JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 16, Number 11, 2009

Pp. 1577–1591
DOI: 10.1089/cmb.2009.0045

Binding Balls: Fast Detection of Binding Sites Using a Property of Spherical Fourier Transform

MATTEO COMIN,[1] CONCETTINA GUERRA,[1] and FRANK DELLAERT[2]

ABSTRACT

The functional prediction of proteins is one of the most challenging problems in modern biology. An established computational technique involves the identification of three-dimensional local similarities in proteins. In this article, we present a novel method to quickly identify promising binding sites. Our aim is to efficiently detect putative binding sites without explicitly aligning them. Using the theory of Spherical Harmonics, a candidate binding site is modeled as a Binding Ball. The Binding Ball signature, offered by the Spherical Fourier coefficients, can be efficiently used for a fast detection of putative regions. Our contribution includes the Binding Ball modeling and the definition of a scoring function that does not require aligning candidate regions. Our scoring function can be computed efficiently using a property of Spherical Fourier transform (SFT) that avoids the evaluation of all alignments. Experiments on different ligands show good discrimination power when searching for known binding sites. Moreover, we prove that this method can save up to 40% in time compared with traditional approaches.

Key words: algorithms, alignment, protein motifs, protein structure.

1. INTRODUCTION

THE PREDICTION OF INTERACTIONS between proteins and ligands is of great interest to the functional annotation of proteins. In applications such as the design of drugs, computational techniques have proven useful to filter out possible three-dimensional arrangements of proteins' complexes. A variety of geometric methods for shape representation and recognition, often developed within other fields, have been applied to solve many instances of the protein structural comparison problem (Artymiuk et al., 2005; Ballester and Richards, 2007; Bock et al., 2007a; Shulman-Peleg et al., 2004; Kinoshita et al., 2001; Sommer et al., 2007; Weskamp et al., 2004).

In this article, we present a novel method to quickly identify promising binding sites, either in a protein cavity or on an entire protein surface. Our aim is to efficiently detect putative binding sites without explicitly aligning them, i.e., *without* actually computing the optimal rotation that best overlaps two binding sites. Instead, with our method we are able to simultaneously evaluate all possible rotations corresponding to a single translation.

[1]Department of Information Engineering, University of Padova, Padova, Italy.
[2]College of Computing, Georgia Tech, Atlanta, Georgia.

To represent a given binding site, we introduce the notion of *Binding Ball*, a spherical description of a query binding site, a concept tightly related to Connolly's representation of molecular surfaces (Connolly, 1983). We define a Binding Ball as a sphere with an associated spherical function representing the molecular surface.

After creating the Binding Ball for a given query binding site, we "roll" it over a protein's surface to be examined, and our main contribution is in the way we can quickly score and select those candidate positions. The Binding Ball's definition is particularly suitable to assess the similarity between the spherical functions representing a binding site and a selected set of atoms, which is traditionally assessed by a scoring function based on the three-dimensional alignment of their atoms or surfaces. This may involve an exhaustive search in the space of all possible rotations and translations that might be computationally demanding. To avoid this search, we design a comparison method to efficiently evaluate the similarity between two binding sites without explicitly computing the best rotation.

While rolling the Binding Ball over a protein's surface, we are able to efficiently score each position, evaluating all possible rotations at that location simultaneously, by making use of a specific property of the Spherical Fourier Transform (SFT). The key insight is that all possible rotations R can be evaluated simultaneously by making use of the properties of the SFT. Informally, the SFT allows any spherical function to be expanded as $f(\theta) = \sum_{l \in N, m \in [-l, l]} c_m^l Y_m^l(\theta)$, where c_m^l are the spherical harmonic coefficients. A similarity score, namely correlation, between two spherical functions for a given rotation R can be computed using the SFT. In this article, we are not interested in the rotation that produces the highest correlation, but in the integration of correlations for all possible rotations R. By exploiting a property of the SFT, we prove that the former is computationally expensive and that the latter is a combination of spherical harmonic coefficients. Hence, a fast screening of the candidate region is possible, resulting in a list of putative regions that are found similar with the query binding site.

The spherical Fourier transformation has been already used in a variety of different contexts, including the alignment of Binding Sites (Cai et al., 2002; Leicester et al., 1994a,b; Makadia and Daniilidis, 2006; Ritchie and Kemp, 1999, 2000); see Section 2 for a more comprehensive discussion. In almost all applications, the rotation that best overlaps the two functions is computed explicitly from the spherical harmonic representation. In this article, we do not align binding sites, but we want to efficiently discover regions that are potentially similar to a given binding site.

Experiments on different ligands show good discrimination power when searching for known binding sites using the average correlation. Moreover, we prove that this method can save up to 40% in time compared with traditional approaches.

2. RELATED WORK

It has long been recognized that geometry plays an important role in structural bionformatics. Several matching strategies have been employed for binding site recognition and classification, including hashing techniques (Shulman-Peleg et al., 2004; Weskamp et al., 2004; Chen et al., 2005b,a), graph-theoretic methods (Artymiuk et al., 2005; Hofbauer et al., 2004; Kinoshita et al., 2001; Shatsky et al., 2006; Weskamp et al., 2004), moment invariants descriptors (Ballester and Richards, 2007; Sommer et al., 2007), shape descriptors such as spin images and shape contexts (Bock et al., 2007a,b, 2008), and clustering (Kuttner et al., 2003). Some approaches combine sequence and structure information (Barker and Thornton, 2003; Binkowski et al., 2003, 2005; Yao et al., 2003), whereas others integrate structural information with physico-chemical properties (Hofbauer and Aszodi, 2005; Kinoshita et al., 2001; Minai et al., 2008; Shulman-Peleg et al., 2004). Some of these techniques have been made available over the web (Ausiello et al., 2008; Jambon et al., 2005; Kinoshita et al., 2001).

Below, we concentrate on reviewing approaches to protein binding site comparison and localization based on spherical harmonics. For a comprehensive survey of other methods, see Laurie and Jackson (2006). The expansion coefficients can be used as a feature vector or signature for describing the shape. Properties of the spherical harmonics relevant to our application are presented in Section 3.2.

As global shape descriptors, the spherical harmonics are suitable to represent closed surfaces, such as those of globular proteins. They have long been used in biochemistry and computational biology for protein modeling, visualization, and comparison. An interesting feature is that they allow the representation of geometry as well as of physico-chemical properties of the molecules, such as hydrophobicity and elec-

trostatic properties. One of the first applications of spherical harmonics to protein rendering is found in Max and Getzoff (1988); although this article does not apply this representation to comparing proteins, it does however mention that structural similarity of two biomolecules can be expressed as the distance of their spherical harmonic coefficients. In Leicester et al. (1988), this idea is further explored and applied to molecular docking. The method evaluates the quality of the fit of candidate molecular interfaces; however, it does not automatically search for possible locations of geometrically complementary surfaces.

