Protein Binding Ligand Prediction Using Moments-Based Methods

Rayan Chikhi, Lee Sael, and Daisuke Kihara

Abstract Structural genomics initiatives have started to accumulate protein structures of unknown function in an increasing pace. Conventional sequence-based function prediction methods are not able to provide useful function information to most of such structures. Thus, structure-based approaches have been developed, which predict function of proteins by capturing structural characteristics of functional sites. Particularly, several approaches have been proposed to identify potential ligand binding sites in a query protein structure and to compare them with known ligand binding sites. In this chapter, we introduce computational methods for describing and comparing ligand binding sites using two dimensional and three dimensional moments. An advantage of moment-based methods is that the tertiary structure of pocket shapes is described compactly as a vector of coefficients of series expansion. Thus a search against an entire PDB-scale database can be performed in real-time. We evaluate two binding pocket representations, one based on two-dimensional pseudo-Zernike moments and the other based on three-dimensional Zernike moments. A new development of pocket comparison method is also mentioned, which allows partial matching of pockets by using local patch descriptors.

Introduction

Functional assignment of proteins is a fundamental and challenging problem in biology and bioinformatics [1]. In recent years structural genomics projects have been solving an increasing number of protein structures which were not able to be characterized by traditional sequence based methods [2, 3]. Therefore, much effort has been devoted recently to the development of function prediction methods based on structural information. Structure-based function prediction methods aim either to

D. Kihara (✉)
Department of Biological Sciences; Department of Computer Science; Markey Center for Structural Biology, College of Science, Purdue University, West Lafayette, IN 47907, USA
e-mail: dkihara@purdue.edu

D. Kihara (ed.), *Protein Function Prediction for Omics Era*,
DOI 10.1007/978-94-007-0881-5_8,

capture global structure or local structure similarity to proteins of known function in the structure database (PDB [4]). Global approaches are motivated by the observation that protein folds are better conserved than primary sequences. Alternatively, local methods aim to capture properties of functional sites, where interactions with ligand molecules or other proteins take place. Ligand binding sites are intrinsically unrelated to global folds, as two sequentially and structurally dissimilar proteins may bind the same ligand molecule [5]. Because two proteins with similar folds often have different functions [6] ligand binding sites are of particular interest in structure-based function prediction. In most cases ligand molecules bind to a protein at its surface pocket regions [7], hence detecting pockets enables identification of binding sites [8–10]. Binding ligand prediction approaches have two logical steps: (i) detection of the pocket region in a given protein surface and (ii) comparison of a pocket against a database of known sites.

Several methods have been developed to predict the location of ligand binding sites in a protein surface. These methods are based on the detection of specific geometric properties on the protein surface. For instance, gaps can be detected on a protein surface using probe spheres [11–13]. Grid-based methods [7, 14, 15] scan protein surface points for various properties, e.g. voids, the visibility, or the depth. Voronoi diagrams have also been applied to identify pockets by recognizing depressed regions [16]. Recent methods combine geometrical criteria with evolutionary information [17–20] and energetics [21–23].

Comparison of binding sites relies on how pockets are represented. These representations are either based on coordinates of residues/atoms or shape of pocket surfaces. In the former representation, protein binding pockets are described as sets of three dimensional coordinates of key residues [24–26], for which pair-wise similarity is computed, for example, with the root mean square deviation (RMSD). The geometric hashing [27] technique defines a distance between two binding sites by the number of spatially matching atoms. Alternatively, in a type of fingerprinting methods, a site is represented by all the distances between residues, which are then grouped by types for fast matching [28, 29]. Similar fingerprinting approaches have also been applied to atoms on the solvent accessible surface [5, 30].

Surface-based representations of binding sites are based on a wide spectrum of computational techniques. Moments-based methods belong to this category and are thoroughly discussed in the next section. Graph-based representation is an alternative choice for representing protein surfaces. Klebe et al. employed subgraph matching algorithm to describe the surface geometry and the electrostatic potential of binding pockets [31]. Kinoshita et al. used a clique detection algorithm [32] for local surface similarity retrieval in their method named eF-Site. Using the eF-Site and its associated tool, eF-Seek, users can search functional sites in an unannotated query structure [32, 33]. Another approach uses the spin-image, a 2D histogram representation for protein surface points, which describes relative geometrical position of each point to the other points [34]. Generally speaking, moment-based descriptors have advantage over graph methods and 2D histograms in terms of lower time complexity (thus faster running time).

In this chapter, we review three-dimensional (3D) and two-dimensional (2D) geometric moments for the representation and comparison of protein ligand binding sites. Concretely, we describe application of the 2D pseudo-Zernike moments and the 3D Zernike descriptors. The 2D pseudo-Zernike (p-Z) moments are employed to describe the projection of the pocket surface on a 2D image. The 3D Zernike descriptors (3DZD) can directly represent 3D pocket surface properties. These moments compactly represent a binding pocket by a vector of coefficients of the series expansion. The rest of this chapter is organized as follows: First, an overview of the theoretical differences between these moments is given. In addition to the p-Z moments and the 3DZD descriptors, we also discuss the spherical harmonics in comparison with the two methods. Then, we describe our recent works on the application of the 2D p-Z moments and the 3DZD for binding ligand prediction for proteins. The methodology quantifies similarity of pockets by the Euclidean distance of the vector of p-Z/3DZD coefficients of pockets and uses a k-NN classifier to make final prediction of binding ligand for a query pocket. Our methods are benchmarked on two datasets. Finally, recent ongoing development in our group on a new pocket comparison method is discussed, which uses local surface patch descriptors to allow matching of flexible binding ligands.

Pocket Surface Shape Descriptors

In this section we briefly describe 3D and 2D moment-based pocket descriptors, which will be used in the subsequent sections. For the 3D descriptors of pockets, we introduce the spherical harmonics and the 3D Zernike descriptors. For the 2D descriptors, we introduce the p-Z moments.

Spherical Harmonics

Spherical harmonics are a set of mathematical moments which are applied for 3D volumetric representation of objects [35]. The object shape is approximated as a spherical function $f(\theta,\phi)$ defined on the unit sphere, which describes the distance to the outermost surface of the object from the center for the direction (θ,ϕ). The function $f(\theta,\phi)$ is then expanded as a series of spherical harmonics

$$f(\theta,\phi) \approx \sum_{l=0}^{l_{\max}} \sum_{m=-l}^{l} c_{lm}\, \mathrm{Re}\left[Y_{lm}(\theta,\phi)\right], \tag{1}$$

where $l_{\max}$ is the moments order, $\mathrm{Re}[Y_{lm}(\theta,\phi)]$ is the real part of the spherical harmonic functions, and c_{lm} are the associated coefficients. The similarity of two objects can be measured by the Euclidean distance of the vectors of coefficients c_{lm} of the two pockets. Since spherical harmonics are not rotationally invariant, in principle pose normalization of object is needed.

