Protein surface descriptors for binding sites comparison and ligand prediction

Rayan Chikhi

Internship report, Kihara Bioinformatics Laboratory, Purdue University, 2007

Abstract. Proteins molecular recognition play an important role in their function. Determining which ligand can bind to a protein is a complex matter due to the nature of protein-ligand interactions and flexibility of binding sites. However, geometric complementarity has often been observed between the ligand and its binding site. Under the assumption that geometrically similar binding sites bind the same ligand, binding sites are mainly studied using three dimensional and graph based representations. In this paper, we present a model for two dimensional ligand binding pockets representation and we apply it to pocket-pocket matching and binding ligand prediction. This model is based on surface mapping of the binding site and makes use of two dimensional Pseudo-Zernike descriptors. Our results show that for certain classes of ligands (HEM, NAD, PO_4), up to 60% of binding sites are correctly predicted to belong to the right class.

1 Introduction

Proteins are large chains of molecules (*amino acids*) and play essential roles in the human body. They are a key part of the immune system, they transport molecules such as oxygen, and are involved in every cellular function. Made of 300 amino acids on average, proteins are large compounds harder to study than molecules. A protein function is often determined by its three dimensional structure, experimentally measured by X-ray crystallography or NMR spectroscopy. An ongoing worldwide effort, the Structural Genomics initiative [1] is solving three dimensional structures of proteins of medical interest, where little to nothing is known about their function. In many cases, a protein is functionally activated by a molecule (*ligand*) binding to it, acting as a switch. Therefore, determining which ligand could bind to a protein is fundamental for protein function identification.

Protein-ligand interactions are known to be based on geometric and electrostatic complementarity. Current methods for comparing binding pockets are focused on geometric properties. Our motivation is to model binding pockets with versatile surface properties, ie. shape, b-factors (correlated with flexibility) and electrostatic potential (though only shape is covered in this paper). Since binding pockets are suitable for a star-like shape approximation, we transform a binding pocket to a spherical function and describe it with two dimensional moments.

In the next section, we give an overview of current research on protein binding sites discovery, matching and comparison. In section 3, we formally describe a new model for binding pockets descriptors based on ray-casting and two dimensional descriptors. In section 4, these descriptors are evaluated for binding pocket comparison and used in the design of a ligand prediction method.

2 Related Work on Ligand Binding Sites

Ligand binding sites have been computationally studied from three main angles: detection of binding pockets using protein surface, match of protein structure against a database of known binding sites patterns, and pocket-pocket comparisons.

2.1 Identification of binding sites

Detection of binding sites consists in predicting where on the protein surface any ligand can bind, which is critical to drug discovery. To the best of our knowledge, binding site detection does not predict which ligand binds to the discovered site.

The most common approach to ligand binding site localization is volumetric search for large cavities, more recent methods also use electrostatic potential and conservation. SURFNET [2] performs a gap search by fitting spheres inside protein convex hull. PocketPicker [3] and LIGSITE [4] methods consist in creating a grid and scanning it for protein-void-protein events in many directions, whereas VisGrid [5] uses visibility of surface points to find pockets. A broad survey of binding pocket detection is presented in [3]. It is worth noting that most successful approach cited in this paper is LIGSITE^{csc}, which achieves an average detection accuracy of 75%.

Another unique binding pocket detection method is local similarity search on protein surface. It consists in using a database of known binding sites and scanning the surface of a protein to find surface patch matches in the database. This was implemented in a very accurate but computationally intensive method using a maximum subgraph algorithm, eF-seek [6].

2.2 Pocket-pocket comparison

Similarity of binding pockets play a crucial role in structural protein function prediction. There is a flood of methods for binding site representation and comparison. Since mechanisms of binding are not yet fully understood, binding sites are commonly defined on geometric criterions.

Among comparison methods, we will restrict our survey to the most popular ones: three dimensional shape matching with spherical harmonics [7], geometric hashing [8] and three dimensional root mean square deviation [9].

