

EFFECTIVE LABELING OF MOLECULAR SURFACE POINTS FOR CAVITY DETECTION AND LOCATION OF PUTATIVE BINDING SITES

Mary Ellen Bock

Dept. of Statistics, Purdue University 150 N. University Street, West Lafayette, IN 47907-2067, USA
E-mail: mbock@purdue.edu

Claudio Garutti

Dept. of Information Engineering, University of Padova, Via Gradenigo 6a, 35131 Padova, Italy
E-mail: garuttic@dei.unipd.it

Concettina Guerra

Dept. of Information Engineering, University of Padova, Via Gradenigo 6a, 35131 Padova, Italy
College of Computing, Georgia Institute of Technology, 801 Atlantic, Atlanta, GA, USA
E-mail: guerra@dei.unipd.it

We present a method for detecting and comparing cavities on protein surfaces that is useful for protein binding site recognition. The method is based on a representation of the protein structures by a collection of spin-images and their associated spin-image profiles. Results of the cavity detection procedure are presented for a large set of non-redundant proteins and compared with SURFNET-ConSurf. Our comparison method is used to find a surface region in one cavity of a protein that is geometrically similar to a surface region in the cavity of another protein. Such a finding would be an indication that the two regions likely bind to the same ligand. Our overall approach for cavity detection and comparison is benchmarked on several pairs of known complexes, obtaining a good coverage of the atoms of the binding sites.

Keywords: protein surfaces comparison; spin-images; binding sites; cavity detection; drug design

1. INTRODUCTION

The automatic recognition of regions of biological interest, such as binding sites, on protein surfaces is a critical task in function determination and drug design. The number of protein structures available is increasing, while the assessment of the function of a protein binding site involves time-demanding experimentation with ligands. To this extent, every tool is welcome that can give function-related information, like putative binding sites, for directing the experimental phase.

Cavity detection is often the first step for functional analysis, since binding sites in proteins usually lie in cavities. In our work, we represent a protein surface using spin-images, and, based on such representation, use a labeling of surface points that is effective in finding cavities and binding sites. Our approach is simple and fast, purely geometric with no dependence on physico-chemical properties. It examines a subset of surface points, generally less than half of the original points, that are likely to lie on cavities. Those are the points, labeled *blocked*, whose normal intersects the protein surface at some other

point. For each blocked point, the procedure generates a trial sphere and constrains the radius of the sphere so that it does not penetrate any neighboring atom, by using the values of the spin-image. The clusters of overlapping spheres correspond to surface cavities.

One use of the method is to compare similarities of a cavity from one protein to a cavity in another protein. The comparison method based on spin-images, introduced for protein surface comparison,^{1,2} can be adapted to find a surface region in one cavity that is geometrically similar to a surface region in the other cavity. Such a finding would be an indication that the two regions likely bind to a common ligand. Typically, the surface region that constitutes the binding site of a ligand in a cavity is only a small part of the total surface area of the cavity and the volume of the cavity is much larger than needed to accommodate the ligand. One extension of the comparison of cavities in proteins is to compare cavities found in two different chains of the same protein. Once again similar surface regions within the two cavities may indicate binding sites for the same lig-

tion in terms of spherical harmonic coefficients. This method is interesting and fast; however, as pointed out by the authors, it requires a registration phase, to align the two shapes, that it is not always very reliable. A geometric hashing approach have been used¹³ to compare and cluster phosphate binding sites in protein-nucleotide complexes, leading to the identification of 10 clusters. These are the structural P-loop, di-nucleotide binding motif [FAD/ NAD(P)-binding and Rossmann-like fold] and FAD binding motif. A cavity-aware match technique¹⁴ which uses C-spheres to represent active clefts which must remain vacant for ligand binding. The technique reduces the number of false positives while maintaining most of the true positive matches found with identical motifs lacking C-spheres. A different instance of the comparison problem^{1,2} is when two complete protein surfaces are compared to discover their most similar regions. The adaptation of this method to surface cavities will be discussed in this paper.

3. SURFACE CHARACTERIZATION

3.1. Spin-image representation of protein surfaces

We represent the molecular surface as a collection of spin-images, each of them associated to a surface point with its normal. Surface points are generated using Connolly's molecular representation.²⁶ Spin-images are semi-local shape descriptors used mostly in the area of computer vision for 3D model retrieval and registration.²⁷ A spin-image provides a high-dimensional description of the appearance of a 3D object in a local reference system. It is an histogram of quantized surface point locations in a local coordinate system associated to a 3D point on the surface and to its normal. Spin-images are discriminative (and as such can be used for recognition), easy to compute and invariant under rigid transformations.

For a surface point P with normal n , let (P, n) be the coordinate system with origin in P and axis n . In this system, every surface point Q is represented by two coordinates (α, β) , where α is the perpendicular distance of Q to n , and β the signed perpendicular distance of Q to the plane T through P perpendicular to n . The spin-image is a two-dimensional histogram of the quantized coordinates (α, β) of the surface points. The image pixels are of size equal to 1

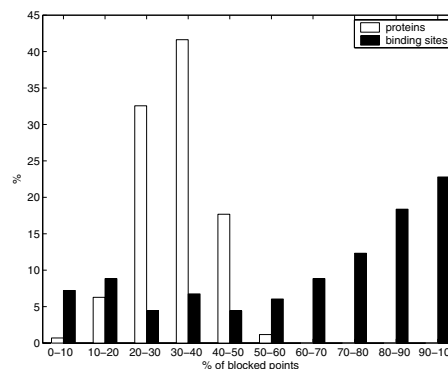


Fig. 1. Histogram of the number of blocked points on protein surfaces and binding sites.

