BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 3 2011, pages 351–358
doi:10.1093/bioinformatics/btq672

Structural bioinformatics

Advance Access publication December 6, 2010

MSPocket: an orientation-independent algorithm for the detection of ligand binding pockets

Hongbo Zhu* and M. Teresa Pisabarro*

Structural Bioinformatics, BIOTEC Technical University of Dresden, Tatzberg 47-51, 01307 Dresden, Germany

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: Identification of ligand binding pockets on proteins is crucial for the characterization of protein functions. It provides valuable information for protein–ligand docking and rational engineering of small molecules that regulate protein functions. A major number of current prediction algorithms of ligand binding pockets are based on cubic grid representation of proteins and, thus, the results are often protein orientation dependent.

Results: We present the MSPocket program for detecting pockets on the solvent excluded surface of proteins. The core algorithm of the MSPocket approach does not use any cubic grid system to represent proteins and is therefore independent of protein orientations. We demonstrate that MSPocket is able to achieve an accuracy of 75% in predicting ligand binding pockets on a test dataset used for evaluating several existing methods. The accuracy is 92% if the top three predictions are considered. Comparison to one of the recently published best performing methods shows that MSPocket reaches similar performance with the additional feature of being protein orientation independent. Interestingly, some of the predictions are different, meaning that the two methods can be considered complementary and combined to achieve better prediction accuracy. MSPocket also provides a graphical user interface for interactive investigation of the predicted ligand binding pockets. In addition, we show that overlap criterion is a better strategy for the evaluation of predicted ligand binding pockets than the single point distance criterion.

Availability: The MSPocket source code can be downloaded from http://appserver.biotec.tu-dresden.de/MSPocket/. MSPocket is also available as a PyMOL plugin with a graphical user interface.

Contact: hongboz@biotec.tu-dresden.de; mayte@biotec.tu-dresden.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on June 18, 2010; revised on November 30 , 2010; accepted on December 1, 2010

1 INTRODUCTION

The prediction of ligand binding sites on proteins provides important information for protein–ligand docking and structural-based rational engineering of small molecules that modulate protein functions (Campbell *et al.*, 2003; Sotriffer and Klebe, 2002). Furthermore, comparative analysis of ligand binding pockets is found to provide valuable information for the understanding of protein–ligand binding specificity (Chen and Honig, 2010).

It has been observed that ligand binding sites often locate in the largest pockets on protein surfaces (London *et al.*, 2010; Nayal and Honig, 2006). Thus, the identification of pockets on protein surfaces plays a key role in the prediction of protein functional sites, in particular, ligand binding sites. A variety of computational approaches have been proposed for the prediction of ligand binding pockets. These methods can be divided into two categories according to the information they utilize to detect pockets: *geometric approaches* that are purely based on the geometric characteristics of proteins, and *comprehensive approaches* that not only consider geometric criteria but also take into account evolutionary information, interaction energy or chemical properties of proteins. A major number of these methods, in both categories, are based on the cubic grid representation of protein structures. Geometric methods like POCKET (Levitt and Banaszak, 1992), LIGSITE (Hendlich *et al.*, 1997) and LIGSITE^{cs} (Huang and Schroeder, 2006) generate 3D grids for proteins and identify surface pockets as the set of solvent grid points that are situated between protein grid points. PocketPicker (Weisel *et al.*, 2007) uses grids to represent proteins and search the environment of each surface grid along 30 directions for defining pockets. Tripathi and Kellogg (2010) introduced the VICE program as part of the HINT toolkit (Kellogg *et al.*, 2005). Similar to PocketPicker, VICE scans grid points along the path in various directions at each grid points and defines pocket grids as those with at least half of the scan directions 'blocked'. The VICE program represents proteins as binary grid maps, in which grid points occupied by atoms are set to one and the rest zero, such that the VICE algorithm is performed on only integers and thus is very efficient. Yu *et al.* (2010) suggested the Roll algorithm, in which a probe sphere of radius 2 Å is used to roll on each slice of the 3D grid representations of proteins. Pockets are defined to be the regions between the probe sphere and the protein surface.