Spherical harmonics cannot be directly applied to the problem of binding site recognition. This problem requires us to analyze and extract local shape features of a protein surface, and it is not straightforward to do that using a global shape descriptor. The first method using the spherical harmonic representation for binding site recognition is presented in Cai et al. (2002). The binding site is represented by a single sphere, placed in the center of the cavity where the binding site resides and then "inflated" until it approximates the shape of the cavity. Their computational approach to compare two spherical functions is the same as that in Ritchie and Kemp (1999, 2000).

In the above applications, spherical harmonics are restricted to single valued functions; i.e., it is assumed that each ray of the unit sphere intersects the molecular surface at one point only. This restriction is removed in Duncan and Olson (1993), where a protein is represented as an elastic surface and modeled by the continuous deformation into a spherical shape.

In Leicester et al. (1994a,b), protein shapes are classified based on the L2 distance in coefficient space. Basically the same idea was used in Morris et al. (2005) and Stockwell and Thornton (2006) to compare and cluster protein binding sites. There, a registration phase is used to align two binding sites prior to comparing them, thus avoiding the determination of the rotation for the spherical harmonic representation at the expense of possible loss of accuracy. Neither approach offers a systematic way of computing distances under rigid transformations, although it is recognized that this is a crucial aspect of the calculation.

In a series of papers, Ritchie and Kemp (1999, 2000) and Ritchie et al. (2008) propose an approach to the docking problem using spherical Fourier correlation. They systematically search for complementary sites on two proteins and exploit the properties of the SFT to efficiently search the space of all possible rotations. In particular, rotations of a molecular surface are obtained by mixing the harmonic coefficients according to a matrix representation of the rotation (Kostelec and Rockmore, 2008). These same ideas have been used in the area of computer vision to compute the correlation of two images as a function of rotations (Makadia and Daniilidis, 2006). Functions defined on the SO(3) group of rotations have fast SO(3)-Fourier transforms (called SOFT) (Kostelec and Rockmore, 2008), which are closely related to the spherical harmonic expansions. An alternate method based on spherical harmonics (Kazhdan et al., 2003) represents a shape by means of a shape signature that is invariant under rigid transformations, thus avoiding the explicit exploration of the rotation space when comparing two 3D objects.

3. BINDING BALL METAPHOR

We model a query binding site (*BS*) as a Binding Ball (*BB*). A Binding Ball is a sphere that touches the protein surface at two points and has its center on one of the normals at the two points. A Binding Ball is constructed such that no surface points are contained in it. Thus, the concept of Binding Ball is somehow similar to the Connolly's representation which uses a probe's radius typically smaller than that of a *BB* because of the different use. For every Binding Ball, we associate to it the set of surface points that are within a given distance t_{bb} from its center. A spherical function f that describes these surrounding points is associated to the *BB*. This spherical representation f will be used to compare the Binding Ball with a candidate region by resorting to Spherical Harmonics.

3.1. Binding ball construction

We now describe more formally the construction of a query Binding Ball. Given a point $p \in BS$ with normal n, the Binding Ball tangent to p is constructed as follows. For any other point $p' \in BS$, $p' \neq pp$, we determine the sphere that touches the molecular surface at p and p' and is centered on the normal n of p. The sphere associated to the point p is chosen as the one with the smallest radius over the spheres constructed for all points p'. Let r_p be such minimum radius. Note that this ensure that the Binding Ball is tangent to p, it touches some p', and it does not include any other point. If r_p is smaller than a given

threshold r_{bb}, then the sphere is discarded and no binding ball is associated to p. Otherwise, the Binding Ball obtained is associated to the point p, and we create the spherical function f from all surface points that are within distance t_{bb} from its center. This procedure is repeated for all points in BS, and among all the BBs thus created we choose the one that captures all atoms of the query binding site.

Note that the selected BB will be placed approximately at the center of the binding site. Moreover, the radius of the BB mimics the ligand volume, and this is of great importance since the BB will be rolled looking for similar regions. After an extensive evaluation, we determined r_{bb} such that the ligand volume is comparable with the BB volume; consequently the number of putative BBs is small, and t_{bb} is such that the number of atoms per BB can cover entirely the binding site. In our experiments, using as thresholds $r_{bb} = 2\mathring{A}$ and $t_{bb} = 10\mathring{A}$, the number of BBs in the cavities considered ranges between 20 and 200.

We recall that the spherical function f associated to the BB is computed for the set of surface points that are within a distance t_{bb} from its center. We compute f using the discretized spherical coordinates for the set of points associated to the BB in the following way. We consider a tessellation of the unit sphere with $O(B^2)$ bins, where B is called the bandwidth. For each bin of the sphere's tessellation, we project the radius passing by the bin's center to the protein's surface and compute the closest point to that radius. The bin's value is the distance between this point and the BB's center. The set of points in this spherical coordinate system produces the function f that will be used to compute the Spherical Harmonics coefficients.

3.2. Spherical harmonics

Given the spherical function f, that represents the atoms of the Binding Ball, here we want to compute the Spherical Harmonics coefficients $\hat{f}$ of f. These coefficients, $\hat{f}$, will be used as a signature for a fast comparison of the function f with other candidate regions. Below we review spherical harmonics following Healy et al. (2003), Makadia and Daniilidis (2006), and the references therein. The SFT is the equivalent on the sphere of traditional Fourier analysis. In particular, a given spherical function on the sphere S^2 is approximated by a truncated series of basis functions $Y_m^l : S^2 \rightarrow C$, commonly known as the *Spherical Harmonics*. The $(2l+1)$ spherical harmonics of degree l are given by:

$$Y_m^l(\theta, \phi) = k_{l,m} P_m^l(cos\theta) e^{im\phi}, \quad m = -l, \ldots, l$$

where P_m^l are the associated Legendre functions, and the normalization constants $k_{l,m}$ are chosen to satisfy the orthogonality of spherical harmonics:

$$k_{l,m} = (-1)^m \sqrt{\frac{(2l+1)(l-m)}{4\pi(l+m)}}$$

Thus, any function $f(\eta) \in \mathcal{B}^2(S^2)$ can be written as:

$$f(\eta) = \sum_{l=1}^{B} \sum_{m=-l}^{l} \hat{f}_m^l Y_m^l(\eta) \text{ where } \hat{f}_m^l = \int_{\eta \in S^2} f(\eta) \overline{Y_m^l}(\eta) d\eta$$

where $\hat{f}_m^l$ are the Fourier coefficients and B is the bandwidth that specifies the largest term of the expansion corresponding to non zero coefficients. The coefficients can be computed in $O(B^2 \log^2 B)$ (Makadia and Daniilidis, 2006).