\AA in our application. A spin-image is rotation invariant since all points on a ring centered on the normal n have the same coordinates.

The spin-image dimensions depend on the point P and its corresponding tangent plane and corresponding normal n to its tangent plane T . The number of columns depends on the maximum distance α_{max} from n of other points on the surface of the object. Let h be the number of rows and k be the number of columns of the spin-image. If $\beta_r = \beta_{max} - \beta_{min}$ then $h = \lceil \beta_r / \epsilon \rceil$ and $k = \lceil \alpha_{max} / \epsilon \rceil$, where ϵ is the pixel size.

3.2. Characterizing cavities in terms of blocked points

We label surface points as *blocked* or *unblocked* depending on the shape of their spin-images. A surface point P with normal n is labeled blocked if n intersects the surface at any other point lying above the tangent plane T at P perpendicular to n ; otherwise it is labeled unblocked. To label a point, only the first column of its spin-image needs to be examined: if it contains a non-zero pixel with positive β , then the point is blocked, otherwise it is unblocked.

Crucial to our cavity detection procedure is the identification of blocked points on the protein surface. Typically, the number of blocked points on a protein surface is smaller than that of unblocked points, i.e. of points whose normal does not intersect the surface at any other point. Not surprisingly, the opposite is true for points of the binding sites.

In Fig. 1 we show the statistics of blocked points of proteins and binding sites (the proteins are taken from a non-redundant dataset³ that will be discussed

in more detail later). For most proteins, less than 50% of the surface points are blocked, while for the majority of the binding sites, more than 70% of points are blocked.

For example, out of 5039 Connolly’s points of protein 1nsf (D2 Hexamerization domain of N-Ethylmaleimide sensitive factor) 1800 are blocked, i.e. approximately 35% of the total. For the binding site of 1nsf with ligand ATP, the percentage of blocked points goes up to 74%. As another example, protein 1mjh, an hypothetical protein binding ATP, has an even higher percentage of blocked points on the binding site, i.e. above 80%.

Furthermore, blocked points are strongly present in cavities, especially in internal cavities. In fact, if a cavity is internal, then the normals at all points of the cavity intersect the protein at some other points of the cavity. If a cavity is external, there might be few unblocked points at the bottom of the cavity. Thus, for cavity detection, we restrict our analysis to blocked points.

The identification of blocked points can be done very easily once the spin-images of surface points have been constructed. If the first column (corresponding to $0 \leq \alpha < \varepsilon$) of a spin-image contains a non-zero pixel with positive β , then the point is blocked, otherwise is unblocked. Here we are assuming that the normal n intersects the surface at some other point Q if n is within ε distance from Q , where ε is the spin-image pixel size.

4. METHODS

4.1. Cavity detection

Our approach in delineating surface cavities considers only blocked points. For each blocked point, it builds the largest sphere that can fit at that point; then it determines the cavities as clusters of overlapping spheres. Given a blocked point P with normal n and spin-image $spin(P)$, the associated sphere is obtained from the biggest (discrete) semi-circle in $spin(P)$, tangent to the cell in O and containing only empty cells of $spin(P)$. Due to the cylindrical symmetry of spin-images, the semi-circle of $spin(P)$ corresponds to the sphere in 3-D. Defining the sphere starting from the spin-image allows fast construction of the spheres.

For a blocked point, we find the sphere as follows. We consider the *horizontal profile* of a spin-image as

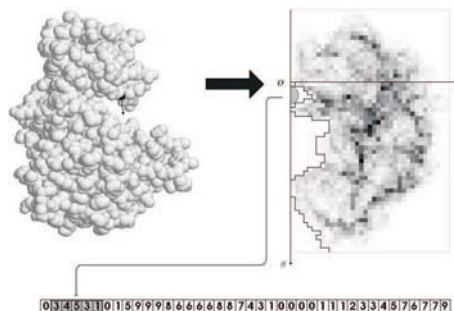


Fig. 2. Determination of the sphere using spin-image horizontal profile.

a one-dimensional array with length $Z + 1$, where Z is a count of the number of successive zero elements along the column 0 (corresponding to $0 \leq \alpha < \varepsilon$) of the spin-image for $\beta \geq 0$ starting at $\beta = 0$. The j^{th} element of the vector is given by the number of contiguous zero-elements in row i of the spin-image starting at column 0 and ending at the first non-zero cell along row i .

Z is a constraint on the largest possible diameter of a sphere that can touch the protein surface at the blocked point (We have assumed ε equal to 1 \AA). The particular values of the elements of the profile further constrain the largest diameter of such a sphere. To calculate the largest possible radius of the sphere, LPR , we initially set the variable R equal to $Z/2$. As we observe the values of the horizontal profile starting at position 1, no constraint is imposed if the value is greater than the current value of R . The smallest position j such that the vector value at j^{th} position is smaller than the current value of R gives the first constraint upon the LPR and this must be calculated. For i positive, a value of i in position i is a constraint of radius i on LPR . More generally, it can be easily shown that a value i at position j is a constraint of c on LPR where $c = (i^2 + j^2)/2j$ if $i \geq j$ and $c = (i^2 + (j - 1)^2)/2(j - 1)$, otherwise. If c is less than R , then R is set to c . For successive positions in the horizontal profile, this computation is repeated if the profile value is smaller than R . Fig. 2 shows an example of determination of the sphere using the spin-image horizontal profile.