Recently, a study on shape variation of binding sites and how they are related to their ligand was published [10], spherical harmonics descriptors were used. The authors concluded that binding sites binding the same ligand show variable shape conformations, and global geometric complementarity alone is not sufficient for molecular recognition.

3 Novel Binding Pockets Descriptors

Since spherical harmonics cannot capture partial shape complementarity, we choose to use an approach where local similarity search has been already well researched: two dimensional descriptors. In this section we describe a novel binding pocket description model based on ray-casting and 2D moments. The binding pocket will be represented as a spherical panoramic picture from its center of gravity, on which we apply descriptors used in content-based image retrieval, Pseudo-Zernike moments.

We first define some terms. The notion of *surface* refers to the Connolly surface [11], commonly used in proteins surface visualization and surface-related computations. We consider a binding pocket (BP) as any connected subset of the protein surface that is not part of the protein convex hull. We define G as the center of gravity of BP, provided it does not lie inside the protein volume; otherwise, G is any of the closest points outside of it. The *opening* of BP is defined as the set of rays starting at G and not intersecting BP.

3.1 Ray-casting of outermost surface

We now describe a ray-casting [12] strategy to represent BP as seen from G. To remove one degree of freedom and later achieve rotation invariance, we make the assumption that a binding pocket orientation is partially defined by its opening. Therefore, our representation is a piecewise continuous surface map relative to a coordinate system defined from BP opening.

A three dimensional cartesian coordinate system $(\overrightarrow{x}, \overrightarrow{y}, \overrightarrow{z})$ specific to BP is defined as follows: origin G, $\overrightarrow{z}$ is a unit vector aligned with the center of mass of all the opening rays. Intuitively, $\overrightarrow{z}$ points toward the pocket opening. The later use of 2D rotationally invariant descriptors enable us to define $(\overrightarrow{x}, \overrightarrow{y})$ arbitrarily.

Using spherical coordinates, we define $f(\theta, \phi)$ as follows: $(\theta, \phi) \in [0, 2\pi], [0, \pi]$ and

$$f(\theta,\phi) = \begin{cases} \max\limits_{i}(d_i) & \text{where a ray starting at G intersects } BP \text{ at distances } (d_i) \text{ from } G \\ 0 & \text{if no intersection occurs.} \end{cases}$$

This can be interpreted as a spherical function describing the outermost surface of BP.

Figure 1 sketches f definition on the intersection of BP with a fictional plane containing G. Since this function is a piecewise continuous spherical function, in order to

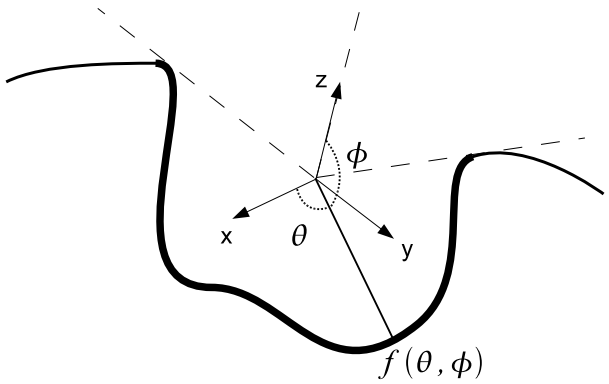

Fig. 1. Mapping of binding pocket (bold line) from its center of gravity. Z axis is aligned with the center of the pocket opening, plane X,Y is arbitrarily oriented.

use two dimensional descriptors f has to be mapped to a plane, the same way the Earth maps are projected.

3.2 Projection and Pseudo-Zernike descriptors

Numerous methods exists for spherical function projection, because no projection can be constructed to preserve spherical properties such as area, shape and distance altogether [13]. We selected a very simple scheme, a special case of equi-rectangular (distance preserving) projection named *plate-carrée* projection. This consists in mapping $f(\theta, \phi)$ a plane where:

$$x = \theta$$
$$y = \phi$$

Experimentally, this projection does not distort shapes of a binding pocket beyond recognition by descriptors. A projected surface of a binding pocket is shown Figure 2.