3.3. Binding site recognition

Here we exploit the two parts that constitute the Binding Ball metaphor, i.e., construction and rolling, for the detection of putative binding sites. In Figure 1a, we show the binding site of the protein 1csn, along with the Binding Ball whose associated set of points coincides with the binding site. In Figure 1b, we show an example of how a Binding Ball is used in a query. The Binding Ball is rolled over a protein's surface and its function f is compared with the surrounding. These two operations, construction and roll, constitute the Binding Ball metaphor. In the following we address the fast comparison of the function f of the query Binding Ball with the selected atoms, in yellow. Similarly to the Binding Ball construction, this selection includes all atoms that are within a distance t_{bb} from the BB center t. A spherical function $g(t)$, representing the selection associated to the position t (in yellow), is computed. The BB rolling and screening is now simply the comparison of two spherical functions, f and $g(t)$.

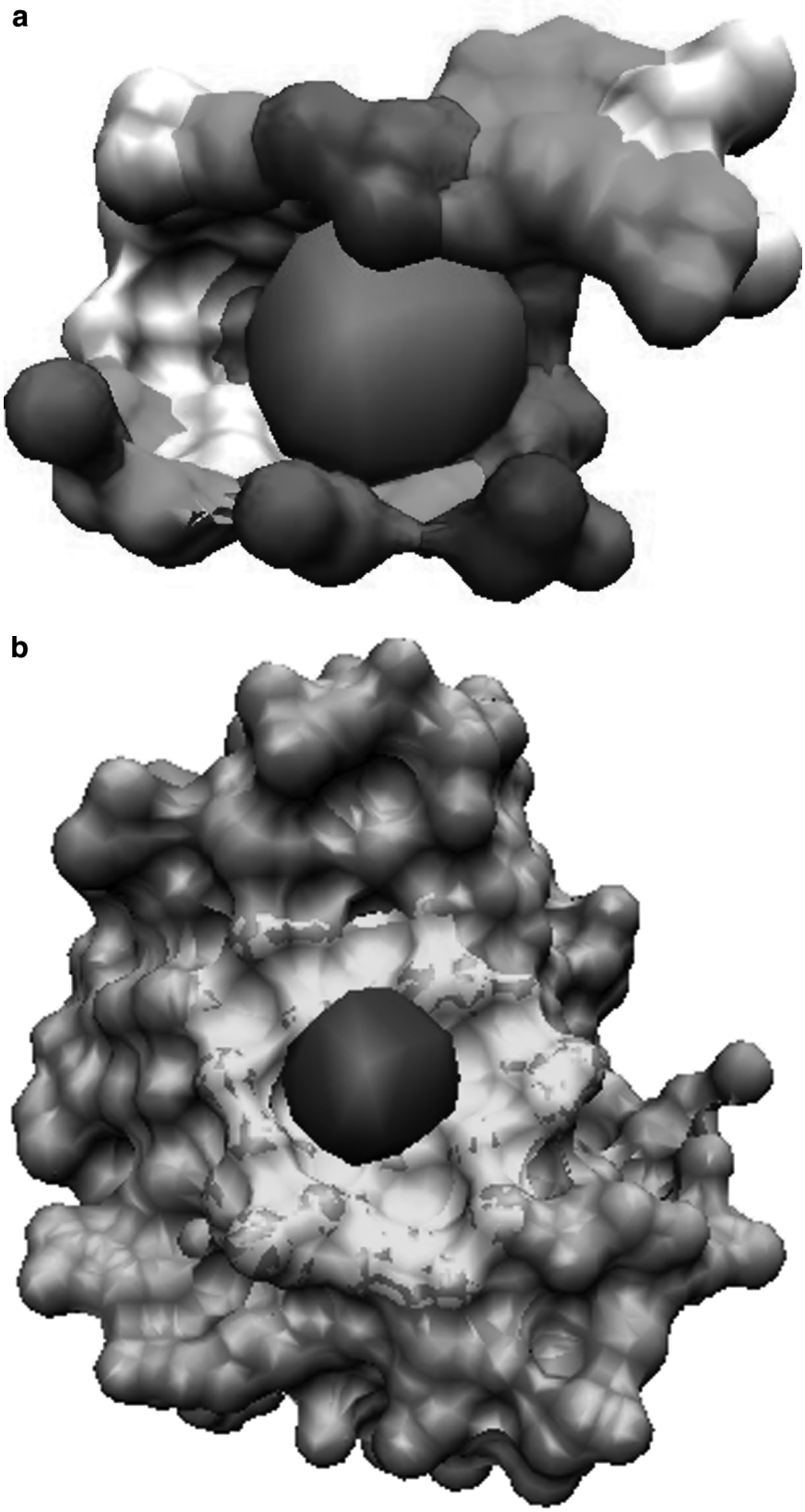

FIG. 1. (**a**,**b**) Examples of construction and use for the Binding Ball of Protein 1csn.

3.4. Scoring with explicit alignment

The SFT can be used to efficiently compute the *correlation* of two spherical functions, f and $g(t)$, in S^2. Given two functions in S^2, $f(\eta)$ and $g(\eta,t)$, the correlation between $f(\eta)$ and $g(\eta,t)$ for a given rotation $R \in SO(3)$ is defined as:

$$s(R,t) = \int_{\eta \in S^2} g(\eta,t) f(R^T \eta) d\eta$$

where R^T is the transpose of R. This correlation is a function defined on the group $SO(3)$ of rotations; thus, it has fast $SO(3)$-Fourier (or SOFT) transform (Kostelec and Rockmore, 2008) that can be computed by pointwise multiplications of the spherical harmonic coefficients of the two functions. A fast inverse SOFT transform then reconstructs the original function $s(R,t)$.

Finding the rotation with the maximum value of correlation is expensive. A complete analysis of its time complexity is presented in Makadia and Daniilidis (2006). They prove that $s(R,t)$ can be derived in three steps: first compute the spherical harmonics coefficients for the two functions $f(\eta)$ and $g(\eta,t)$ in time $O(B^2 \log^2 B)$; then combine the coefficients in time $O(B^3)$; finally compute the inverse SOFT transform in time $O(B^3 \log^2 B)$. Clearly, the time complexity of the overall procedure is $O(B^3 \log^2 B)$ and is dominated by the last term corresponding to the inverse SOFT transform.

3.5. Fast scoring

Here we address the problem of binding site detection given a query binding site, using a similarity measure that does not require the explicit alignment of the two surfaces.