For a molecule with a set B of blocked points, we generate spheres only for the subset B' of points of B with a Z value below a given threshold (10 \AA , in our tests). Blocked points with larger Z values are not typical of cavities, since they can also be found

at the top of a region if their normal intersects the surface at a far away region.

Our overall approach is simple and fast. The time required to generate all spheres is $O(b \times d)$, where b is the number of considered blocked points, typically much smaller than the number m of all surface points, and $d = 10$ is the maximum Z value of the spin-images. If we take into account the pre-processing phase needed to create m spin-images, the overall time complexity of our procedure becomes $O(m \times \max\{m, D\} + b \times d)$, where D is the size of the spin-image. This represents a computational advantage with respect to methods for cavity detection that generate m^2 trial spheres, one for each pair of surface points, and check the non penetration of other surface points into each sphere, obtaining an overall time complexity of $O(m^3)$. Notice that the complexities of both approaches can be improved by the use of clever techniques for neighbor finding operations. In our approach, these could lead to a faster creation of spin-images, if only local points are chosen to contribute to the construction of the spin-image of a given point. In the other approaches, fast neighbor finding operations could speed up the check of the non penetration constraint.

Once all spheres of blocked points are obtained, those with LPR below a certain threshold (1 \AA in our experiments) are removed so that small gaps between atoms are not considered. From the remaining spheres, a clustering procedure determines collections of interpenetrating spheres corresponding to the points of the surface cavities. The clusters are identified as the connected components of the undirected graph $G = (V, E)$, in which the vertices are the blocked points, and an edge connects two vertices if their spheres overlap. The overall procedure is outlined below.

PROCEDURE: Cavity Detection

- (1) For a given protein surface, determine the set of blocked points B and its subset B' consisting of points with Z less than a predefined threshold $ThZ = 10$.
- (2) For each point b of B' , build the sphere touching the surface at b from its spin-image profile, as described above.
- (3) Prune the set B' by removing all points with a radius of the sphere $r < 1A$.
- (4) Find the connected components G_1, \dots, G_n of G using Breadth First Search.

The vertices of each connected component of G form a cluster corresponding to a surface cavity. Note that point density has an impact on the choice of the parameters. In our work, we generated one point every square angstrom. The threshold values for ThZ and r were assessed by performing cavity detection on 30 random proteins from the dataset³ using different values of the parameters.

4.2. Finding similar binding sites on two proteins

We now give an outline of our overall approach for detecting similar binding sites on two protein surfaces.

- (1) Build the spin-image representation of the surface points of the two proteins.
- (2) For each protein, find the surface cavities based on the spin-image profiles of blocked points and select the largest cavity(ies).
- (3) Compare pairs of cavities, one per protein, by identifying and grouping sets of corresponding points based on the correlation of their associated spin-images. Return the regions on the two cavities that are most similar.

Step 1 and 2 have been described in the previous sections. For comparing pairs of cavities in step 3 we use an adaptation of the recognition method based on spin-images,^{1,2} and here referred to as *MolLoc*, that allows the discovery of similar regions on protein surfaces. MolLoc takes as input a pair of proteins and finds the regions on the two surfaces that most resemble each other.

Basically, for two given proteins g and g' , MolLoc builds individual point correspondences (Q, Q') , $Q \in g$ and $Q' \in g'$, if their spin-images have a high correlation value. A high correlation value is taken as an indication of structural similarity of the local regions surrounding the two points and contributing to the spin-images. Once point correspondences are identified, they are clustered into groups of consistent correspondences. The consistency criterion is purely geometric and enforces the rigidity constraint of three dimensional objects. It states that the angles between normals at two surface points on one protein and the distances between the two points must be preserved between the corresponding points of the other protein.

Although effective in identifying surface similarity, MolLoc suffers from high computational complexity. For a pair of large proteins, the execution time can be up to two hours. A number of heuristics have been proposed to cope with this problem. One heuristics consists of mapping surface points into cells of a 3D grid, and restricting the matching procedure to points contained into pairs of grid cells and into their neighboring cells.

We use the same basic matching procedure for comparing two surface cavities obtaining execution times that are of the order of minutes or even seconds. No mapping of points into a 3D grid is necessary, which is also instrumental in producing more accurate results.

5. DATA AND RESULTS

5.1. Cavity detection

We conducted experiments for cavity detection on a dataset of 244 previously defined³ The protein structures are taken from the PDB. Of these proteins, 112 are enzymes (45.9%), 129 nonenzymes (52.9%), and three "hypothetical" (1.2%) proteins, according to PDBsum²⁸ and Uniprot²⁹ These PDB entries contained 464 ligands not covalently bound to the protein and then for each complex protein-ligand there is a binding site. The binding sites of these complexes are determined in the following way. For a ligand binding to a protein, the binding site consists of the atoms of the protein that are (i) closer than a given threshold (5 Å in our experiments) to at least one atom of the ligand, and (ii) have at least one surface point that is *blocked by the ligand*. A surface point is said to be blocked by the ligand if its normal intersects (is close to) at least one atom of the ligand. The surface points and their normals are generated using Connolly's program²⁶ The obtained binding sites are generally identical (or very similar) to those derived with the CSU software³⁰ that analyzes the interatomic contacts in protein complexes.