Kahraman and colleagues [8] defined the Interact Cleft Model for ligand binding sites by employing spherical harmonics as follows. For a ligand binding pocket, the volume of a ligand binding pocket is defined by SURFNET [11] spheres within 0.3 Å to protein atoms interacting with the bound ligand. The software HBPLUS [36] is used to determine such atoms. To achieve rotation invariance, a coordinate system is defined at the center of gravity of the pocket volume. The moment of inertia tensor for the pocket volume V is a matrix of components

$$I_{i,j} = \int_V (r^2 \delta_{i,j} - r_i r_j) dV, \tag{2}$$

where $i, j = x, y, z$ and r is the vector from the center of gravity to a point in the volume. The pocket is rotated so that its moment of inertia tensor is diagonal with maximal values in x followed by y then followed by z. The outermost surface of these spheres is then expanded as a spherical harmonics series $f(\theta,\phi)$ where the order $l_{\max}$ is set to 16.

3D Zernike Descriptors

We have applied the 3D Zernike descriptors (3DZD), which also give a series expansion of a 3D function. It allows a compact and rotationally invariant representation of a 3D object. Mathematical foundation of the 3DZD was laid by Canterakis [37], then Novotni and Klein [38] have applied it to 3D shape retrieval. Here we provide a brief mathematical derivation of the 3DZD. Refer to the two papers [37, 38] for more technical details.

The surface of a ligand binding pocket is extracted using the Connolly surface [39] of protein heavy atoms within 8 Å to any heavy atom of the bound ligand, then placed on a 3D grid. To represent a surface shape, each grid cell (voxel) is assigned the value of 1 if it contains the protein surface and the value of 0 otherwise. For representing other physicochemical properties, such as the electrostatic potentials and hydrophobicity values, values are also assigned only to the surface voxels. The resulting voxels-values mapping is considered as a 3D function, $f(x)$, which is expanded into a series in terms of Zernike-Canterakis basis [38] defined by the following collection of functions:

$$Z_{nl}^m(r, \vartheta, \varphi) = R_{nl}(r) Y_l^m(\vartheta, \varphi), \tag{3}$$

with $-l < m < l$, $0 \leq l \leq n$, and $(n - l)$ even. The function $Y_l^m(\vartheta, \varphi)$ are the spherical harmonics [40] and $R_{nl}(r)$ are radial functions defined by Canterakis, constructed so that $Z_{nl}^m(r, \vartheta, \varphi)$ can be converted to polynomials, $Z_{nl}^m(\mathbf{x})$, in Cartesian coordinates. Now 3D Zernike moments of $f(\mathbf{x})$ are defined as the coefficients of the expansion in this orthonormal basis, i.e. by the formula

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

Finally, the rotational invariance is obtained by defining the 3DZD series, F_{nl}, as norms of vectors Ω_{nl}:

$$F_{nl} = \sqrt{\sum_{m=-l}^{m=l} (\Omega_{nl}^{m})^2} \tag{5}$$

The parameter n is called the order of 3DZD, which determines the resolution of the descriptor. As stated above, n defines the range of l and a 3DZD is a series of invariants (Eq. (5)) for each pair of n and l, where n ranges from 0 to the specified order. We use order $n = 20$ in the pocket comparison. which was shown to provide sufficient accuracy in a previous works of shape comparison [38]. The order $n = 20$ yields 121 invariant numbers (Eq. (5)).

As for the surface electrostatic potentials, 3DZD is computed separately for the pattern of positive values and for the negative values and later concatenated into a single vector. The separation of negative patterns and positive patterns is done by creating an input grid only of negative values and only of positive values and calculating 3DZD for each grid separately [46].

The obtained 3DZD is normalized to a unit vector by dividing each moment by the norm of the whole descriptor. This normalization is found to reduce dependency of 3DZD on the number of voxels used to represent a protein [46]. An example of the invariant values of the 3DZD of a ligand binding pocket (Fig. 1a) is shown in Fig. 1b.

In our previous works, we have applied the 3DZD successfully to various protein and ligand structure analyses [41–43], including rapid protein global shape analysis (http://kiharalab.org/3d-surfer) [44, 45], quantitative comparison for protein surface physicochemical property [46], small ligand molecule comparison [47], protein–protein docking prediction [48], and comparison of low-resolution electron density maps [49].

2D Pocket Model with Pseudo-Zernike Moments

We have also developed a new computational pocket model using two dimensional moments [50]. The key aspect of this method is the projection on a 2D plane of a spherical panoramic picture computed from the center of the binding pocket. The 3D to 2D dimensional reduction relies on the finding that pockets can be quite reliably pre-aligned using their opening.

Here, the shape of a pocket is extracted using the same procedure as 3DZD. A 3D Cartesian coordinate system $(\vec{x}, \vec{y}, \vec{z})$ is defined relative to a binding pocket, following the representation in Fig. 1c. The origin of the coordinate system is the center of gravity of the binding pocket, provided the latter is not inside the protein volume; otherwise, the origin is defined as any of the closest points outside. The *opening* of a binding pocket is the set of rays starting at the center of gravity which do not intersect the volume of the pocket. The unit vector of the x-axis is defined