In our setting, the correlation between two functions, f and $g(t)$, represents the similarity of the Binding Ball with the candidate position t for a given rotation R. Since our objective is to efficiently screen all possible position t and the associated $g(t)$, irrespective of the rotation R generating the maximum value of the correlation, we can marginalize out the unknown rotation. Thus, we propose the use of a different scoring function that avoids the complexity of the computation of the maximum correlation over all rotations. As we will show in Section 4, it also yields better results than the maximum correlation. We evaluate the similarity between the BB and the position t using the score:

$$Score(t) = \int_{R \in SO(3)} s(R,t)^2 dR = \int_{R \in SO(3)} s(R,t)\overline{s(R,t)} dR$$

where the integral is taken over all rotations and $\overline{s(R,t)}$ denotes the complex conjugate of $s(R,t)$. Since we are dealing with real functions f and $g(t)$ then the correlation $s(R,t)$ is also real and $s(R,t) = \overline{s(R,t)}$. Although somewhat improperly, from now on we will refer to this score as the *average correlation*. This score can be computed efficiently by exploiting the properties of the SFT. In particular, we prove that

$$\int_{R \in SO(3)} s(R,t)\overline{s(R,t)} dR = \sum_{l=1}^{B} \sum_{m=-l}^{l} \sum_{p=-l}^{l} \hat{s}^l_{mp} \overline{\hat{s}^l_{mp}}$$

where $\hat{s}^l_{mp}$ are the coefficients of the $SO(3)$ Fourier Transform of $s(R,t)$.

Proof. We know that $\hat{s}^l_{mp} = \overline{\hat{g}^l_m} \hat{f}^l_p$ and that $s(R,t)$ can be expanded as

$$s(R,t) = \sum_{l=1}^{B} \sum_{m=-l}^{l} \sum_{p=-l}^{l} \overline{\hat{g}^l_m} \hat{f}^l_p U^l_{pm}(R)$$

where the set of functions U^l_{pm} are a basis for functions defined on $SO(3)$. We recall that $U^l_{pm}(\alpha,\beta,\gamma) = e^{-ip\gamma} P^l_{pm}(cos(\beta)) e^{-im\alpha}$ where P^l_{pm} are generalized associated Legendre polynomials. By substitution we can write

$$\int_{R \in SO(3)} \sum_{l=1}^{B} \sum_{m=-l}^{l} \sum_{p=-l}^{l} \sum_{l_2=1}^{B} \sum_{m_2=-l_2}^{l_2} \sum_{p_2=-l_2}^{l_2} \overline{\hat{g}^l_m} \hat{f}^l_p \hat{g}^{l_2}_{m_2} \overline{\hat{f}^{l_2}_{p_2}} U^l_{pm}(R) \overline{U^{l_2}_{p_2 m_2}(R)} dR$$

since the coefficients $\hat{g}_m^l$ and $\hat{f}_p^l$ are rotation independent we can move them outside the integration, as in

$$\sum_{l=1}^{B}\sum_{m=-l}^{l}\sum_{p=-l}^{l}\sum_{l_2=1}^{B}\sum_{m_2=-l_2}^{l_2}\sum_{p_2=-l_2}^{l_2}\overline{\hat{g}_m^l}\hat{f}_p^l\hat{g}_{m_2}^{l_2}\overline{\hat{f}_{p_2}^{l_2}}\int_{R\in SO(3)}U_{pm}^l(R)\overline{U_{p_2m_2}^{l_2}(R)}dR$$

Given the orthogonality of the functions $U_{pm}^l(R)$

$$\int_{R\in SO(3)}U_{pm}^l(R)\overline{U_{p_2m_2}^{l_2}(R)}dR=\delta_{l,l_2}\delta_{m,m_2}\delta_{p,p_2}$$

this yields non-zero terms only when $l=l_2$, $m=m_2$ and $p=p_2$, implying that almost all terms of the above summations are zero. The remaining ones can be expressed as:

$$\sum_{l=1}^{B}\sum_{m=-l}^{l}\sum_{p=-l}^{l}\overline{\hat{g}_m^l}\hat{g}_m^l\hat{f}_p^l\overline{\hat{f}_p^l}=\sum_{l=1}^{B}\sum_{m=-l}^{l}\sum_{p=-l}^{l}\hat{s}_{mp}^l\overline{\hat{s}_{mp}^l}$$

Hence the claim. ■

From the above proposition, it follows that the average correlation can be computed in time $O(B^3)$. By contrast, the traditional approach to compute the maximum correlation requires $O(B^3\log^2 B)$, as reported in Section 3.2.

3.6. Algorithm summary

The input is the set of the surface points (Connolly's points) of the binding site *BS* of protein *A* and of the largest cavity *C* of protein *B*. The binding site *BS* is represented by a single binding ball *BB*, also called Binding Ball query.

Outline of the algorithm is as follows:

1. Generate the query Binding Ball, *BB*, for protein *A* and compute the associated function *f*.
2. Roll *BB* over *C*. For each position *t* compare *f* with *g*(*t*) using as score the average correlation, *Score*(*t*), of the two associated spherical functions.
3. Return the *k* top-ranking positions *t*, and the selections of atoms that generated *g*(*t*).

After constructing the query Binding Ball, *BB*, we scan the cavity *C* by rolling and scoring the *BB* over *C*. Similarly to the construction of *BB*, for every point $t \in C$ we place the *BB* on the point's normal. If it does not intersect with any other surface's point, we select a set of neighboring points within a distance t_{bb} and compute *g*(*t*). We compare *f* against all *g*(*t*) using the procedure based on spherical harmonics described above. The positions *t*, and the associated sets of points, that produce the highest scores are selected as candidate binding sites.

In the next section, we will exploit this algorithm to retrieve similarity information for a query binding ball in a dataset of proteins binding known ligands. We will also evaluate the performance of our approach in terms of the speed up with respect to the standard reconstruction of *s*(*R*,*t*).

4. RESULTS

In the experiments discussed below, we use the measures of coverage and accuracy to evaluate the performance of our method for binding site detection. The *coverage* is defined as the percentage of residues of a binding site that is found in our solution. The *accuracy* is the percentage of the residues in the solution that are in the binding site. Better performances correspond to larger coverage and accuracy values.