The ligands in the data set form a very heterogeneous set, including sugars, co-factors, substrate analogs, peptides, etc. They also show great variability in the size and shape of their binding sites. The number of atoms in the binding sites varies from 3 to 141, where the binding site of ligand NAG-21 in the complex 1o7d has only 3 atoms, and that of ligand CDN in the complex 1nek has 141 atoms.

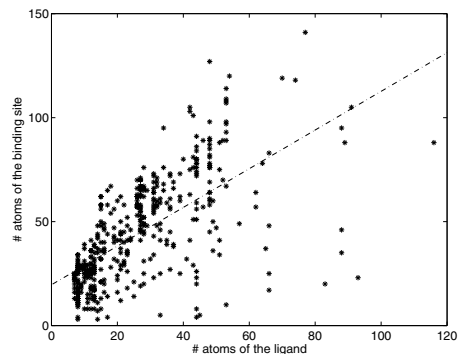


Fig. 3. The figure plots the number of atoms of the binding sites versus the number of atoms of the ligands for all 244 proteins of the dataset. The dotted line is the least square line.

Although there is a correlation between the number of atoms of the binding sites and of the ligands, as shown in Fig. 3, the binding sites of the same ligand and with different proteins may vary significantly in size. For example, the binding sites of ligand MPD in protein complexes 1d3c, 1h6g, 1hty, 1i78, 1lvo, 1nvm, 1oo0, 1srq consist of a number of atoms ranging from 3 to 28. A ligand can have more than one binding site with the same protein, and these binding sites can also vary considerably in size. The ligand UPL (unknown branched fragment of phospholipid) has 27 binding sites on the same protein (1lsh), of which the smallest has only 4 atoms, while the biggest has 56 atoms. The ligand of the dataset that shows the largest variability is FAD (flavin-adenine dinucleotide), where the biggest of its 11 binding sites has 114 atoms and the smallest has just 10 atoms.

Our cavity detection algorithm was run on the whole data set of 244 proteins. For each protein, it returned all cavities with more than a threshold number of atoms, ranked according to the number of atoms they contain. Thus rank one identifies the largest cavity, rank two the second largest cavity, and so on. This number is taken as an approximate measure of extension of the cavity. The number of cavities found on a protein vary considerably, depending on the size of the protein and its shape. In analyzing our solutions, we use the measure of *coverage* of the residues (atoms) of the binding site, i.e. the percentage of residues (atoms) of the binding site found in the cavity. A residue belongs to a cavity if at least one of the surface points close to it belongs to the cavity.

If the binding site of a ligand is known, we call

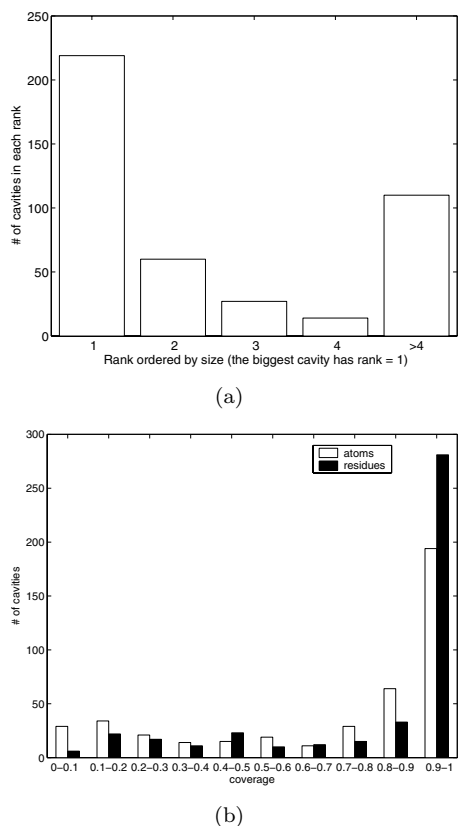


Fig. 4. 4(a) distribution of rank of cavities containing the ligand. 4(b) coverage of binding sites.

best-coverage cavity the cavity with the biggest coverage (in terms of atoms) of the binding site. In discussing our results, we consider only the best-coverage cavity for each complex of the dataset, and refer to it simply as cavity in the following.

Fig. 4(a) shows the distribution of ranks of best-coverage cavities (those containing the ligand). Of the 464 binding sites, 224 are in the largest cavity. As shown in Fig. 4(b), the values of coverage of residues of the binding sites are generally very good, with the majority of cavities achieving a coverage above 90%. This is true also for the coverage of the atoms of the binding site, even though such values are generally lower than those obtained for residues.

The results of our procedure for the whole dataset are available at <http://www.unipd.it/~garuttic/cavity/cavities07.xml>. Fig. 4(a) shows the distribution of the best-coverage cavities according to their rank. It can be seen that in most cases our method identifies the binding site in the biggest cavity. Moreover, according to Fig. 4(b), we