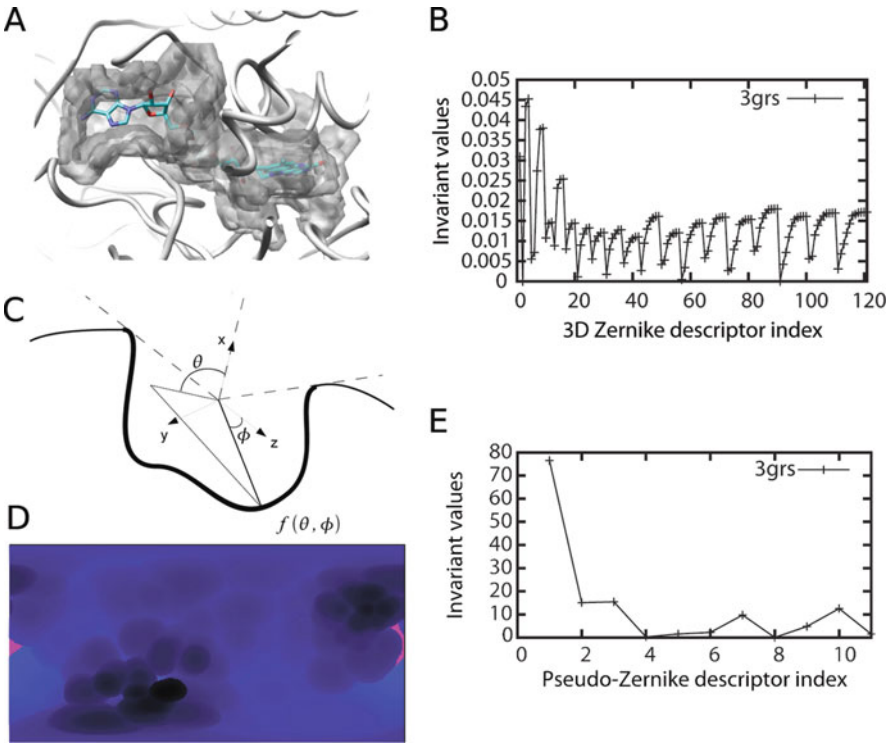

Fig. 1 Examples of the binding pocket representation with the 3DZD and 2D pseudo-Zernike descriptors. **a**, flavin adenine dinucleotide (FAD) binding pocket in PDB entry, 3grs. Pocket surface region within 8.5 Å to the ligand is shown. **b**, the 3DZD of the pocket. **c**, the coordinate system for projecting the pocket to the 2D map. **d**, the projected pocket. The distance from the center of the pocket to the pocket surface is represented in a color code from *blue* (closer) to *black* (more distant). *Pink* region shows aperture of the pocket. The x-axis is for θ and the y-axis shows ϕ. **e**, the pseudo-Zernike descriptor of the pocket

as a collinear vector to the average vector of the pocket opening. In cases where the opening is empty, the x-axis is arbitrarily defined. The remaining two axes, $\vec{y}$ and $\vec{z}$, are defined arbitrarily such that the basis $(\vec{x}, \vec{y}, \vec{z})$ is orthogonal. Optionally, an additional pre-alignment step can be applied. The z-axis is rotated such that its principal moment of inertia is maximized over all possible directions on the plane orthogonal to the x-axis. Simulations showed that this pre-alignment step is not necessary when using rotationally invariant descriptors, such as the p-Z moments used here (See fig. 3 in Chikhi et al. [50]).

A spherical function $f(\theta, \phi)$ is defined for the outermost surface of the binding pocket. Practical computation of $f(\theta, \phi)$ can be done using ray-casting. Rays are shot in every direction (θ, ϕ) from the center of gravity of the pocket to the pocket surface, and a value for the direction (e.g. distance from the center) is taken from the surface point which first intersects the ray. If a ray never intersects the protein surface, a null value is assigned to the ray direction. Note that the function $f(\theta, \phi)$

can also describe any surface property, such as geometry or electrostatic potential [50]. Then, the function f is mapped to a 2D plane in order to be described using two dimensional moments.

Since no 3D to 2D projection preserves area, shape, and distance properties altogether, there is no solution to perfectly map function f to a 2D plane without distortion. It was found that a simple distance preserving projection, the *plate-carrée* projection, is sufficient for the purpose of pocket matching. By mapping $f(\theta, \phi)$ to a planar image using this projection, the bottom of the pocket ($\theta = \pi$) is projected to the center of the image and the opening of the pocket $\left(\theta = 0,\ \phi = \frac{\pi}{2}\right)$, is projected to the sides (Fig. 1d). The resolution of the picture is 360 × 180, as coordinates are mapped to integer values of (θ, ϕ), resulting in 64,800 rays shot from the pocket center of gravity to each (θ, ϕ) direction. Rotations around the x-axis of the pocket correspond to rotations around the center of the image. However, since the $\vec{z}$ axis is arbitrarily defined under the 2D pocket model, the ϕ coordinate has no reference. Conveniently, the p-Z moments are mathematically invariant around the center of the image. And practically, as we will see in the results, these moments can robustly describe a projected pocket despite the lack of reference for the $\vec{z}$ axis.

2D Pseudo-Zernike Moments

The p-Z moments [51] have been employed for describing an image shape in pattern recognition applications, and they are shown to be less sensitive to noise than conventional 2D Zernike moments [52, 53]. The p-Z moments use a set of complete and orthogonal basis functions defined over the unit circle ($x^2+y^2 \leq 1$) as follows:

$$V_{n,m}(x,y) = e^{im\theta} R_{nm}(r) = e^{im\theta} \sum_{s=0}^{n-|m|} \frac{(-1)^s (2n+1-s)! \rho^{(n-s)}}{s!(n+|m|+1-s)!(n-|m|-s)!} \tag{6}$$

where $\rho = \sqrt{x^2+y^2}$, $\theta = \tan^{-1}(y/x)$, and $n \geq 0$, $|m| \leq n$. Using the polynomials, the p-Z moments of the order n and the repetition m for a 2D image $f(x,y)$ are defined as:

$$A_{n,m} = \frac{n+1}{\pi} \int_{x^2+y^2 \leq 1} f(x,y) V^*_{n,m}(x,y) dxdy \tag{7}$$

The asterisk ($*$) denotes the complex conjugate. In this study, the order of moments $n = 4$ is used for most of the computation. An example of the pseudo Zernike values is shown in Fig. 1e.

Theoretical Comparison of Moments in Shape Descriptors

We briefly discuss differences between these three moments from a mathematical point of view. Obviously, the 2D p-Z moments describe a 2D function, hence a

pocket 3D structure needs to be initially projected to a 2D image. On the other hand, the spherical harmonics and the 3DZD represent 3D objects, thus they can directly handle the 3D coordinates of a pocket. The coordinate system defined in Fig. 1c makes the p-Z moments rotationally invariant around the center of the image. However, a disadvantage arises from distortions caused by the projection, although in the benchmark study the 2D pocket model showed comparable performance with the 3DZD [50].

Comparing the spherical harmonics and the 3DZD, the 3DZD has a radial function $R_{nl}(r)$, (Eq. (3)), while the spherical harmonics do not. This difference results in an advantage of the 3DZD over the spherical harmonics in describing 3D pockets which intersect with a ray of a certain direction (θ, ϕ) for multiple times at different distance, r. The 3DZD can naturally handle such shapes (non star-like shapes), because it can assign a different value at each r. On the other hand, naïve use of the spherical harmonics can only take one value per direction. Therefore, usually only the outermost (or innermost) surface of an object is described by the spherical harmonics. To describe non star-like shapes, Funkhouser et al. used multiple concentric spherical shells [35].

Another advantage of the 3DZD over the spherical harmonics is that it is invariant to rotation of the object (Eq. (5)), while the direct use of the spherical harmonics is not. Thus, the 3DZD does not need pose normalization (pre-alignment) of objects for comparing and computing the similarity. This is advantageous in constructing a database of pockets since the 3DZD of pockets can be pre-computed and stored. The spherical harmonics can obtain the rotational invariance by the aforementioned use of concentric spherical shells. However, there are several drawbacks to this approach. As radial consistency of objects is not preserved (shells can be rotated with no impact on the descriptors), a certain amount of shape information is lost. Furthermore, because of polar sampling, spherical harmonics descriptors are not practically robust to rotation [54]. Also the 3DZD is more compact than the spherical harmonics by one order of magnitude [55], because adjacent spherical shells in the spherical harmonics descriptors are highly correlated.