4.1. Detection of binding sites of ligand ATP

In our first experiments, we considered a set of 21 proteins all binding ligand ATP and use as query the active site of the Catalytic Subunit of cAMP-dependent Protein-Kinase (pdb code 1atp, chain E) also binding ATP. The target proteins are from different families according to the structural classification SCOP (Murzin et al., 1995) (Table 1, column 4). Comparison of the sequence and structure of 1atp with the other

Table 1. Sequence Identities and RMSD of the Proteins of the Dataset when Compared to 1atp

Protein	*RMSD*	*Seq. id.*	*Family*
1a49	5.6 Å	2.5%	Pyruvate kinase beta-barrel domain
1a82	4.7 Å	5.6%	Nitrogenase iron protein-like
1ayl	7.1 Å	5.0%	PEP carboxykinase C-terminal domain
1csn	2.4 Å	19.0%	Protein kinases, catalytic subunit
1b8A	6.2 Å	3.8%	PEP carboxykinase C-terminal domain
1e2q	5.7 Å	4.5%	Nucleotide and nucleoside kinases
1e8xA	4.4 Å	7.6%	Phoshoinositide 3-kinase (PI3K) helical domain
1f9aC	6.0 Å	6.2%	Adenylyltransferase
1f9w	7.1 Å	5.7%	Motor proteins
1g5t	5.5 Å	0.0%	RecA protein-like (ATPase-domain)
1gn8A	5.6 Å	6.2%	Adenylyltransferase
1hck	1.9 Å	29.6%	Protein kinases, catalytic subunit
1j7k1	5.2 Å	5.6%	Helicase DNA-binding domain
1jjv	6.4 Å	7.5%	Nucleotide and nucleoside kinases
1kay	6.1 Å	5.7%	Actin/HSP70
1kp2A1	5.6 Å	7.5%	N-type ATP pyrophosphatases
1mjhA	6.1 Å	4.7%	Universal stress protein-like
1nsf	4.8 Å	8.3%	Extended AAA-ATPase domain
1phk	1.4 Å	34.3%	Protein kinases, catalytic subunit
1yag	7.6 Å	3.8%	Actin/HSP70

Column 4 indicates the family of a protein according to SCOP.

proteins of the set shows significant variation among overall sequence identities (from 0% to 34.3%) as well as root mean square deviation (RMSD) of all $C\alpha$ atoms (from 1.4 to 7.6 Angstroms). The sequence identities and RMSD values shown in Table 1 have been determined using the alignment method *Combinatorial Extension* (CE) applied to the complete structures (Shindyalov and Bourne, 1998). Of the set of proteins only three belong to the same SCOP family as 1atp, namely 1phk, 1csn, and 1hck. As expected, those are the proteins that have the highest degree of overall sequence and structure similarity with 1atp.

We did the experiments between the query binding site of 1atp represented by a BB and the largest cavity of each target protein, as described in the previous sections. In Table 2, for each pairwise comparison, we report the top 5 BBs of the cavity that are found most similar to the query BB according to the average correlation.

Table 2 (column 7) shows the coverage values for all pairwise comparisons of protein 1atp; they are computed as the ratio of the values in columns 6 and 2. Not surprisingly, some of the best coverage results are obtained for the three proteins 1phk, 1csn, and 1hck that have the same SCOP classification as 1atp (*protein kinase*). However, similar good values of coverage are reported for proteins 1a49, 1f9a, 1b8a, 1gn8, 1mjh, and 1yag that belong to different SCOP families and share low degree of sequence and structure similarity with the query. On average, our coverage results are good. The average coverage computed over all 5 top-ranked BBs of all pairwise comparisons is 0.78 (as reported at the bottom of the table). We observe that in a pairwise comparison the best coverage value over all 5 top BBs (in bold in column 7 of Table 2) is generally achieved by the top-ranked BB. Indeed, the average coverage value computed only for the top BB of all pairwise comparisons is 0.85. This observation may be useful in future work that will address the problem of aligning binding sites based on spherical harmonics since it will allow us to restrict the processing to the top BB.

There are cases where our procedure fails to locate the binding site of ATP on the target protein. We consider a failure to occur when the coverage of the binding site for the top-ranked BB is below the value of 0.25. Failure occurs with three proteins: 1fmw (coverage 0), 1e8x (coverage 0.13), and 1kay (coverage 0.23). It might appear surprising that the correlation values in correspondence to these failures are relatively high. For instance, the correlation value for the top BB of 1fmw is 2617.8. It turns out that protein 1fmw, the motor domain of dictyostelium myosin, hosts in its largest cavity another ligand BL7 in addition to ligand ATP (PDB complex 3bz9). The compound BL7 consists of several rings of atoms, two of which are structurally similar to the adenine ring of ATP. We found that the top BB obtained for 1fmw has a good

TABLE 2. PAIRWISE COMPARISONS USING AS QUERY THE BINDING SITE OF 1ATP BOUND TO ATP

Protein	*No. of residues of the BS*	*Rank of BB*	*No. of residues of the BB*	*Average correlation*	*No. of residues of BB ∩ BS*	*Coverage*	*Accuracy*
1a49A	32	1	46	2593.5	30	**0.97**	0.67
		2	49	2538.2	32	0.97	0.63
		3	38	2524.0	30	0.78	0.66
		4	45	2518.3	31	0.97	**0.69**
		5	44	2514.5	31	0.94	0.68
1a49C	34	1	45	2606.6	34	**1.00**	0.76
		2	45	2586.1	31	0.91	0.69
		3	41	2577.9	29	0.85	0.71
		4	47	2577.7	33	0.97	0.70
		5	44	2557.	34	1.00	0.77
1a82	23	1	32	2040.2	16	0.70	0.50
		2	28	1970.3	17	0.74	0.61
		3	27	1946.7	15	0.65	0.56
		4	30	1941.4	21	**0.91**	0.70
		5	31	1900.8	14	0.61	0.45
1ayl	27	1	33	2490.9	23	**0.85**	0.70
		2	34	2457.	23	0.85	0.68
		3	32	2452.4	23	0.85	0.72
		4	33	2445.4	23	0.85	0.70
		5	31	2410.6	19	0.70	0.61
1b8aB	22	1	42	2580.3	22	**1.00**	0.52
		2	39	2572.6	17	0.77	0.44
		3	36	2525.9	19	0.86	0.53
		4	39	2415.8	22	1.00	0.56
		5	38	2406.1	22	1.00	0.58
1csn	28	1	35	2271.0	28	**1.00**	0.80
		2	33	2237.1	28	1.00	0.85
		3	34	2232.3	28	1.00	0.82
		4	32	2215.1	24	0.86	0.75
		5	33	2189.0	28	1.00	0.85
1e2q	19	1	29	2262.8	11	**0.58**	0.38
		2	29	2238.6	11	0.58	0.38
		3	31	2196.1	10	0.53	0.32
		4	29	2178.7	11	0.58	0.38
		5	26	2155	11	0.58	0.42
1e8x	23	1	37	2813.1	3	0.13	0.08
		2	38	2791.6	2	0.09	0.05
		3	37	2729	2	0.09	0.05
		4	33	2719.8	4	**0.17**	0.12
		5	39	2701.2	2	0.09	0.05
1f9aC	24	1	31	2410.1	24	**1.00**	0.77
		2	30	2406.1	23	0.96	0.77
		3	27	2391	23	0.96	0.85
		4	26	2351.1	23	0.96	0.88
		5	25	2291.5	23	0.96	0.92
1fmw	24	1	39	2617.8	0	0.00	0.00
		2	39	2610.2	0	0.00	0.00
		3	33	2588.3	1	**0.04**	0.03
		4	39	2555.2	0	0.00	0.00
		5	36	2503.1	1	0.04	0.03
1g5t	13	1	23	2151.2	12	**0.92**	0.52
		2	24	2059.9	12	0.92	0.50
		3	24	2054	12	0.92	0.50
		4	23	1987.9	8	0.62	0.35
		5	24	1968.4	10	0.77	0.42