can infer that most of the times the binding site is completely included in the cavity. In Tab. 1 we show the top 20 cavities according to their values of coverage. The values of coverage for SURFNET-Consurf are not reported in this and in the other tables because they are not available. Thus, all these cavities tightly include the binding site, and in the first seven cases they coincide with it. It can be seen that, for these 20 entries, we locate the binding site in one of the four biggest cavities on 14 cases out of 20, which is competitive with the 8 out of 20 of SURFNET-Consurf. Moreover, in all the entries but one, our procedure find that the best-coverage cavity has rank less than or equal to that of SURFNET-Consurf. The only exception is for protein 1p6o with ligand HPY-411, but it can be noted that this protein has several cavities with similar dimensions, and thus the ranking can be significantly different even with similar algorithms. Tabs. 2 and 3 show the top 20 cavities according to their size, defined as cavity volume and number of atoms of the cavity, respectively. The results of Tab. 2 do not show any significant differences between the two methods, since all the cavities but two have rank one and big size in both methods. The two exceptions are complex 1ei6 with ligand PPF-412 (chain D), and complex 1r72 with ligand NAD-5. In the first case we find a small cavity that completely includes the binding site, which can be considered an improvement with respect to the big cavity found with SURFNET-Consurf, while in the second case the small cavity found has a 25% coverage on a binding site of 8 atoms and thus contains only two atoms of the binding site. The results of Tab. 3 show the biggest cavities that we find. They all have rank one, high coverage, and a considerable number of atoms (more than 600). Also the cavities found with SURFNET-Consurf have big size, but eight of them have rank higher than one, which suggests that these cavities are smaller than ours. This analysis suggests that the results that we obtain are close to those of SURFNET-Consurf, with a fast and still accurate geometrical method, without including any information about residues conservation.

From the analysis of the results, we can observe that for ligands with a large number of atoms in contact, our procedure identifies the binding site in a cavity with rank lower than four in most cases; otherwise it tends to find the binding site in a smaller cavity with rank larger than four (see

Table 1. The 20 cavities with the best values of coverage found by our procedure. *PdbID* is the ID of the complex in the PDB. *Chain* is the chain used in the experiment. *Rank* is the identifier of the cavity of the protein with the best-coverage of the binding site. *Cov* and *# Atoms of the cavity* refer to the best-coverage cavity. *Cov* is the coverage expressed in terms of atoms. *# Atoms of the b.s.*, *# Atoms of the ligand* and *Name of the ligand* refer to the ligand as indicated in the PDB. *Ligand name* is expressed in the format `resname:chain:seqnumber`.

PdbID	Chain	Rank	Rank SURFNET- ConSurf	Cov	#Atoms of the b.s.	#Atoms of the cavity	Cavity Vol in SURFNET- ConSurf (\AA^3)	#Atoms ligand	Ligand
1ejj	A	4	> 4	1.00	24	24	NA	11	3PG::601
1fw9	A	2	4	1.00	25	25	189	10	PHB::199
1h2r	SL	> 4	> 4	1.00	16	16	NA	8	NFE::1004
1l9g	A	3	> 4	1.00	25	25	NA	8	FS4::201
1p6o	AB	2	> 4	1.00	18	18	NA	8	HPY::410
1p6o	AB	2	1	1.00	18	18	279	8	HPY::411
1qft	A	2	> 4	1.00	27	27	NA	8	HSM::173
1otw	AB	> 4	> 4	1.00	42	46	NA	24	PQQ::501
1p0z	A	2	4	1.00	38	42	366	13	FLC::1632
1o7d	ABCDE	> 4	> 4	1.00	26	29	NA	8	TRS:A:2
1otw	AB	> 4	> 4	1.00	42	48	NA	24	PQQ::500
1lrh	AD	3	> 4	1.00	37	44	NA	14	NLA::8190
1lrh	AD	> 4	> 4	1.00	35	42	NA	14	NLA::5190
1r9l	A	2	2	1.00	29	40	292	8	BET::1001
1i9g	A	1	2	1.00	62	90	1141	27	SAM::301
1l5j	A	> 4	> 4	1.00	25	37	NA	7	F3S::868
1dl5	AB	3	4	1.00	63	95	748	26	SAH::1699
1hnn	A	1	3	1.00	63	101	358	26	SAH::2001
1us5	A	1	> 4	1.00	29	48	NA	10	GLU:A:1315
1o0r	A	1	1	1.00	72	120	1284	36	GDU::404

Fig. 5(a) and Fig. 5(b)). Consider the case of ligand MPD (2-METHYL-2,4-PENTANEDIOL) binding to 14 chains of 8 different proteins. When the binding site is large, as in the complex 1srq where it consists of 89 atoms, then it is found in the cavity with rank one; by contrast, in the complex 1d3c with only 12 atoms in contact, the binding site is found in the cavity ranked 14. Among the 210 cavities with rank one, 142 have a binding site with more than 40 atoms (see Fig. 5(a)). There are few ligands for which the binding sites are approximately of the same size. An example is ligand ATP whose binding sites are about 40 atoms and are, in all cases, contained in the top cavity, with rank one.

Fig. 5 shows the distribution of binding sites (ligands) by cavity rank and number of atoms of binding site (ligand). The bigger the number of atoms of the binding site, the better the rank of the corresponding cavity. In fact, on 88 binding sites that have less than 20 atoms, only 17 binding sites lie in the biggest cavity, 5 in the second biggest cavity, two in the third and one in the fourth, while 63 binding sites are located in a cavity smaller than the fourth. The results improve if the number of atoms of the

binding site increase. Thus 64% of the binding sites that have 20 or more atoms but less than 40 lie in one of the four biggest cavities, and this percentage increases to 88% for the binding sites that have 40 or more atoms but less than 60 and 95% for the binding sites that have 60 or more atoms but less than 80. Finally, all but four of the 29 binding sites that have 80 or more atoms but less than 100 lie in one of the three biggest cavities, and all the 14 binding sites that have 100 atoms or more lie in the biggest cavity. Fig. 5(b) shows analogous results for the ligands.