Overall, the 3DZD is an improvement over spherical harmonics descriptors. The p-Z moments have not yet been formally studied for the description of 3D objects using a single projection, since this approach seems to be relatively specific to the description of binding pockets.

Binding Ligand Prediction Using the Pocket Descriptors

Using the 3DZD and the p-Z descriptors discussed above, we built a binding ligand prediction method for protein structures named Pocket-Surfer. Since both representations describe a pocket as a vector of coefficients, similarity of two pockets can be quantified by computing the Euclidean distance of their descriptors (vectors).

The 3DZD and the p-Z descriptors contain shape information of the pockets. However, the information of the size of the pockets is lost since the pockets are first fit to a unit sphere (for 3DZD) or a unit circle (for the pseudo Zernike descriptors)

in the process of computing the moments. Therefore, we add the size information of a pocket into the vector as follows:

$$\text{Descriptor}(P) = (w \cdot S_{\text{P}}, A_1^{\text{P}}, A_2^{\text{P}}, \ldots, A_k^{\text{P}}, \ldots, A_N^{\text{P}}), \tag{8}$$

where S_{P} is the size of the pocket P weighted by a factor w, $A_k{}^{\text{P}}$ is the kth value of the pocket moments (either 2D or 3D), and N is the total number of values of the moments. As the pocket size Sp, we used the average distance from the center of gravity G of the pocket to the pocket surface.

Equipped with the pocket descriptors and a similarity metric (i.e. Euclidean distance), pockets in a database are sorted according to the distance to a query pocket. Using the k nearest pockets to the query, the binding ligand for the query pocket (pocket type) is predicted using a k-nearest neighbors (k-NN) classifier as follows. The scoring function for a binding pocket of a ligand type F is defined as

$$\text{Pocket\_score}(F) = \sum_{i=1}^{k} \left(\delta_{l(i),F} \log \left(\frac{n}{i} \right) \right) \cdot \frac{\sum_{i=1}^{k} \delta_{l(i),F}}{\sum_{i=1}^{n} \delta_{l(i),F}}, \tag{9}$$

where $l(i)$ is a function that returns the ligand type (AMP, FAD, etc.) of the ith closest pocket to the query, n is the total number of pockets in the database, and the indicator function $\delta_{X,Y}$ equals to 1 if X is of type Y, and is null otherwise. The role of the first term in this scoring function is to assign higher scores to pockets with higher ranks, within the top k results. The second term is a normalization factor of the score by considering the number of pockets of the type F in the database. The numerator is the number of pockets of the type F retrieved within top k and the denominator is the number of all the pockets of the type F in the database. Using Eq. (9), the score is computed for all the pocket types and they are sorted by the score.

To summarize the Pocket-Surfer procedure, a flow-chart is presented in Fig. 2. Given a query protein structure, ligand binding pockets on the protein surface are detected (i.e. predicted) by geometrical criteria using a method like LigSite [18] or VisGrid [7]. In the benchmark study, known ligand binding pockets are used (i.e. pockets are extracted as the surface regions which are in contact to the binding ligand molecule) to test the pocket comparison and the ranking ability of the procedure. Then, the Connolly surface [39] of pockets is constructed. Next, the pocket descriptor (Eq. (8)), either the 3DZD or the p-Z descriptor, is computed for the query pocket. Finally, the distance from the query to all pocket descriptors precomputed and stored in a database is computed, and the ligand type for the query is predicted using Eq. (9). Pocket-Surfer has been implemented as a web server at http://kiharalab.org/pocket-surfer/. Currently the pocket database to be searched holds only a limited number of pockets used in the benchmark study of the published paper [50]. Expansion of the database is under way to make the server bear practical use of binding ligand prediction.

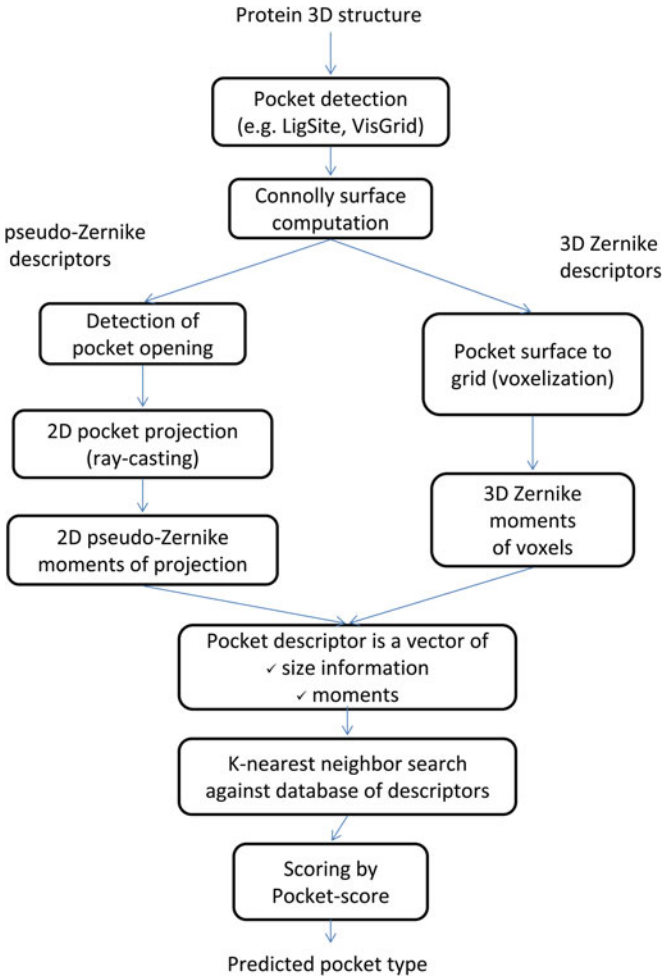

Fig. 2 Schematic flow chart of the binding ligand type prediction procedure using 2D pseudo-Zernike or 3D Zernike descriptors

Benchmark Results of Binding Ligand Prediction

In a recent paper [50], we benchmarked the performance of binding ligand with the p-Z, spherical harmonics, and 3DZD pocket models on two datasets. We briefly summarize the results in this section. The first dataset (the Kahraman set, named after the author [8] who compiled this dataset) consists of 100 evolutionary-distant proteins binding one of nine different ligand molecules (see the legends of Fig. 3). This dataset is used to train parameters and compare the performance of 3DZD and p-Z with spherical harmonics. The second dataset (the Huang set [18]) is independent from the first one in terms of proteins and ligand types. It contains 175 proteins,