The next step is to describe the two dimensional BP projection with image moments. Because of the rapid growth of content-based image retrieval, there is a flood of descriptors that can be used to quantify similarity of images. By definition of $f(\theta, \phi)$, only the shape descriptors are of interest. Among them, we choose to use Pseudo-Zernike [14] moments.

The Pseudo-Zernike moments use a set of complete and orthogonal basis functions defined over the unit circle as follows:

$$V_{n,m}(x, y) = e^{jm\theta} \sum_{s=0}^{n-|m|} \frac{(-1)^s (2n + 1 - s)! \rho^{(n-l)}}{s!(n + |m| + 1 - s)!(n - |m| - s)!}$$

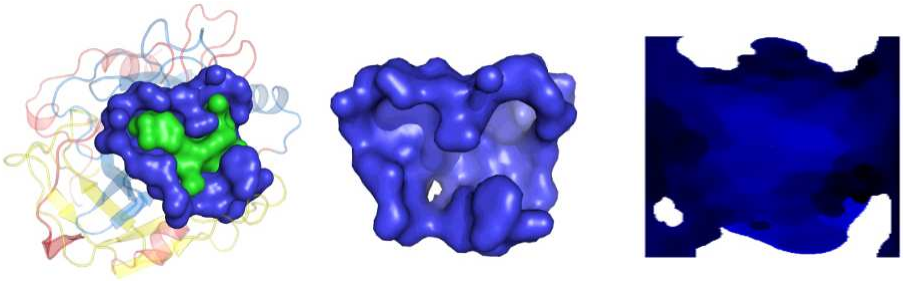

Fig. 2. Overview of the binding pocket representation process: the ligand binding site of a protein (on the left, PDB:1dwd protein) is represented by the whole cavity surface (middle), which is sphere-mapped from its center of gravity and projected (right)

where $\rho = \sqrt{x^2 + y^2}, \theta = tan^{-1}(y/x)$. Pseudo-Zernike moments $(A_{n,m})_{0 \leq n+m \leq i}$ of i-th order are computed for an image $f(x, y)$ with the following formula:

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

This choice was motivated by three reasons. First, comparative studies show that these moments are robust for shape description [15,16], they have been extensively used for face recognition [17]. Second, these moments are rotationally invariant around the center of the image, which is a required property due to the coordinate system we used to model binding pockets. Third, they are orthogonal over the unit circle. In a binding pocket, the active site is likely to be buried inside the cavity, at the opposite to the opening. Due to the adequate position and orientation of our coordinate system, the Pseudo-Zernike basis is likely to capture the active site shape, therefore providing an ideal local similarity criteria.

4 Applications and results

In this section our descriptor model is used to compare actual ligand binding sites and predict which ligand is most likely to bind to a binding pocket, by searching for similar pockets.

4.1 Evaluation of Descriptors

To assess the quality of our descriptors we compare them with spherical harmonics using the protein data set derived in [10], under the Interact Cleft Model. This model has been built with pseudo-spheres within 0.3 Å of proteins atoms interacting with the bound ligand. Since our binding pocket model is related to protein surface, we derived

a similar model named Ligand BP Approximation by keeping protein atoms within 5 Å of the bound ligand (distance was experimentally chosen in order to include atoms in a buffer region). Pseudo-Zernike moments are computed to the 7th order on the binding pocket model described in Section 3.

Then, Pseudo-Zernike moments are compared with spherical harmonics moments using all-against-all distance matrices, shown Figure 3. It appears that our model is able to reflect similarity of sites from the same ligand set (green dots in the diagonal square) while suffering from a very low specificity (green dots also appear outside diagonal squares). Oppositely, Interact Cleft Model (spherical harmonics) is able to separate PO_4 sites from every other ligand except GLC, but is not able to capture similarity of any other type.