The biggest cavity does not contain any binding site in 80 of the 244 proteins considered in the experiments. For example, 1b11 (feline immunodeficiency virus protease complexed with T1-3-093) has a binding site with ligand INT in the cavity with rank two, while the cavity with rank one does not contain any ligand (see Fig. 6(a)). The ligands are located close to β -sheets 53-57, 62-68, 89-92 and 37-39, while the biggest cavity extends from the N-terminal valine to residue 114 close to C-terminal methionine, including residue 108 of alpha-helix 104-110. Also the ligand C8E in the complex 1bxw is not located in the largest cavity (see Fig. 6(b)). The largest cavity is

Table 2. The 20 cavities with the biggest cavity volume according to SURFNET-ConSurf. *PdbID* is the ID of the complex in the PDB. *Chain* is the chain used in the experiment. *Rank* is the identifier of the cavity of the protein with the best-coverage of the binding site. *Cov* and *# Atoms of the cavity* refer to the best-coverage cavity. *Cov* is the coverage expressed in terms of atoms. *# Atoms of the b.s.*, *# Atoms of the ligand* and *Name of the ligand* refer to the ligand as indicated in the PDB. *Ligand name* is expressed in the format **resname:chain:seqnumber**.

PdbID	Chain	Rank	Rank SURFNET- ConSurf	Cov	#Atoms of the b.s.	#Atoms of the cavity	Cavity Vol in SURFNET- ConSurf (\AA^3)	#Atoms ligand	Ligand
1n35	A	1	1	0.92	49	759	19763	28	CH1::1291
1n35	A	1	1	0.89	45	759	19763	28	CH1::1295
1n35	A	1	1	0.81	31	759	19763	28	CH1::1294
1l3i	ABCD	1	1	0.97	62	1080	12820	26	SAH::803
1l3i	ABCD	1	1	0.97	58	1080	12820	26	SAH::802
1l3i	ABCD	1	1	0.96	57	1080	12820	26	SAH::804
1l3i	ABCD	1	1	0.93	57	1080	12820	26	SAH::801
1p91	AB	1	1	0.97	67	632	10221	27	SAM::1401
1p91	AB	1	1	0.95	63	632	10221	27	SAM::2401
1f48	A	1	1	0.96	57	689	8993	27	ADP::590
1f48	A	1	1	0.90	50	689	8993	27	ADP::591
1sr9	AB	1	1	1.00	30	412	8477	8	KIV::701
1itw	A	1	1	0.96	27	325	8213	13	ICI:A:743
1jv1	AB	1	1	0.95	62	1499	6810	39	UD1::901
1ei6	AD	4	1	1.00	24	63	6643	7	PPF:D:412
1p0h	A	1	1	0.71	76	247	6351	48	COA::601
1eyr	AB	1	1	0.89	47	380	6322	50	CDP::1001
1eyr	AB	1	1	0.81	47	380	6322	50	CDP::2001
1r72	AB	3	1	0.25	8	32	6224	44	NAD::5
1ueu	A	1	1	0.88	48	239	5745	29	CTP::501

at the bottom of a β -barrel, while the ligand sticks outside from the center of the barrel and does not have a geometrically tight binding with the protein. In both cases our biggest cavities coincide with those found by the CASTp server (<http://sts.bioengr.uic.edu/castp>), which is also based on geometric criteria only.^{9,10}

5.2. Finding similar binding sites on two proteins

We benchmarked our method on several pairs of proteins or chains from another representative set.^{11,2} The set includes 46 proteins, 12 proteins with a chain binding to ATP and 10 with a chain binding to other adenine-containing ligands. Other proteins are from diverse functional families that can bind estradiol, equilin and retinoic acid. Other different protein families from the set are: HIV-1, anhydrase, antibiotics, fatty acid-binding proteins, chorismate mutases and serine proteases. In analyzing our solutions, we use the measure of coverage, i.e. the percentage of residues of the binding site found in the solution, and of accuracy, i.e. the percentage of residues in the solution that belong to the active site. A residue be-

longs to a solution if at least one of the surface points close to it belongs to the solution.

We performed comparisons of a query protein or chain surface with other proteins of the data set of 46 proteins or chains to retrieve those with high score when matched with the query. The score of a comparison is defined as the number of correspondences between points on the pair of matching regions identified on two cavities. We also compute the root mean square deviation (rmsd) of the rigid transformation that best aligns the corresponding points in the pair of regions for the two surfaces. The results shown here are obtained using the Catalytic Subunit of cAMP-dependent Protein-Kinase (pdb code 1atp, chain E) as query protein. This chain binds ATP. As already observed in the previous section, the ATP binding pockets in different proteins show great structural variability, although their size in terms of number of atoms/residues is about the same.

In Tab. 4 we show the values of coverage and accuracy obtained when comparing the cavity with rank one of 1atp with those of proteins 1phk, 1csn, 1mjh, 1hck and 1nsf. For the same pairs of proteins, we show also the values of coverage of the binding

Table 3. The 20 cavities with the biggest number of cavity atoms according to our procedure. *PdbID* is the ID of the complex in the PDB. *Chain* is the chain used in the experiment. *Rank* is the identifier of the cavity of the protein with the best-coverage of the binding site. *Cov* and *# Atoms of the cavity* refer to the best-coverage cavity. *Cov* is the coverage expressed in terms of atoms. *# Atoms of the b.s.*, *# Atoms of the ligand* and *Name of the ligand* refer to the ligand as indicated in the PDB. *Ligand name* is expressed in the format **resname:chain:seqnumber**.