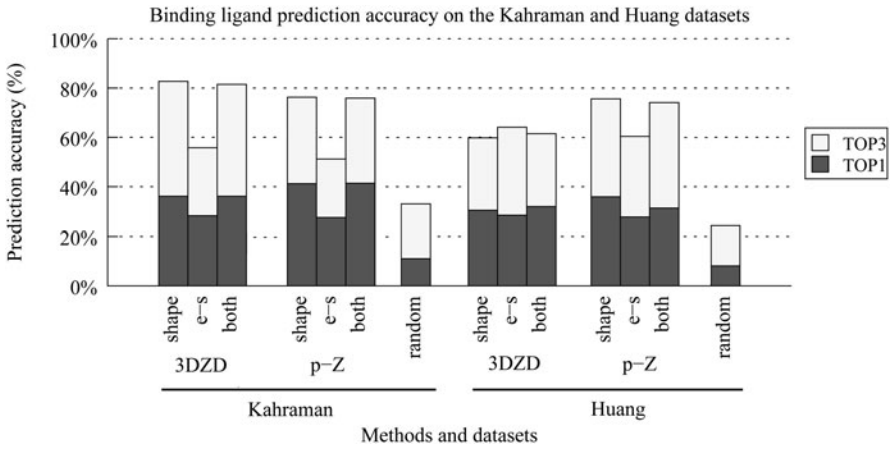

Fig. 3 The average binding ligand prediction accuracy for the Kahraman and the Huang dataset with the 3DZD and the 2D p-Z descriptors. The Kahraman dataset contains 100 proteins, each of which binds one of the following nine different ligands: adenosine-monophosphate (AMP) (9), adenosine-5′-triphosphate (ATP) (14), flavin adenine dinucleotide (FAD) (10), flavin mononucleotide (FMN) (6), glucose (GLC) (5), heme (HEM) (16), nicotinamide adenine dinucleotide (NAD) (15), phosphate (PO4) (20), and steroid (STR) (5). In the second parentheses the number of entries is shown. The Huang dataset consists of 175 proteins, which bind either of twelve ligand molecules: adenosine (ADN) (11), biotin (BTN) (12), fructose 6-phosphate (F6P) (12), fucose (FUC) (14), galactose (GAL) (36), guanine (GUN) (12), mannose (MAN) (18), O1-methyl mannose (MMA) (10), 2-phenylimidazole (PIM) (5), palmitic acid (PLM) (26), retinol (RTL) (5), and 2′-deoxyuridine 5-monophosphate (UMP) (13). The average Top-1 and Top-3 success rates of binding ligand prediction for all ligand type are reported. Results are shown for the shape descriptors, the electrostatics descriptors and both combined. For the combination of the shape and the electrostatic (e-s) potential descriptors, the average Euclidean distance by the pocket shape and the electrostatic potential descriptors are used

each of which binds one of twelve ligand molecules. Based on the performance on the Kahraman dataset, the descriptors parameters for p-Z (resp. 3DZD) descriptors were set to $w = 4.5$ (0.04) and $n = 4$ (20). The number of neighbors used in the k-NN classifier was set to $k = 24$. Using the two datasets, performance of the binding ligand prediction was examined for the pocket shape descriptors that combine the pocket and size shape information (Eq. (8)) and also for the electrostatic potential descriptors. To compute the surface electrostatic potential descriptors, the electrostatic potential on the protein surface is mapped on the 2D image for the p-Z while on the voxels of the 3D grid for the 3DZD [50].

First, we compared the performance of the shape descriptor of 3DZD and the p-Z with that of the spherical harmonics on the Kahraman dataset. The value for the spherical harmonics was taken from the paper by Kahraman et al. [8]. Both 3DZD and the p-Z performed slightly better than the spherical harmonics in terms of the Area Under the Curve (AUC) values of the receiver operating characteristic (ROC) curve [56]. The AUC values of the 3DZD, the p-Z and the spherical harmonics were 0.81, 0.79 and 0.77, respectively. We have also examined the performance of the

p-Z with pocket pre-alignment (Eq. (2)) but the improvement was only 0.75%. Thus the p-Z is practically robust enough to rotation.

Figure 3 shows the Top-1 and Top-3 success rate of the 3DZD and the p-Z averaged over all ligand types. For the Top-3 success rate, a ligand for a pocket is considered to be correctly predicted if the correct ligand is included within the top 3 scoring ligand types according to the *Pocket_score* (Eq. (9)). For the Kahraman dataset, the best Top-1 success rate was achieved by the pocket shape descriptor of the p-Z (41.2%), while the shape descriptor of the 3DZD was the best for the Top-3 success rate (82.7%). The pocket shape descriptors (left bars) performed significantly better than the electrostatic potential descriptors (middle bars) for both 3DZD and the p-Z. Because of this, combining them did not improve the performance (81.5% by the 3DZD and 75.9% by the p-Z). This observation is in agreement with a previous report that electrostatic potential is variable within families of binding pockets [57]. For the Huang dataset (Fig. 3, right), both 3DZD and the p-Z showed lower success rate by the shape descriptor as compared with the Kahraman dataset. On the other hand, the electrostatic potential descriptors of both 3DZD and the p-Z showed a higher success rate on this dataset relative to the Kahraman set. As a result, for the 3DZD, the combination of the shape and the electrostatic potential descriptors showed improvement over the shape descriptor. The best Top-1 (35.9%) and the Top-3 success rate (75.6%) were achieved by the p-Z shape descriptor.

Figure 4 shows the Top-1 and Top-3 accuracy for individual pocket types on both datasets. PO_4 was predicted very well because it is distinguished by its smaller size from the other ligands. Some ligands, such as FMN, were poorly predicted. FMN is the most flexible ligand among the three smallest ligands in the dataset (GLC, FMN, and STR) with an average RMSD of 1.08 Å. The success rate largely differs from ligand to ligand and the trends are consistent for the 3DZD and the p-Z. This implies that the difference in the performance for each ligand is attributed not to the characteristics of the approaches but to the actual similarity of pockets of particular ligand types.

Performance with Ligand-Free Pockets and Predicted Pockets

In practical situations of binding ligand prediction, one of the two cases may arise: (1) the binding pocket in a query structure is known, but it is in a ligand-free conformation or (2) a binding pocket is unknown, hence it needs to be predicted. The challenge for the first case is the difference in shapes of ligand-free and ligand-bound binding pockets. To assess this difference, we searched the Huang dataset with ligand-free pockets and determined pocket retrieval accuracy with the p-Z and 3DZD pocket shape descriptors. For the p-Z (resp. 3DZD) descriptors, in 11 (resp. 7) out of 12 ligands the ligand-free pockets are retrieved with a similar or often better AUC value than the closest ligand-bound pockets [50]. The RMSD value of the ligand-bound and ligand-free proteins ranges from 0.19 to 2.48 Å with an average value of 0.86 Å. This is consistent with a recent study [58] that reports the average