(*continued*)

TABLE 2. (CONTINUED)

Protein	*No. of residues of the BS*	*Rank of BB*	*No. of residues of the BB*	*Average correlation*	*No. of residues of BB ∩ BS*	*Coverage*	*Accuracy*
1gn8A	23	1	34	2344.3	23	**1.00**	0.68
		2	33	2296.5	23	1.00	0.70
		3	29	2285.	22	0.96	0.76
		4	31	2274.4	23	1.00	0.74
		5	32	2233.8	22	0.96	0.69
1hck	24	1	32	2391	23	0.96	0.72
		2	33	2381.8	23	0.96	0.70
		3	31	2364.7	24	**1.00**	0.77
		4	34	2335.3	23	0.96	0.68
		5	33	2316.1	23	0.96	0.70
1j7k	23	1	25	2182.4	21	0.91	0.84
		2	26	2141.2	22	**0.96**	0.85
		3	27	2087.1	20	0.87	0.74
		4	26	2020.4	9	0.39	0.35
		5	24	2011.6	22	0.96	0.92
1jjv	14	1	33	1865.8	8	**0.57**	0.24
		2	32	1841.0	8	0.57	0.25
		3	34	1771.1	6	0.43	0.18
		4	36	1730.1	7	0.50	0.19
		5	30	1692.6	8	0.57	0.27
1kay	30	1	31	2314.4	7	0.23	0.23
		2	43	2286.9	26	**0.87**	0.60
		3	42	2275.1	22	0.73	0.52
		4	38	2259.3	23	0.77	0.61
		5	46	2235.5	25	0.83	0.54
1kp2	26	1	43	2669.1	24	0.92	0.56
		2	42	2622.8	22	0.85	0.52
		3	42	2577.3	26	**1.00**	0.62
		4	38	2533.6	21	0.81	0.55
		5	43	2533.4	24	0.92	0.56
1mjhA	24	1	33	2211.6	24	**1.00**	0.73
		2	31	2198.8	23	0.96	0.74
		3	33	2163.4	24	1.00	0.73
		4	32	2141.4	23	0.96	0.72
		5	31	2103.7	23	0.96	0.74
1nsf	25	1	35	2162.8	24	**0.96**	0.69
		2	31	2099.0	24	0.96	0.77
		3	30	2007.8	20	0.80	0.67
		4	26	2002.2	20	0.80	0.77
		5	26	1994.1	20	0.80	0.77
1phk	27	1	29	2255.0	10	0.37	0.34
		2	33	2249.8	15	0.56	0.45
		3	31	2207.6	27	**1.00**	0.87
		4	32	2204	27	1.00	0.84
		5	31	2194.8	27	1.00	0.87
1yag	27	1	37	2402.6	26	**0.96**	0.70
		2	36	2372.2	25	0.93	0.69
		3	38	2320.4	26	0.96	0.68
		4	35	2318.5	26	0.96	0.74
		5	35	2304.4	26	0.96	0.74
Average						0.78	0.58

For each pairwise comparison, we report the top 5 BBs of the protein that are found most similar to the query BB according to the average correlation (column 5). For each target protein, we list the number of residues of the binding site with ATP (column 2), the number of residues in each of the top 5 BBs (column 4), and the number of residues common to the binding site and to the considered BB (column 6). The BBs with the best coverage are shown in bold.

coverage (0.5) of the binding site of BL7; furthermore, the residues of the binding site of BL7 contained in the top BB are those in contact with the two rings similar to the adenine ring of ATP. The structural similarity of these two surface regions is correctly detected by our procedure. Similar considerations apply to the other two proteins 1e8x and 1kay, whose binding sites with ATP have high structural similarity with that of 1fmw.

Next we evaluate the performance of our method in terms of accuracy. The accuracy in column 8 of Table 2 is computed as the ratio of values in columns 6 and 4. We observe from the table that the accuracy values of the top BBs although reasonably good are generally not close to the maximum obtainable value of 1.00. This is somewhat expected because a binding site is rarely a spherical region and is often smaller than the surface region represented by any BB that includes the binding site.

Finally, we assess the goodness of the average correlation as scoring function relative to the classical maximum correlation. Our choice of average correlation was mostly based on computational considerations; as we have shown in the previous section, the time complexity of our approach improves over the classical one. This translates in shorter execution times, as we will see below. However, it turns out that this score is valid also if we take into account the quality of the results. We repeated the experiments on the same data set using the maximum correlation and obtained on average coverage of 70.4% as opposed to 78% in our case (Table 2).

4.2. Detection of binding sites of ligand EQU

A second dataset was collected containing proteins from different SCOP families binding the steroid EQUILENIN (EQU). We used as query the binding site extracted from the complex logz. On this set, we conducted two types of experiments. The first was similar to the one described above, namely the query binding site was searched in the largest cavity of the another protein. The results are reported in Table 3, which shows only the top-ranked BB. The higher coverage values with respect to the results of ATP are explained by the known limited conformational variability of this type of ligand. In the second type of experiments, the search was not limited to the cavity but done over the entire protein surface. As the data in Tables 3 and 4 show, in both cases the method is very effective, and the results are only slightly better for the case of cavities. This is an important fact because, although in the majority of cases a ligand resides in the largest cavity, sometimes it may be located on a relatively flat area. Thus, a search over the entire surface will allow to locate the active site in all cases.

4.3. Functional prediction

One important application of protein surface matching is functional prediction. When a novel protein with unknown function is discovered, a huge set of proteins with known function and binding sites can be screened, searching for a candidate binding site in the new protein. More specifically, if a surface region of the novel protein is similar to that of the binding site of another protein, the function of the novel protein can be inferred and its molecular interaction predicted. Although our method was not explicitly designed

TABLE 3. THE QUERY BINDING SITE IS THAT OF EQU IN COMPLEX WITH 1OGZ

Protein	*No. of residues of the BS*	*No. of residues of top-ranked BB*	*Average correlation*	*No. of residues of BB ∩ BS*	*Coverage*	*Accuracy*
1csq	15	22	1645	15	1.00	0.68
1gs3	17	31	1950	17	1.00	0.55
1ogx	17	30	1689	17	1.00	0.57
1ogz	16	28	2154	16	1.00	0.57
1oh0	16	23	1814	16	1.00	0.70
1oho	16	27	1920	16	1.00	0.59
1qjg	14	27	2224	14	1.00	0.52
1w6y	16	22	1750	16	1.00	0.73
Average					0.99	0.62

The search is done on the largest cavity of the protein listed in column 1. Columns 4 and 5 show coverage of binding site and accuracy of top BB in terms of residues.