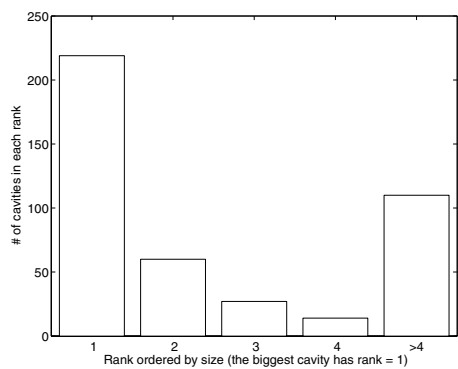
PdbID	Chain	Rank	Rank SURFNET- ConSurf	Cov	#Atoms of the b.s.	#Atoms of the cavity	Cavity Vol in SURFNET- ConSurf (\AA^3)	#Atoms ligand	Ligand
1jv1	AB	1	1	0.95	62	1499	6810	39	UD1::901
1jv1	AB	1	3	0.97	60	1499	3746	39	UD1::902
1l3i	ABCD	1	1	0.97	62	1080	12820	26	SAH::803
1l3i	ABCD	1	1	0.97	58	1080	12820	26	SAH::802
1l3i	ABCD	1	1	0.96	57	1080	12820	26	SAH::804
1l3i	ABCD	1	1	0.93	57	1080	12820	26	SAH::801
1m98	AB	1	3	1.00	103	775	1334	42	HEQ::351
1m98	AB	1	2	0.98	105	775	1311	42	HEQ::350
1m98	AB	1	2	0.74	35	775	1311	23	SUC::401
1nek	ABCD	1	3	0.88	141	766	2211	77	CDN::308
1nek	ABCD	1	3	0.85	52	766	2211	23	UQ2::306
1n35	A	1	1	0.92	49	759	19763	28	CH1::1291
1n35	A	1	1	0.89	45	759	19763	28	CH1::1295
1n35	A	1	1	0.81	31	759	19763	28	CH1::1294
1lvo	AB	1	3	0.86	28	714	1426	8	MPD::4002
1lvo	AB	1	4	0.85	27	714	892	8	MPD::4001
1f48	A	1	1	0.96	57	689	8993	27	ADP::590
1f48	A	1	1	0.90	50	689	8993	27	ADP::591
1f2u	ABCD	1	1	1.00	72	670	3526	31	ATP:A:901
1f2u	ABCD	1	1	0.99	69	670	3526	31	ATP:C:901

Table 4. Comparison of 1atp (cAMP-dependent Protein-Kinase) with 1phk (Subunit of glycogen phosphorylase kinase), 1csn (Casein kinase-1), 1mjh:B ("Hypothetical" protein MJ0577), 1hck (Cyclin dependent PK) and 1nsf (Examerization domain of N-ethylmaleimide-sensitive fusion protein).

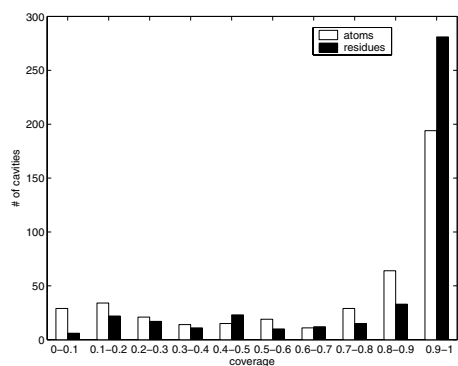
Pdb ID	# residues in binding site	Coverage MolLoc ²	Coverage Cavity comparison	Accuracy Cavity comparison
1atp	23	78%	91%	80%
1phk	26	69%	90%	76%
1atp	23	70%	78%	75%
1csn	26	62%	80%	91%
1atp	23	26%	34%	100 %
1mjh:B	25	24%	32%	88%
1atp	23	39%	56%	92 %
1hck	24	42 %	58 %	87 %
1atp	23	43%	60%	93%
1nsf	23	35%	43%	76%

site obtained by the comparison method based on spin-images² and here referred to as *MolLoc*. We do not report the accuracy values for MolLoc; although the solution regions had a significant overlap with the binding sites, they spanned areas much larger than the binding sites. Indeed the goal of MolLoc

was to identify similar regions on protein surfaces, not to find binding sites. For the proteins 1atp and 1csn, which both bind to the ligand ATP, the two most similar regions on each protein are part of the binding site and this explains also the high values of coverage for MolLoc. In both proteins, the binding



(a)



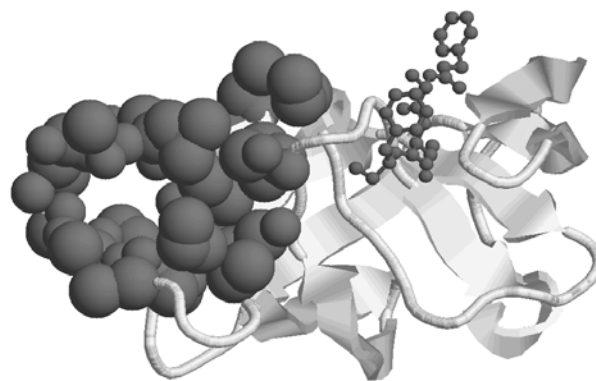
(b)

Fig. 5. Distribution of binding sites (ligands) by cavity rank and number of atoms of the binding site (ligand).