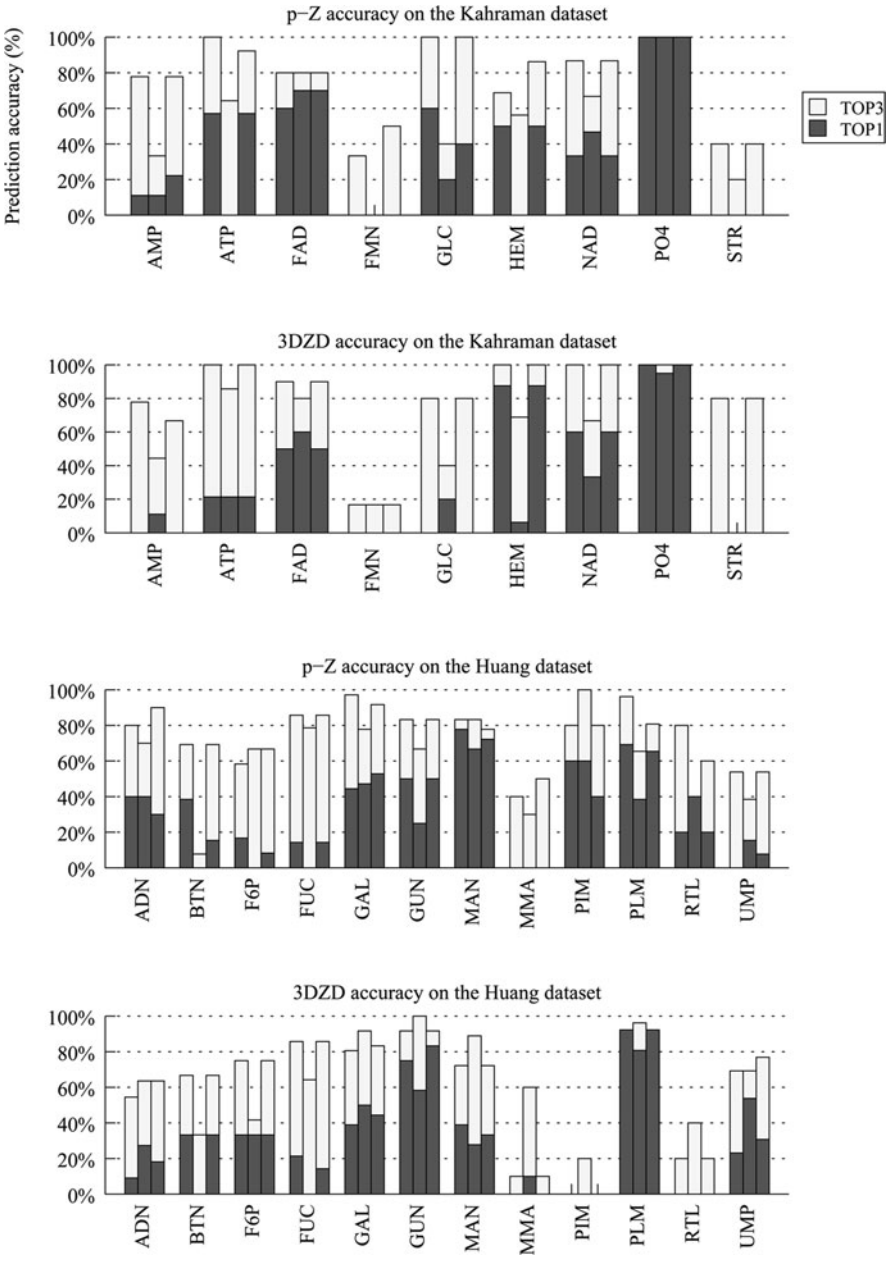

Fig. 4 The Top-1 and Top-3 success rates of binding ligand prediction for individual ligand types in the Kahraman and the Huang dataset. Results are shown for the shape descriptors, the electrostatics descriptors and both combined

RMSD between ligand-bound and ligand-free form is 0.74 Å. Our results indicate that the p-Z and 3DZD descriptors are robust enough with respect to the actual range of conformational difference between ligand-bound and ligand-free forms of binding pockets.

For simulating the second case, the situation where binding pockets are not known beforehand, we examined how well the p-Z and 3DZD descriptors perform with predicted pockets. We predicted pockets by running the LIGSITE [18] program for each protein in the Kahraman dataset and queried against the dataset (thus, dataset of known pockets) [50]. This resulted in a significant deterioration of the performance: the AUC value of the p-Z and the 3DZD dropped from 0.79 to 0.52 (p-Z) and from 0.88 to 0.53 (3DZD). The Top-3 success rates of the 3DZD dropped from 82.7 to 38.9% while for the p-Z it dropped from 77.3 to 41.0%. We note that inaccuracies in binding pocket prediction largely accounts for the unsuccessful retrieval of predicted pockets, hence more accurate prediction methods [17, 20] are likely to improve results.

Computational Time of Pocket-Surfer

We estimated the running times for computing the p-Z and the 3DZD descriptors and searching against a database of binding pockets. For computing the p-Z descriptor of a pocket, pocket projection and p-Z moments computation steps typically take about 10 s [50]. Surface voxelization and 3DZD descriptors are computed in around 40 s. Searching a query descriptor against a database of 100 descriptors takes around 12 milliseconds for the p-Z descriptors and 20 milliseconds for the 3DZD due to different moment orders [50]. By extrapolation to a PDB-scale database, searching a query pocket can be done in about a few seconds with the p-Z and 3DZD. This is significantly faster than the other methods of similar purpose. Hence, the p-Z and the 3DZD pocket descriptors realize real-time pocket database searches, where users can retrieve a search result instantly sitting in front of a computer.

Pocket Comparison with Local Surface 3D Zernike Descriptors

At the last of this chapter, we will briefly describe our recent ongoing development of binding ligand prediction method which considers similarity of local surface regions in pockets. Shape of pockets for the same ligand molecule can significantly vary due to several reasons, including the flexibility of ligand molecules and binding of solvent molecules [8]. Therefore, pockets which bind the same ligand may be better detected by scoring the local similarity of pockets. Comparing local regions of pockets can be done by segmenting the pockets into local patches and comparing the patches separately. The outline of the algorithm of local pocket surface comparison

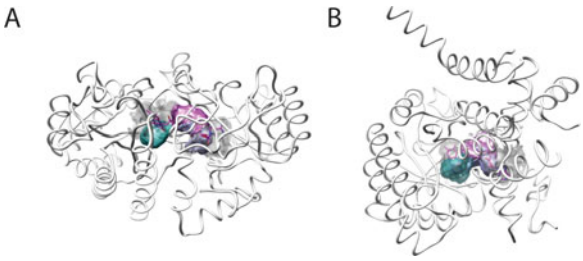

Fig. 5 Local binding site matching of FAD binding proteins. Three local binding patches from **a**, protein 1e8g; and **b**, protein 1k87. Each patch color indicates equivalent position relative to the ligand molecule

method works as follows [59]: First, seed points are evenly distributed on the pocket surface. Then, the shape of surface patch region which is within a sphere centering at each seed point is encoded by the 3DZD. Thus, a whole pocket shape is described as a set of 3DZDs each of which encodes local patch shape. For example, ATP binding pockets are represented as 29.5 overlapping local surface patches on average, while NAD binding pockets have on average 36.8 surface patches of a 5 Å radius. The surface electrostatic potentials and other properties can be also computed in the same manner. To compute the similarity of two pockets, we seek for a set of pairs of surface patches, each taken from the two pockets, which maximizes the overall score for the set. The score will consider the similarity of patches in each pair, the relative position of the patches in each pocket, and the size of the pocket.