Table 4. The Query Binding Site Is that of EQU in Complex with 1ogz

Protein	*No. of residues of the BS*	*No. of residues of top-ranked BB*	*Average correlation*	*No. of residues of BB ∩ BS*	*Coverage*	*Accuracy*
1csq	15	23	1644	15	1.00	0.65
1gs3	17	31	1950	17	1.00	0.55
1ogx	17	30	1690	17	1.00	0.57
1ogz	16	29	2124	16	1.00	0.55
1oh0	16	29	1777	16	1.00	0.55
1oho	16	27	1920	16	1.00	0.59
1qjg	14	21	2068	11	0.79	0.52
1w6y	16	22	1750	16	1.00	0.73
Average					0.92	0.58

The search is done on the entire surface of the protein listed in column 1. Columns 4 and 5 show coverage and accuracy of the top BB.

Table 5. Comparison of the Results Obtained with the Query Binding Site of EQU (Columns 3–5) and with the Query Binding Site of ATP (Columns 6–8) on the Same Set of Five Proteins Binding ATP

		EQU vs. ATP			*ATP vs. ATP*		
Protein	*No. of residues in the BS*	*No. of residues of top BB*	*Average correlation*	*Coverage*	*No. of residues of top BB*	*Average correlation*	*Coverage*
1a49C	34	43	2362	0.88	45	2607	1.00
1b8aA	22	40	2322	0.77	42	2580	1.00
1jjv	14	29	1677	0.36	33	1866	0.57
1kay	30	41	2021	0.70	43	2287	0.87
1j7k	23	26	1963	0.39	25	2182	0.91
Average			2069.00	0.62		2304	0.87

Table 6. Comparison of the Execution Times of Our Method (Columns 2–3) with the Classical Method Based on Spherical Harmonics (Columns 4–5)

	Our method, time (sec)		*Classical method, time (sec)*	
Protein	*Cavity only*	*Entire surface*	*Cavity only*	*Entire surface*
1csq	31.8	68.5	41.1	87.1
1gs3	27.5	65.3	33.8	97.9
1ogx	37.7	89.8	48.8	120.9
1ogz	17.4	75.1	20.9	105.3
1oh0	26.4	70.8	30.4	90.1
1oho	23.4	70.7	32.1	95.9
1qjg	24.1	75.0	33.8	106.7
1w6y	22.0	69.7	27.2	87.4
Average	26.3	73.1	33.5	98.9

Execution times are shown for two instances of the comparison: the search for a binding site is done on the protein cavity only (columns 2, 4) or the entire surface (columns 3, 5).

for functional prediction, it may provide a good initial step towards the solution of this problem. In fact, as shown by the following experiment, the correlation values could provide a basis for prediction. In the experiment, we use the binding sites of two different ligands (ATP and EQU) to query a set of five target proteins binding ATP. Table 5 shows the results of all pairwise comparisons of the two queries with the target proteins. We find that the average correlation over all pairwise comparisons of a single query is larger for the query that binds ATP, i.e., the same ligand bound to the target structures.

A thorough evaluation of our proposed method would have required comparisons with other methods proposed in the literature. Although there exist several related methods, as surveyed in Section 2, the problems they address often differ in significant ways. For instance, in most cases the problem is to compare/classify binding sites. Another problem that has received attention is the alignment of binding sites or cavities on two protein surfaces, in other words the determination of a set of corresponding atoms/residues on two proteins based on some geometric or chemico-physical constraints. By contrast, the problem we considered here is the fast screening of a structure to locate a candidate binding site using as query a known binding site. No alignment is required nor used as a step towards the detection of the binding site.

4.4. Execution times

A major strength of our approach is its reduced computational complexity with respect to the classical approach based on spherical harmonics, as discussed in Section 3.5. In Table 6, we list the CPU time per structure pair using the binding site of EQU as query; the program is run on a AMD Athlon XP 2600. We report the execution times in the two instances of the comparison: when the search involves the largest cavity only (column 2) or the entire surface (column 3). We also show the execution times of the classical approach in which the maximum rather than the average correlation is computed. The CPU time for such computations for the two instances of the comparison are shown in Table 6 (columns 4 and 5, respectively). When considering the entire surface of the target protein, the average CPU time per pair of our approach, reported at the bottom of the Table (column 3), is much lower than that to compute the maximum correlation (column 5); in fact, the time saving is on average 40%.

5. CONCLUSION

We have developed a method for a fast detection of candidate binding sites on a protein surface based on a property of spherical harmonics. We have shown that this approach is fast and effective: it achieves high coverage of the binding site on average (0.78 on the benchmark dataset) with execution times of few seconds.

Our scoring scheme, if compared with the traditional approach based on the maximum correlation, not only is less computationally demanding, but also achieves higher coverage results. The high computational efficiency allows the screening of large datasets of structures in search of functionally related proteins. Thus, we expect this procedure to be a valuable aid in assisting biologists in the difficult task of assigning a functional role to proteins.

To evaluate the goodness of results, a variety of different criteria have been used in the literature. For example, many authors use the RMSD of the aligned residues, or the number of aligned residues. Some approaches use our same evaluation criterion, i.e., the coverage of the binding site, but computed for the obtained alignment. In general, the coverage of an alignment is conceivably worse than that of our screening method because of the additional constraints. In fact, the alignment coverage values of MolLoc reported in Bock et al. (2008) for some of the pairs considered in our tests (1atp paired with 1phk, 1csn, 1nsf) are consistently smaller than our values. Future work on aligning the candidate binding sites based on spherical harmonics will allow a more comprehensive evaluation.

ACKNOWLEDGMENT

Support for Guerra and Comin was provided by the University of Padova under the Research Project CPDA077213/07, and by the Fondazione Cariparo.

DISCLOSURE STATEMENT

The authors declare they have no competing interests.

REFERENCES

Artymiuk, P., Spriggs, R., and Willett, P. 2005. Graph theoretic methods for the analysis of structural relationships in biological macromolecules. *J. Am. Soc. Inform. Sci. Technol.* 56, 518–528.

Ausiello, G., Gherardini, P.F., Marcatili, P., et al. 2008. Funclust: a web server for the identification of structural motifs in a set of non-homologous protein structures. *BMC Bioinform.* 9, S2.

Ballester, P., and Richards, W. 2007. Ultrafast shape recognition to search compound databases for similar molecular shapes. *J. Comput. Chem.* 28, 1711–1723.

Barker, J., and Thornton, J. 2003. An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. *Bioinformatics* 19, 1644–1649.

Binkowski, T., Adamian, L., and Liang, J. 2003. Inferring functional relationships of proteins from local sequence and spatial surface patterns. *J. Mol. Biol.* 332, 505–526.