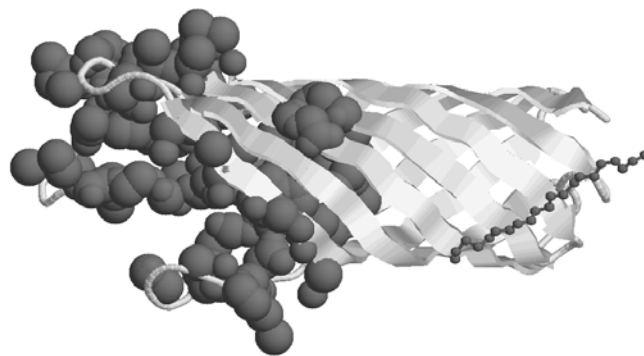
sites are located in the top cavity. The new method improves on coverage while at the same time obtaining a good accuracy for all pairwise comparisons. The execution time is drastically reduced w.r.t. MolLoc. While MolLoc took about two hours to execute, the new method took less than two minutes.

There are cases when we cannot expect our algorithm to identify the common regions that correspond to the active sites on a pair of cavities. However, if a large cavity is broken into several smaller cavities by physico-chemical considerations about binding sites, then one runs the risk of losing part of the binding site, which will make it harder to identify common binding sites when comparing cavities in two proteins.

From the observations in the previous section about the difference in size of different binding sites for the same ligand, it is evident that any matching procedure based on purely geometric criteria will fail to recognize binding sites for those cases.



(a)



(b)

Fig. 6. Proteins 1b11(6(a)) and 1bxw(6(b)). The biggest cavities are displayed in spacefill.

6. CONCLUSIONS

We have presented a method for binding site recognition that is effective and fast. It uses only geometric criteria and a description of the protein surfaces by means of a collection of two-dimensional arrays, the spin images, each describing the spatial arrangement of the protein surface points in the vicinity of a given surface point. As mentioned, there are several cases where our recognition procedure fails to identify the correct binding sites. When a ligand binds different proteins at sites that vary significantly in size and shape, most of existing approaches are inadequate to identify the binding location. The problem is further complicated by the simultaneous presence of several ligands within the same cavity. We think our work can contribute one more step towards the solution of

the problem, when only geometric features are considered.

REFERENCES

1. M. E. Bock *et al.*, *Proc. Combinatorial Pattern Matching CPM 2005*, 417–428 (2005).
2. M. E. Bock *et al.*, *J. Comp. Biol.* **14**(3), in press (2007).
3. F. Glaser *et al.*, *Comput. Syst. Bioinformatics Conf.* **62**, 479–488 (2006).
4. G. P. Brady Jr and P. F. Stouten, *J. Computer Aided Mol. Des.* **14**, 383–401 (2000).
5. R. A. Laskowski, *J. Mol. Graph.* **13**, 323–330 (1995).
6. I. D. Kuntz *et al.*, *J. Mol. Biol.* **161**(2), 269–288 (1982).
7. R. A. Laskowski *et al.*, *J. Mol. Biol.* **351**, 614–626 (2005).
8. D. G. Levitt and L. J. Banaszak, *J. Mol. Graphics* **10**, 229–234 (1992).
9. J. Liang *et al.*, *Proteins* **33**, 1–17 (1998).
10. J. Liang *et al.*, *Proteins* **33**, 18–29 (1998).
11. A. Shulman–Peleg *et al.*, *J. Mol. Biol.* **339**, 607–633 (2004).
12. R. J. Morris *et al.*, *Bioinformatics* **21**(10), 2347–2355 (2005).
13. A. Brakoulias and R. M. Jackson, *Proteins* **56**, 250–260 (2004).
14. B. Y. Chen *et al.*, *Comput. Syst. Bioinformatics Conf.*, 311–323 (2006).
15. J. A. Barker and J. M. Thornton, *Bioinformatics* **13**, 1644–1649 (2003).
16. T. A. Binkowski *et al.*, *J. Mol. Biol.* **332**, 505–526 (2003).
17. T. A. Binkowski *et al.*, *Prot. Sci.* **14**, 2972–2981 (2005).
18. N. Kinoshita *et al.*, *J. Struct. Funct. Genomics* **2**, 9–22 (2001).
19. G. Kleywegt, *J. Mol. Biol.* **285**, 1887–1897 (1999).
20. N. Kobayashi N. and Go, *J. Mol. Biol.* **26**, 135–144 (1997).
21. Y. Y. Kuttner *et al.*, *Proteins: Struct. Funct. Bioinf.* **52**, 400–411 (2003).
22. L. Lo Conte *et al.*, *J. Mol. Biol.* **285**, 1021–1031 (1999).
23. R. Najmanovich *et al.*, *Bioinformatics* **23**(2), 104–109 (2007).
24. A. Via *et al.*, *J. Mol. Biol.* **57**, 1970–1977 (2000).
25. H. Yao *et al.*, *J. Mol. Biol.* **326**, 255–261 (2003).
26. M. L. Connolly, *J. Appl. Cryst.* **16**, 548–558 (1983).
27. A. E. Johnson and M. Hebert, *IEEE Trans. Patt. Anal. Machine Intell.* **21**(5), 433–449 (1999).
28. R. A. Laskowski, *Nucleic Acids Res.* **29**, 221–222 (2001).
29. The UniProt Consortium, *Nucleic Acids Res.* **35**, D193–197 (2007).
30. V. Sobolev *et al.*, *Bioinformatics* **15**, 327–332 (1999).
31. H. M. Berman *et al.*, *Nucl. Acids Res.* **28**, 235–242 (2000).