Figure 5 shows an example of a pair of FAD binding pockets for which the local pocket surface method yields a better result. Using the global 3DZD, querying the FAD binding pocket of protein 1e8g against the Kahraman dataset retrieved the first FAD binding pocket at the 7th rank (1jqi). In contrast, the local surface comparison method retrieved a FAD binding pocket, 1k87, at the 2nd rank. It is shown in Fig. 5 that the overall pocket shape of 1e8g and 1k87 is quite different because FAD molecule is in a stretched form in 1e8g but bent in 1k87. Despite of the different overall shape, the local patch comparison method could identify the similarity between the two by detecting similarity of the patch pairs shown in the same color.

We have further applied the local protein surface representation by the 3DZD for characterization and classification of protein surface properties [60]. Here, the aim is to annotate entire protein surfaces but not only to compare pocket regions. We extracted local surface patches, which was defined as the surface within a sphere of a 6 Å radius, from 609 representative proteins. This yielded in total of 118,009 patches. A patch was characterized by two features, the shape and the electrostatic potential, and both are described by the 3DZD. We classified the patches using the emergent self-organizing map (ESOM) [61]. The classification resulted in 30–50 clusters of local surfaces of different characteristics. These clusters can be used as surface "alphabet", with which protein surface can be labeled and classified. For example, surface regions of certain biological function, e.g. DNA binding or protein-binding, can be described as a set of the surface alphabets.

Summary

In this chapter, we described moment-based approaches for representing shape of protein surfaces, which are applied for binding ligand prediction by comparing binding pockets. 2D and 3D Zernike moments are able to capture various local protein surface properties of binding pockets. While several other methods exist for binding sites representation and comparison, the moments-based methods benefit from fast computational speed for database search, as well as good retrieval accuracy. However, structure-based function prediction methods are in general vulnerable to structural variability of proteins. To accommodate this problem, we are developing the local pocket surface comparison method where two pockets are compared in terms of matching pairs of local sites.

Comparison of the tertiary structure of proteins, both global and local, is more complicated than comparison of one dimensional protein sequences. Therefore, there have not been as many structure-based methods developed as the sequence-based methods. The p-Z and the 3DZD we introduced in this chapter have potential to change this situation, as they provide very convenient, compact and rotation invariant representation of protein global and local surfaces.

Acknowledgements This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health (R01 GM075004). DK also acknowledges funding from the National Science Foundation (DMS800568, IIS0915801, EF0850009).

References

1. Watson, J.D., Laskowski, R.A., Thornton, J.M. Predicting protein function from sequence and structural data. Curr. Opin. Struct. Biol. **15**: 275–284 (2005).
2. Chandonia, J.M., Brenner, S.E. The impact of structural genomics: expectations and outcomes. Science **311**: 347 (2006).
3. Hawkins, T., Kihara, D. Function prediction of uncharacterized proteins. J. Bioinform. Comput. Biol. **5**: 1–30 (2007).
4. Berman, H.M., et al. The protein data bank. Nucleic Acids Res. **28**: 235–242 (2000).
5. Minai, R., Matsuo, Y., Onuki, H., Hirota, H. Method for comparing the structures of protein ligand-binding sites and application for predicting protein-drug interactions. Proteins: Struct. Funct. Bioinform. **72**: 367–381 (2008).
6. Orengo, C.A., Jones, D.T., Thornton, J.M. Protein superfamilies and domain superfolds. Nature **372**: 631–634 (1994).
7. Li, B., et al. Characterization of local geometry of protein surfaces with the visibility criterion. Proteins **71**: 670–683 (2007).
8. Kahraman, A., Morris, R.J., Laskowski, R.A., Thornton, J.M. Shape variation in protein binding pockets and their ligands. J. Mol. Biol. **368**: 283–301 (2007).
9. Liang, J., Edelsbrunner, H., Woodward, C. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. Protein Sci. **7**: 1884–1897 (1998).
10. Laskowski, R.A., Luscombe, N.M., Swindells, M.B., Thornton, J.M. Protein clefts in molecular recognition and function. Protein Sci. **5**: 2438–2452 (1996).
11. Laskowski, R.A. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. J. Mol. Graph. **13**: 323–328 (1995).

12. Levitt, D.G., Banaszak, L.J. POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. J. Mol. Graph. **10**: 229–234 (1992).
13. Kawabata, T., Go, N. Detection of pockets on protein surfaces using small and large probe spheres to find putative ligand binding sites. Proteins **68**: 516–529 (2007).
14. Weisel, M., Proschak, E., Schneider, G. PocketPicker: analysis of ligand binding-sites with shape descriptors. Chem. Cent. J. **1**: 7 (2007).
15. Hendlich, M., Rippmann, F., Barnickel, G. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. J. Mol. Graph. Model. **15**: 359–363, 389 (1997).
16. Kim, D., et al. Pocket extraction on proteins via the Voronoi diagram of spheres. J. Mol. Graph. Model. **26**: 1104–1112 (2008).
17. Tseng, Y.Y., Dundas, J., Liang, J. Predicting protein function and binding profile via matching of local evolutionary and geometric surface patterns. J. Mol. Biol. **387**: 451–464 (2009).
18. Huang, B., Schroeder, M. LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. BMC Struct. Biol. **6**: 19 (2006).
19. Ota, M., Kinoshita, K., Nishikawa, K. Prediction of catalytic residues in enzymes based on known tertiary structure, stability profile, and sequence conservation. J. Mol. Biol. **327**: 1053–1064 (2003).
20. Capra, J.A., Laskowski, R.A., Thornton, J.M., Singh, M., Funkhouser, T.A. Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. PLoS Comput. Biol. **5**: e1000585 (2009).
21. Laurie, A.T., Jackson, R.M. Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. Bioinformatics **21**: 1908–1916 (2005).
22. Elcock, A.H. Prediction of functionally important residues based solely on the computed energetics of protein structure. J. Mol. Biol. **312**: 885–896 (2001).
23. An, J., Totrov, M., Abagyan, R. Pocketome via comprehensive identification and classification of ligand binding envelopes. Mol. Cell Proteomic. **4**: 752–761 (2005).
24. Porter, C.T., Bartlett, G.J., Thornton, J.M. The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucleic Acids Res. **32**: D129–D133 (2004).
25. Arakaki, A.K., Zhang, Y., Skolnick, J. Large-scale assessment of the utility of low-resolution protein structures for biochemical function assignment. Bioinformatics **20**: 1087–1096 (2004).
26. Ferre, F., Ausiello, G., Zanzoni, A., Helmer-Citterich, M. SURFACE: a database of protein surface regions for functional annotation. Nucleic Acids Res. **32**: D240–D244 (2004).
27. Gold, N.D., Jackson, R.M. Fold independent structural comparisons of protein-ligand binding sites for exploring functional relationships. J. Mol. Biol. **355**: 1112–1124 (2006).
28. Kalidas, Y., Chandra, N. PocketMatch: a new algorithm to compare binding sites in protein structures. BMC Bioinformatics **9**: 543 (2008).
29. Xiong, B., et al. BSSF: a fingerprint based ultrafast binding site similarity search and function analysis server. BMC Bioinformatics **11**: 47 (2010).
30. Binkowski, T.A., Joachimiak, A. Protein functional surfaces: global shape matching and local spatial alignments of ligand binding sites. BMC Struct. Biol. **8**: 45 (2008).
31. Schmitt, S., Kuhn, D., Klebe, G. A new method to detect related function among proteins independent of sequence and fold homology. J. Mol. Biol. **323**: 387–406 (2002).
32. Kinoshita, K., Nakamura, H. Identification of protein biochemical functions by similarity search using the molecular surface database eF-site. Protein Sci. **12**: 1589–1595 (2003).
33. Ramensky, V., Sobol, A., Zaitseva, N., Rubinov, A., Zosimov, V. A novel approach to local similarity of protein binding sites substantially improves computational drug design results. Proteins **69**: 349–357 (2007).