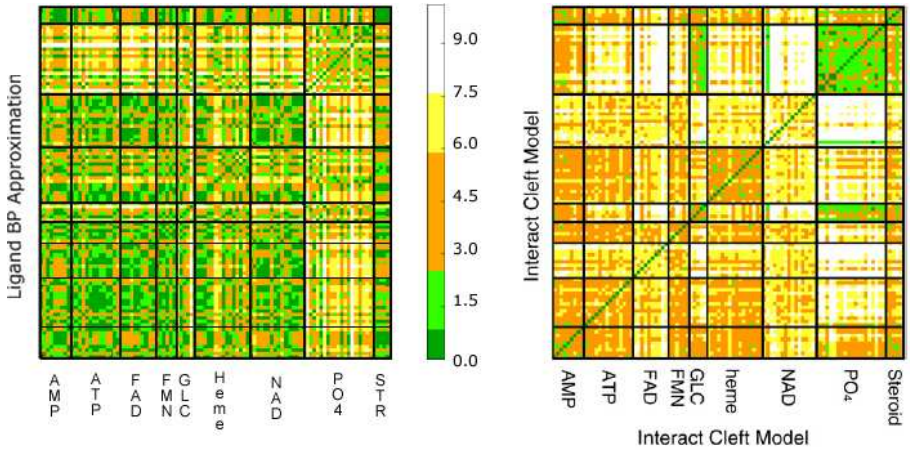

Fig. 3. All-against-all distances matrices of Pseudo-Zernike (left) and spherical harmonics (right, reproduced from [10]) descriptors representing shape similarity of binding pockets from the Thornton-Kahraman [10] protein set. A green dot reflects low distance between two descriptors, whereas orange-yellow reflects high distance. The actual color scale for Pseudo-Zernike descriptors is shown.

A closer inspection of the distance matrix reveals that, even if most coefficient distances are not clearly reflecting separation of ligand families, most of the dark green dots are often in the diagonal square. In the next subsection we design a scoring function that predicts binding pocket ligand type based on this observation.

4.2 Binding ligand prediction from pocket shape

We present a framework to predict the binding site ligand type given a query pocket and a database of binding sites.

Our approach for binding ligand prediction is based on the assumption that, given a query binding pocket, sites binding the same ligand are likely to often show among the k closest pockets in the database. Therefore, even if the closest match is not a binding pocket of the same type, we examine the $k = 20$ nearest neighbors out of $n = 100$ binding pockets, and give a score to every ligand. The scoring function is defined for a ligand F regarding ranks and proportion of pockets binding F:

$$score(F) = \sum_{i=0}^{20} (\mathbb{1}_{l(i)}(F) log(\frac{n}{i})) \frac{\sum_{i=0}^{20} \mathbb{1}_{l(i)}(F)}{\sum_{i=0}^{n} \mathbb{1}_{l(i)}(F)}$$

where $l(i)$ returns ligand type of the i-th nearest neighbor. The ligand with the highest score is predicted to bind with the query pocket.

We applied this scoring function to predict the ligand of every binding site from the Thornton-Kahraman data set, using the remaining of the data set as the reference data. The results are shown Table 1 in two lines, Top 1 is for highest scoring ligand being the correct binding ligand, Top 3 allows the correct answer to lie in the first three highest scoring ligands.

Ligand	AMP	ATP	FAD	FMN	GLC	HEM	NAD	PO_4	STR
Top 1						43.8%	46.7%	60%	
Top 3		50%	60%		60%	43.8%	66%	70%	50%

Table 1. Success rate for binding ligand prediction using our binding pocket model, 7-th order Pseudo-Zernike moments and the scoring function. The protein database we used is the Thornton-Kahraman data set (100 proteins, each binding one of the 9 ligands shown on table). Top 1 means that the correct prediction is the highest scoring ligand, Top 3 extends to the second and third highest scoring ligands. Scores under 15% for Top 1 and 33% for Top 3, corresponding to random predictions are not shown.