Binkowski, T., Adamian, L., and Liang, J. 2005. Protein surface analysis for function annotation in high-throughput structural genomics pipeline. *Protein Sci.* 14, 2972–2981.

Bock, M., Garutti, C., and Guerra, C. 2007a. Discovery of similar regions on protein surfaces. *J. Comput. Biol.* 14, 285–299.

Bock, M., Garutti, C., and Guerra, C. 2007b. Effective labeling of molecular surface points for cavity detection and location of putative binding sites. *Proc. VI Int. Conf. Comput. Syst. Bioinform.* 263–274.

Bock, M., Garutti, C., and Guerra, C. 2008. Cavity detection and matching for binding site recognition. *Theoret. Comput. Sci.* doi:10.1016/j.tcs.2008.08.018.

Cai, W., Shao, X., and Maigret, B. 2002. Protein ligand recognition using spherical harmonic molecular surfaces: towards a fast and efficient filter for large virtual throughput screening. *J. Mol. Graphics Model.* 20, 313–328.

Chen, B., Bryant, D., Fofanov, V., et al. 2005a. Cavity-aware motifs reduce false positives in protein function prediction. *Comput. Syst. Bioinform. Conf.* 311–323.

Chen, B., Fofanov, V., Kristensen, D., et al. 2005b. Algorithms for structural comparison and statistical analysis of 3D protein motifs. *Proc. Pac. Symp. Biocomp.* 334–345.

Connolly, M. 1983. Analytical molecular surface calculation. *J. Appl. Cryst.* 16, 548–558.

Healy, D.M., Kostelec, P., Rockmore, D., et al. 2003. Ffts for the 2-sphere-improvements and variations. *J. Fourier Anal. Appl.* 9, 341–385.

Duncan, B., and Olson, A. 1993. Approximation and characterization of molecular surfaces. *Biopolymers* 33, 219–229.

Hofbauer, C., and Aszodi, A. 2005. Sh2 binding site comparison: a new application of the surfcomp method. *J. Chem. Inform. Model.* 45, 414–421.

Hofbauer, C., Lohninger, H., and Aszodi, A. 2004. Surfcomp: a novel graph-based approach to molecular surface comparison. *J. Chem. Inform. Comput. Sci.* 44, 837–847.

Jambon, M., Olivier, A., Combet, C., et al. 2005. The SuMo server: 3D search for protein functional sites. *Bioinformatics* 21, 3929–3930.

Kazhdan, M., Funkhouser, T., and Rusinkiewicz, R. 2003. Rotation invariant spherical harmonic representation of 3D shape descriptors. *SGP '03 Proc. 2003 Eurographics/ACM SIGGRAPH Symp. Geometry Process.* 156–164.

Kinoshita, N., Furui, J., and Nakamura, H. 2001. Identification of protein functions from a molecular surface database, ef-site. *J. Struct. Funct. Genomics* 2, 9–22.

Kostelec, P., and Rockmore, D. 2008. Ffts on the rotation group. *J. Fourier Anal. Appl.* 14.

Kuttner, Y., Sobolev, V., Raskind, A., et al. 2003. A consensus-binding structure for adenine at the atomic level permits searching for the ligand site in a wide spectrum of adenine-containing complexes. *Proteins* 52, 400–411.

Laurie, A., and Jackson, R. 2006. Methods for the prediction of protein-ligand binding sites for structure-based drug design and virtual ligand screenicng. *Curr. Protein Pept. Sci.* 7, 395–406.

Leicester, S., Finney, J., and Bywater, R. 1988. Description of molecular surface shape using Fourier descriptors. *J. Mol. Graphics* 6.

Leicester, S., Finney, J., and Bywater, R. 1994a. A quantitative representation of molecular surface shape. i: Theory and development of the method. *J. Math. Chem.* 16, 315–341.

Leicester, S., Finney, J., and Bywater, R. 1994b. A quantitative representation of molecular surface shape. ii: Protein classification using Fourier shape descriptors and classical scaling. *J. Math. Chem.* 16, 343–365.

Makadia, A., and Daniilidis, K. 2006. Rotation recovery from spherical images without correspondences. *IEEE Trans. PAMI* 28, 1170–1175.

Max, N., and Getzoff, E. 1988. Spherical harmonic molecular surfaces. *Comput. Graphics Appl. IEEE* 8, 42–50.

Minai, R., Matsuo, Y., Onuki, H., et al. 2008. Method for comparing the structures of protein ligand-binding sites and application for predicting protein-drug interaction. *Proteins* 72, 367–381.

Morris, R.J., Najmanovich, R., Kahraman, A., et al. 2005. Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons. *Bioinformatics* 21, 2347–2355.

Murzin, A., Brenner, S., Hubbard, T., et al. 1995. Scop: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536–540.

Ritchie, D., and Kemp, G. 1999. Fast computation, rotation, and comparison of low resolution spherical harmonic molecular surfaces. *J. Comput. Chem.* 20, 383–395.

Ritchie, D., and Kemp, G. 2000. Protein docking using spherical polar Fourier correlations. *Proteins* 39, 178–194.

Ritchie, D., Kozakov, D., and Sandor, V. 2008. Accelerating and focusing protein-protein docking correlations using multi-dimensional rotational fft generating functions. *Bioinformatics* 24, 1865–1873.

Shatsky, M., Shulman-Peleg, A., Nussinov, R., et al. 2006. The multiple common point set problem and its application to molecule binding pattern detection. *J. Comput. Biol.* 13, 407–428.

Shindyalov, I., and Bourne, P. 1998. Protein structure alignment by incremental combinatorial extension (ce) of the optimal path. *Protein Eng.* 11, 739–747.

Shulman-Peleg, A., Nussinov, R., and Wolfson, H.J. 2004. Recognition of functional sites in protein structures. *J. Mol. Biol.* 339, 607–633.

Sommer, I., Miller, O., Domingues, F., et al. 2007. Moment invariants as shape recognition technique for comparing protein binding sites. *Bioinformatics* 23, 3139–3146.

Stockwell, G.R., and Thornton, J. M. 2006. Conformational diversity of ligands bound to proteins. *J. Mol. Biol.* 356, 928–944.

Weskamp, N., Kuhn, D., Hullermeier, E., et al. 2004. Efficient similarity search in protein structure databases by k-clique hashing. *Bioinformatics* 20, 1522–1526.

Yao, H., Kristensen, D.M., Mihalek, I., et al. 2003. An accurate, sensitive, and scalable method to identify functional sites in protein structures. *J. Mol. Biol.* 326, 255–261.

Address correspondence to:
Dr. Matteo Comin
Department of Information Engineering
University of Padova
Via Gradenigo 6/A
Padova 35131, Italy

E-mail: ciompin@dei.unipd.it