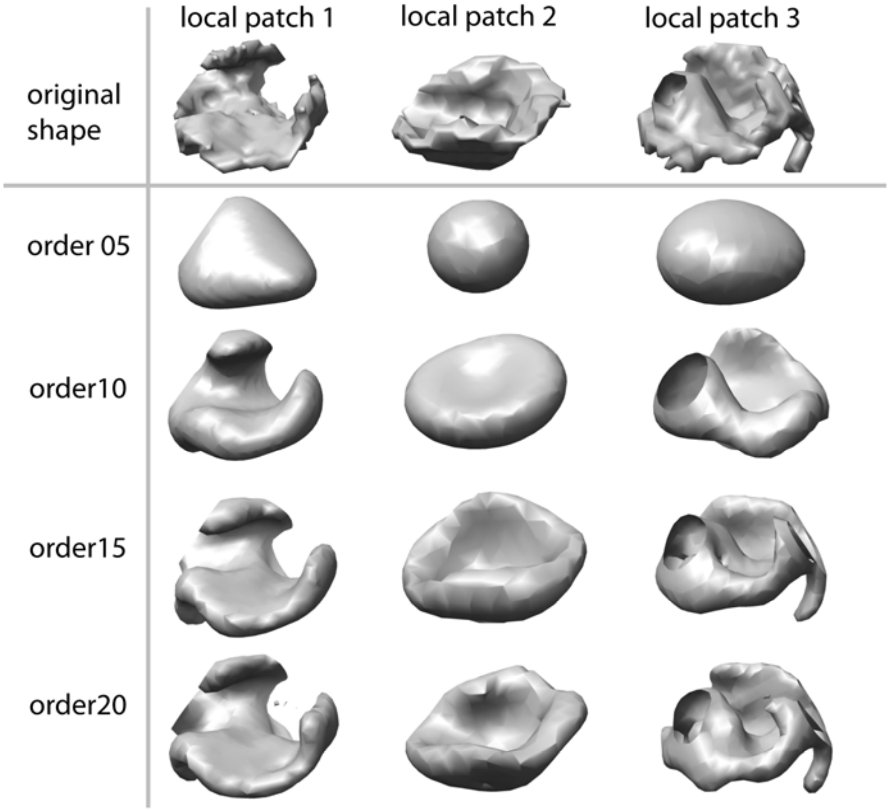


Figure 4

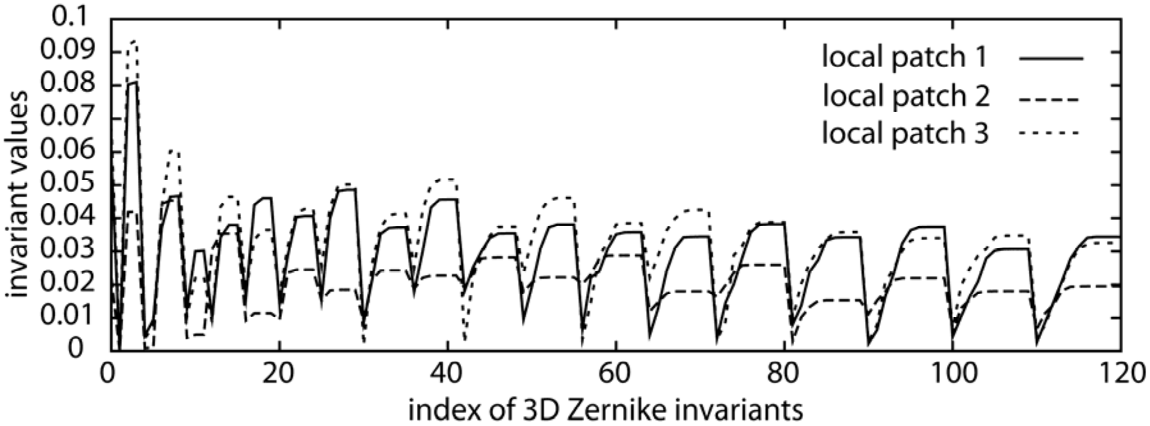


Figure 5

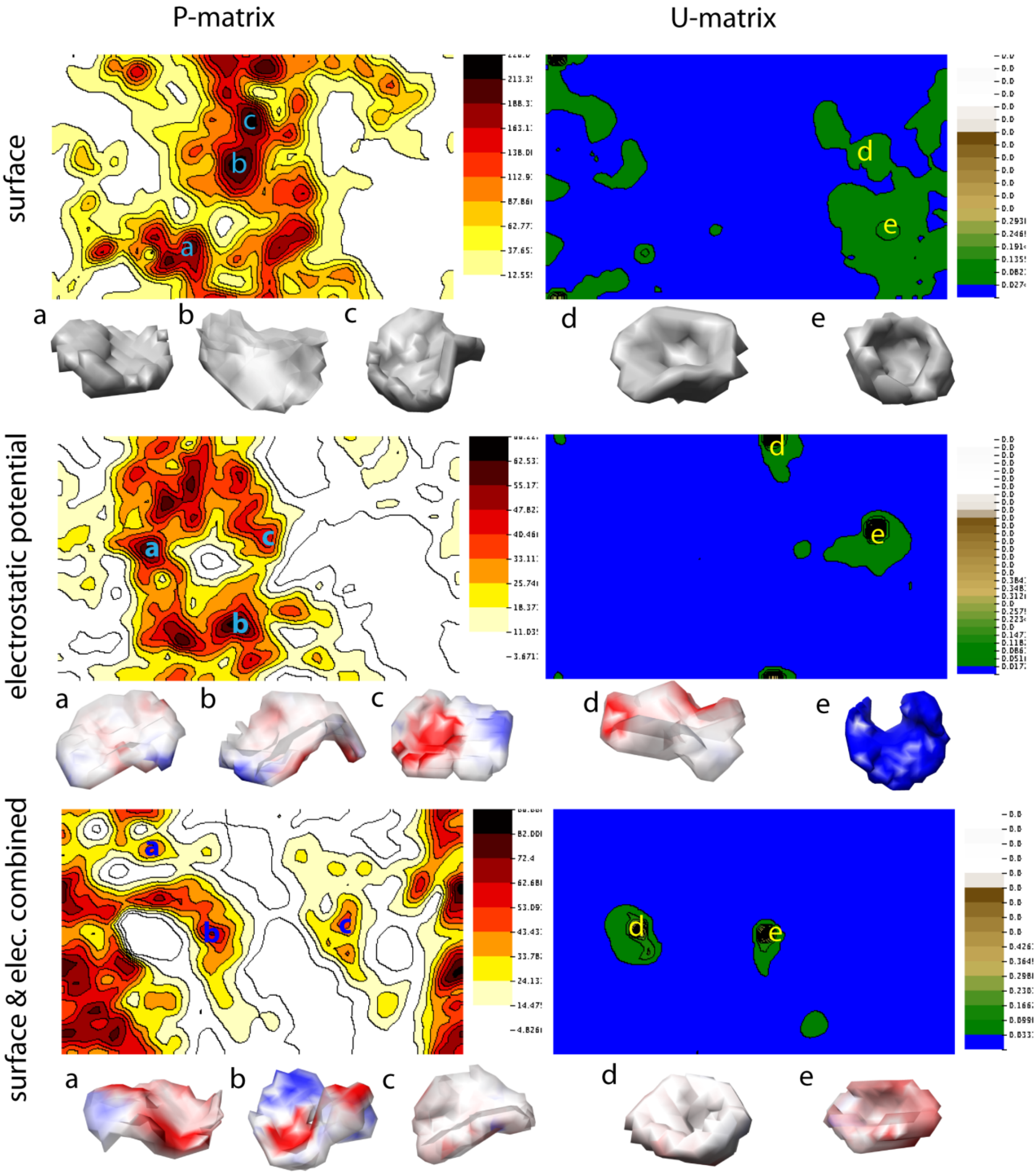

P-matrix
U-matrix
surface
c
b
a
213.3
188.3
163.1
138.0
112.9
87.86
62.77
37.65
12.55
d
e
0.293
0.246
0.191
0.135
0.082
0.027
a
b
c
d
e
electrostatic potential
a
c
b
62.53
55.17
47.82
40.46
33.11
25.74
18.37
11.03
3.671
d
e
0.384
0.348
0.312
0.257
0.223
0.147
0.118
0.086
0.051
0.017
a
b
c
d
e
surface & elec. combined
a
b
c
82.00
72.4
62.68
53.09
43.43
33.78
24.13
14.47
4.826
d
e
0.426
0.364
0.298
0.230
0.166
0.099
0.033
a
b
c
d
e