Our prediction method performs well at identifying the PO_4 ligand, which is due to a clear separation already shown in the distance matrix Figure 3. HEM is also known as a rigid ligand with similar binding pocket shapes, however NAD is flexible but the overall shape of the pocket is well preserved, as most of the closest matches belong to the same ligand type. Top 1 (resp. 3) predictions that scored higher than 11% (resp. 33%) are superior than a random classifier, which has one (resp. three) chances out of 9 to predict the right ligand.

5 Conclusion

In this paper, we design a representation of ligand binding sites using a planar-projected spherical function as an input for two dimensional Pseudo-Zernike descriptors. These descriptors are applied to an actual data set and compared with spherical harmonics. Both show an average selective power on most ligands in the set.

As a proof of concept, we explicit a method to predict which ligands are most likely to bind to a binding site, using a similarity search with known ligand binding sites. This method successfully predicts a correct result for 60% of the PO_4 binding sites, and finds the correct ligand among the best 3 predictions of most ligands with average 60% success rate.

Future directions of research include adapting surface maps to other protein surface properties such as b-factors and electrostatic potential, and combining descriptors to achieve a clearer separation of sites binding different ligands.

References

1. Chandonia, J.M., Brenner, S.E.: The Impact of Structural Genomics: Expectations and Outcomes. Science **311**(5759) (2006) 347–351
2. A, L.R.: Surfnet: A program for visualizing molecular surfaces, cavities and intermolecular interactions. J. Mol. Graph **13** (1995) 323–330
3. Weisel M, Proschak E, S.G.: Pocketpicker: Analysis of ligand binding-sites with shape descriptors. Chemistry Central Journal **1** (2007) 7
4. Huang, B., Schroeder, M.: Ligsitecsc: Predicting ligand binding sites using the connolly surface and degree of conservation. BMC Structural Biology **6** (September 2006) 19+
5. Bin Li, Srinivasan Turuvekere, M.A.D.L.K.R..D.K.: Characterization of local geometry of protein surfaces with the visibility criterion. Proteins: Struct. Funct. Bioinformatics (2007)
6. Kinoshita, K., Nakamura, H.: Identification of the ligand binding sites on the molecular surface of proteins. Protein Sci **14**(3) (2005) 711–718
7. Morris, R.J., Najmanovich, R.J., Kahraman, A., Thornton, J.M.: Real spherical harmonic expansion coefficients as 3d shape descriptors for protein binding pocket and ligand comparisons. Bioinformatics **21**(10) (2005) 2347–2355
8. Gold, N.D., Jackson, R.M.: Fold independent structural comparisons of protein-ligand binding sites for exploring functional relationships. Journal of Molecular Biology **355**(5) (February 2006) 1112–1124
9. Ferre, F., Ausiello, G., Zanzoni, A., Helmer-Citterich, M.: SURFACE: a database of protein surface regions for functional annotation. Nucl. Acids Res. **32**(suppl_1) (2004) D240–244
10. Kahraman, A., Morris, R., Laskowski, R., Thornton, J.: Shape variation in protein binding pockets and their ligands. J Mol Biol (2007)
11. Connolly, M.L.: Shape complementarity at the hemoglobin a1b1 subunit interface. Biopolymers **25**(7) (1986) 1229–1247
12. Roth, S.D.: Ray casting for modeling solids. Computer Graphics and Image Processing **18**(2) (February 1982) 109–144

13. Snyder, J.: Map Projections-A Working Manual. United States Government Printing (February 1983)
14. Zernike, F.: Beugungstheorie des schneidenverfahrens und seiner verbesserten form, der phasenkontrastmethode. Physical
15. Mehtre, B.M., Kankanhalli, M.S., Lee, W.F.: Shape measures for content based image retrieval: A comparison. Information Processing & Management **33**(3) (May 1997) 319–337
16. Zhang, D., Lu, G.: Content-based shape retrieval using different shape descriptors: A comparative study. icme **00** (2001)
17. Wee, C.Y., Paramesran, R.: On the computational aspects of zernike moments. Image Vision Comput. **25**(6) (2007) 967–980