34. Bock, M.E., Garutti, C., Guerra, C. Cavity detection and matching for binding site recognition. Theor. Comput. Sci. **408**: 151–162 (2008).
35. Kazhdan, M., Funkhouser, T., Rusinkiewicz, S. Rotation invariant spherical harmonic representation of 3D shape descriptors. Proc. 2003 Eurographics/ACM SIGGRAPH Symp. Geometry Process. **43**: 156–164 (2003).
36. McDonald, I.K., Thornton, J.M. Satisfying hydrogen bonding potential in proteins. J. Mol. Biol. **238**: 777–793 (1994).
37. Canterakis, N. 3D Zernike moments and Zernike affine invariants for 3D image analysis and recognition. Proceedings of the 11th Scandinavian Conference on Image Analysis, Kanger-Iussuaq, Greenland, pp. 85–93 (1999).
38. Novotni, M., Klein, R. 3D Zernike descriptors for content based shape retrieval. ACM Symposium on Solid and Physical Modeling, Proceedings of the 8th ACM Symposium on Solid Modeling and Applications, pp. 216–225 (2003).
39. Connolly, M.L. Shape complementarity at the hemoglobin alpha 1 beta 1 subunit interface. Biopolymers **25**: 1229–1247 (1986).
40. Dym, H., McKean, H. *Fourier series and integrals*. New York, NY: Academic (1972).
41. Sael, L., Kihara, D. Protein surface representation and comparison: New approaches in structural proteomics. *Biological data mining.* Chen, J., Lonardi, S. (eds.). Kumar, V. (series ed.). Boca Raton, FL: Chapman & Hall/CRC Press, Chapter 3, pp. 89–109 (2009).
42. Venkatraman, V., Sael, L., Kihara, D. Potential for protein surface shape analysis using spherical harmonics and 3D Zernike descriptors. Cell Biochem. Biophys. **54**: 23–32 (2009).
43. Kihara, D., Sael, L., Chikhi, R. Molecular surface representation using 3D Zernike descriptors for protein shape comparison and docking. Curr. Protein Peptide Sci. (2011) (In Press).
44. La, D., et al. 3D-SURFER: software for high-throughput protein surface comparison and analysis. Bioinformatics **25**: 2843 (2009).
45. Sael, L., et al. Fast protein tertiary structure retrieval based on global surface shape similarity. Proteins: Struct. Funct. Bioinform. **72**: 1259–1273 (2008).
46. Sael, L., La, D., Li, B., Rustamov, R., Kihara, D. Rapid comparison of properties on protein surface. Proteins: Struct. Funct. Bioinform. **73**: 1–10 (2008).
47. Venkatraman, V., Chakravarthy, P.R., Kihara, D. Application of 3D Zernike descriptors to shape-based ligand similarity searching. J. Cheminform. **1**: 19 (2009).
48. Venkatraman, V., Yang, Y.D., Sael, L., Kihara, D. Protein-protein docking using region-based 3 D Zernike descriptors. BMC Bioinformatics **10**: 407 (2009).
49. Sael, L., Kihara, D. Protein surface representation for application to comparing low-resolution protein structure data. BMC Bioinformatics **11**:S2 (2010).
50. Chikhi, R., Sael, L., Kihara, D. Real-time ligand binding pocket database search using local surface descriptors. Proteins: Struct. Funct. Bioinform. **78**: 2007–2028 (2010).
51. Bhatia, A.B., Wolf, E. On the circle polynomials of Zernike and related orthogonal sets. Proc. Camb. Philos. Soc. **50**: 40–48 (1954).
52. Zernike, F. Beungungsthoerie des Schneiden-verfahrens und seiner verbesserten Form. Physica **1**: 689–701 (1934).
53. Teh, C.H., Chin, R.T. On image-analysis by the methods of moments. IEEE Trans. Pattern Anal. Mach. Intell. **10**: 496–513 (1988).
54. Laga, H., Takahashi, H., Nakajima, M. Spherical wavelet descriptors for content-based 3D model retrieval. IEEE International Conference on Shape Modeling and Applications (SMI2006), Sendai, Japan, pp. 75–85 (June 2006).
55. Novotni, M., Klein, R. Shape retrieval using 3D Zernike descriptors. Comput. Aided Des. **36**: 1047–1062 (2004).
56. Fawcett, T. An introduction to ROC analysis. Pattern Recognit. Lett. **27**: 861–874 (2006).
57. Kahraman, A., Morris, R.J., Laskowski, R.A., Favia, A.D., Thornton, J.M. On the diversity of physicochemical environments experienced by identical ligands in binding pockets of unrelated proteins. Proteins: Struct. Funct. Bioinform. **78**: 1120–1136 (2010).

58. Brylinski, M., Skolnick, J. What is the relationship between the global structures of apo and holo proteins? Proteins **70**: 363–377 (2008).
59. Sael, L., Kihara, D. Binding ligand prediction for proteins using partial matching of local surface patches. Int. J. Mol. Sci. **11**(12): 5009–5026 (2010).
60. Sael L., Kihara, D. Characterization and classification of local protein surfaces using self-organizing map. Int. J. Knowl. Discov. Bioinfo. **1**: 32–47 (2010).
61. Ultsch, A. Maps for the visualization of high-dimensional data spaces. Proceedings of the Workshop on Self Organizing Maps, Hibikino, Kitakyushu, Japan, pp. 225–230 (2003).