Figure 6

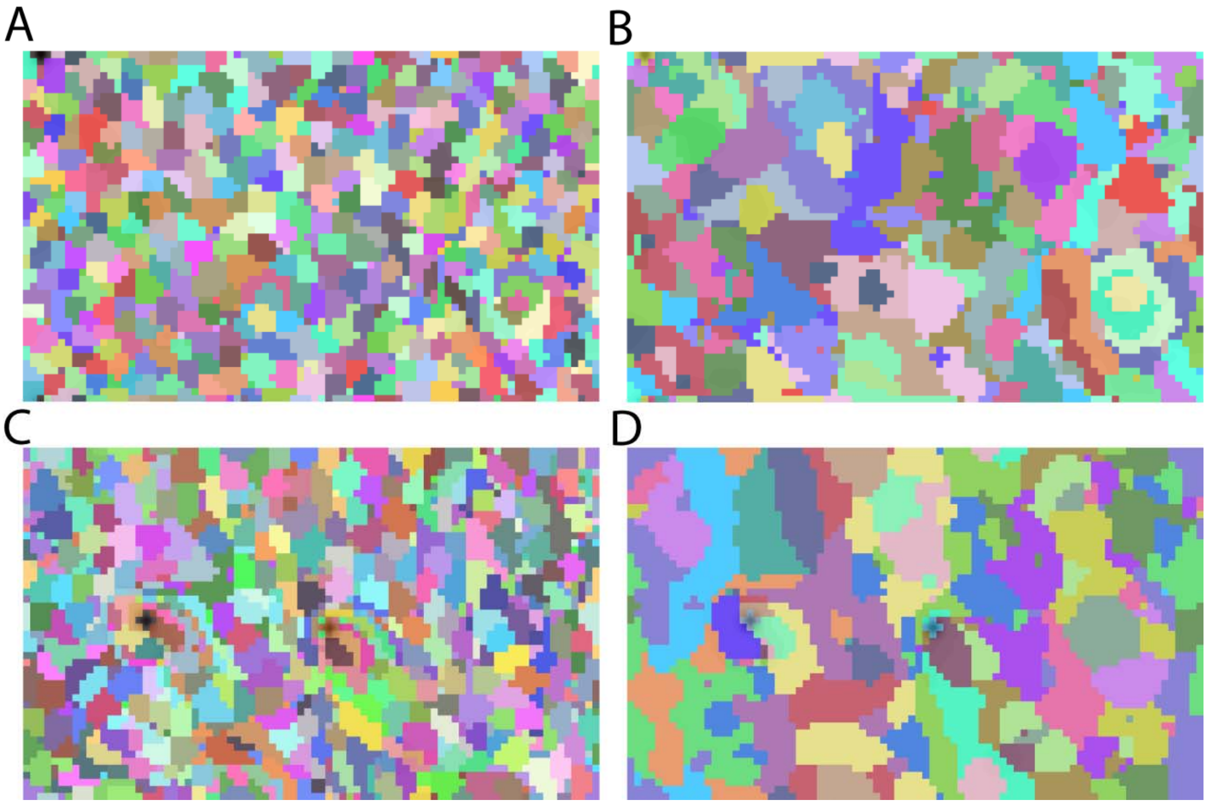


Figure 7

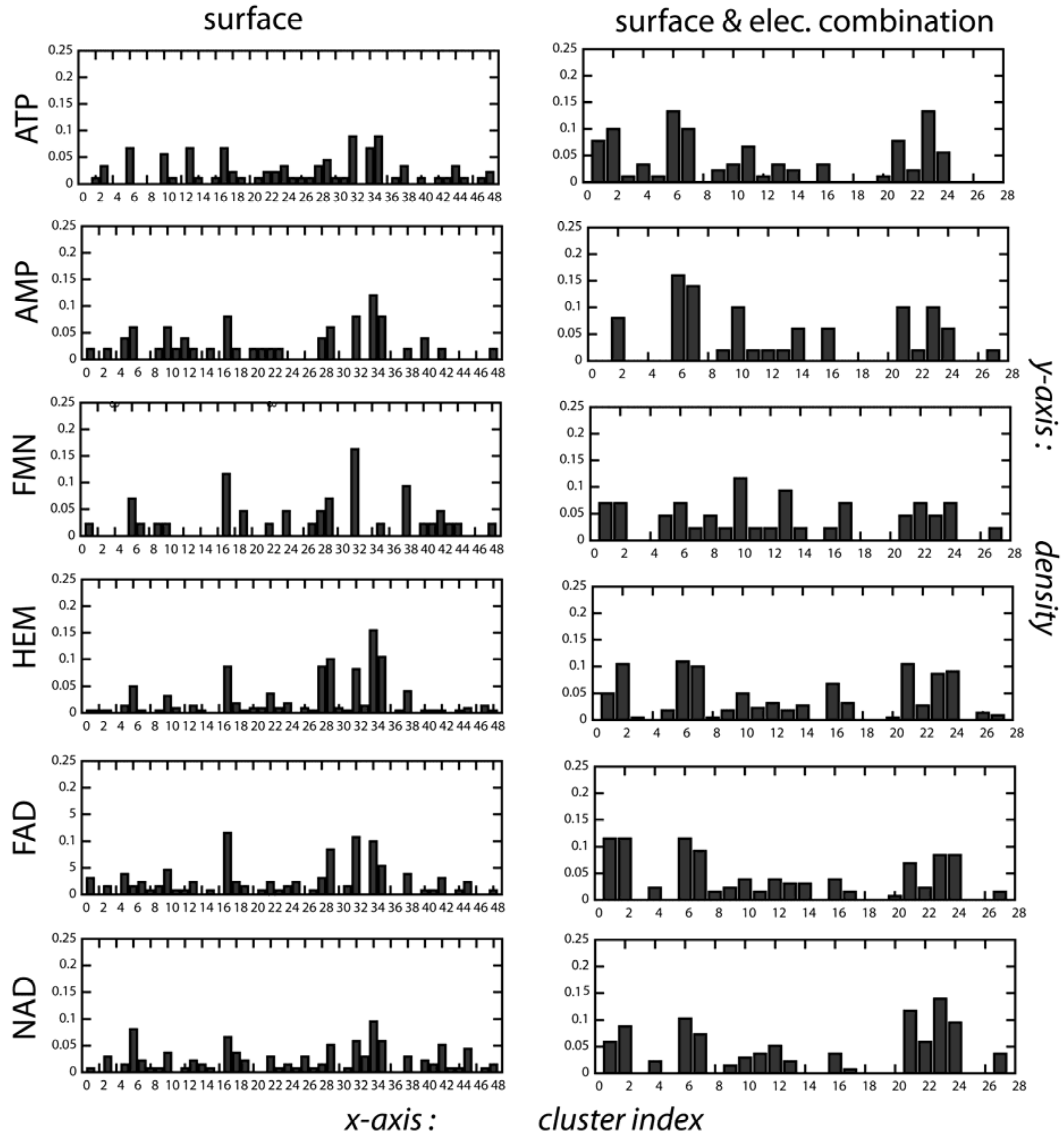
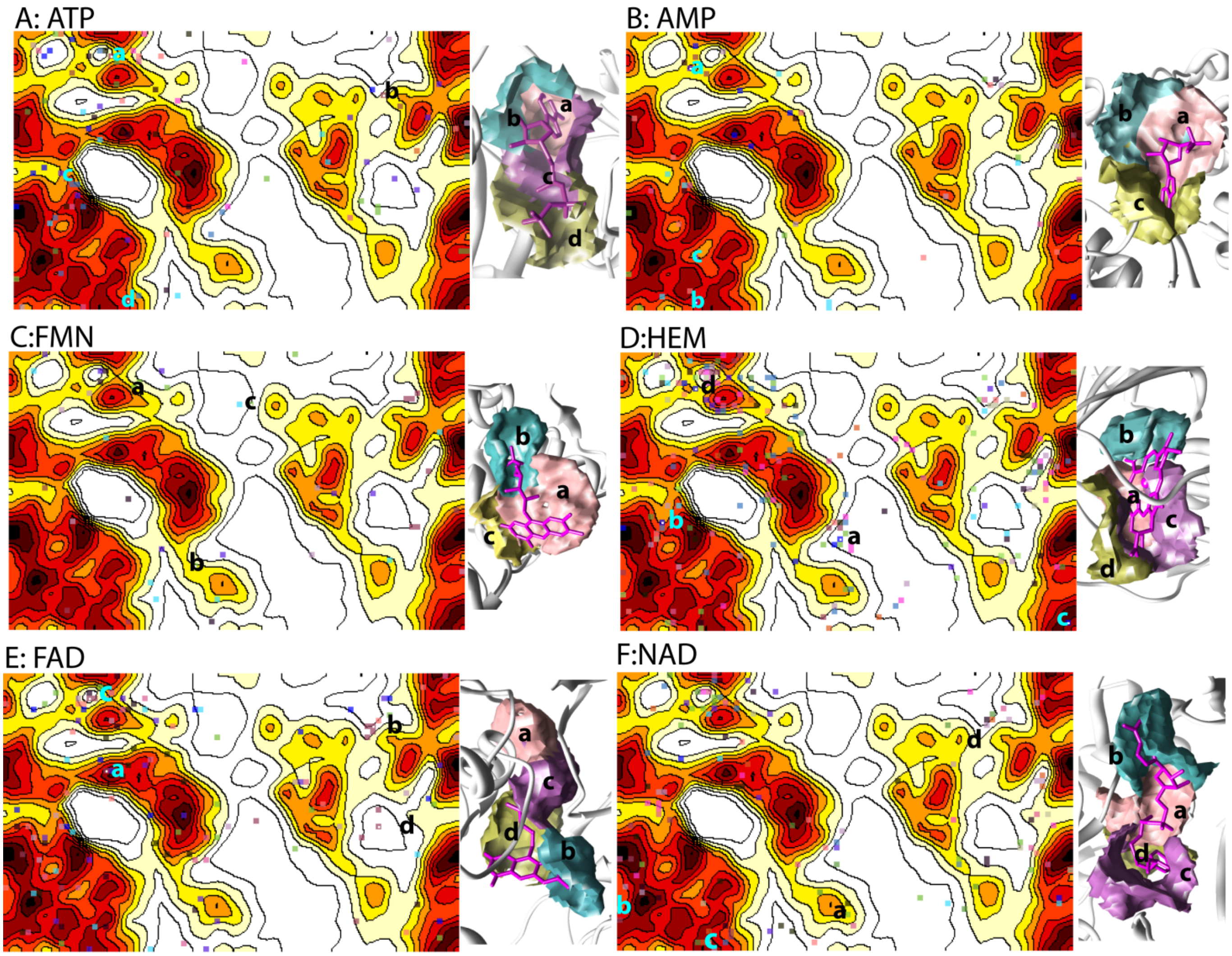

surface
surface & elec. combination
ATP
AMP
FMN
HEM
FAD
NAD
0
0.05
0.1
0.15
0.2
0.25
5
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28
y-axis :
density
x-axis :
cluster index

Figure 8


A: ATP
a
b
c
d
b
a
c
d
B: AMP
a
c
b
b
a
c
C:FMN
a
c
b
b
a
c
D:HEM
d
b
a
c
b
a
c
d
E: FAD
c
a
b
d
a
c
d
b
F:NAD
d
b
a
c
b
a
d
c