The grid representation of proteins is dependent on the orientation of proteins in the coordinate system. Inconsistent results may be observed for grid-based methods if the atomic coordinates of proteins are transformed. One solution to address the problem of inconsistent results is to increase the grid resolution and generate finer grid representations. However, this is at the cost of decreased efficiency of the methods. Alternatively, geometric methods that do not use cubic grid systems have also been suggested. SURFNET (Laskowski, 1995) places probe spheres in gaps between protein surface atoms and defines pockets using the probe spheres. CAST (Liang *et al.*, 1998) represents a protein as Delaunay triangulation and derives the alpha shape of the protein

*To whom correspondence should be addressed.

to detect pockets using discrete-flow theory. PASS (Brady and Stouten, 2000) coats a protein using probe spheres repeatedly and searches pockets that are filled with buried probes. Petsalaki *et al.* (2009) proposed a novel idea for predicting ligand binding sites. First, spatial position specific scoring matrices (S-PSSMs) are derived from known protein–ligand complexes. Then, the surface of the putative binding protein is scanned using the S-PSSMs to locate potential ligand binding sites. The method reaches a positive predictive value of 85% on a benchmark dataset of 405 known protein–ligand complexes. Fpocket (Le Guilloux *et al.*, 2009) uses alpha spheres (Liang *et al.*, 1998) to fill the space within and around the protein. Each alpha sphere is defined by four atoms that are in touch with the sphere and contains no atoms inside it. In addition, each alpha sphere is labeled as polar or apolar according to the physicochemical property of the atoms with which it is in contact. Pockets are predicted to correspond to the ensembles of alpha spheres of intermediate radii, which are required to be predominantly apolar.

Here we present MSPocket (Molecular Surface Pocket), a novel geometric program for searching pockets on protein solvent excluded surfaces. MSPocket does not employ any cubic grid representation of proteins. Thus, MSPocket results are inherently independent of the orientations of proteins in coordinate systems in terms of its core algorithm. MSPocket identifies surface pocket regions according to the normal vector directions at the vertices on the surface. In an evaluation to predict ligand binding pockets, MSPocket achieves an accuracy of 75% on a dataset used in the evaluation of several existing methods. If top three predictions are considered, the accuracy is 92%. There are currently three approaches that reach the best performance in terms of detecting ligand binding pockets on a benchmark dataset of proteins. Among the three approaches, the program Fpocket (Le Guilloux *et al.*, 2009) not only applies geometric criteria, but also considers the electronegativity of protein atoms to prune pockets. The other two approaches use pure geometric criteria: VICE (Tripathi and Kellogg, 2010) and POCASA (Yu *et al.*, 2010). In VICE, protein structures were first processed by performing molecular modeling and adding hydrogen atoms using the Sybyl program suite (www.tripos.com). VICE does not work as an independent program and needs Sybyl to function at the moment (personal contact with the VICE authors). POCASA uses the original atomic coordinates of proteins as input, and it is freely available. We consider POCASA as our reference approach and perform a thorough comparison with it. The comparison results show that MSPocket achieves comparable performance to POCASA with the additional feature of being orientation independent. In addition, some results of MSPocket are observed to be different from those of POCASA. This suggests that the two approaches may be adopted as complementary tools to achieve higher success rate. Moreover, in contrast to the widely used evaluation method of calculating the minimum distance between the mass center of predicted pockets and the ligand atoms, we demonstrate that measuring the overlap between the ligand and predicted pockets provides a more comprehensive assessment of prediction results.

2 METHODS

MSPocket detects pockets on the solvent excluded surface (SES) computed by using MSMS, which is a tool for efficient computation of the analytical model of protein SES (Sanner *et al.*, 1996). MSMS generates a set of surface vertices as a sampling of the protein SES. Each surface vertex is associated with an atom in the protein. The surface vertices are triangulated. The coordinates and the normal vectors of all surface vertices are produced, which are utilized by MSPocket for identifying surface vertices located in concave regions on the SES.

2.1 Workflow

The workflow of the MSPocket approach is described in the following steps:

(i) Calculating protein SES using MSMS. Normally, in the SES generated by MSMS, there is an *external component* and several *internal components*. The *external component* of the SES is the exterior region of the protein SES. The *internal components* are normally interior regions surrounded by the protein (see Fig. 2b). In the MSPocket algorithm, internal components are directly reported as *cavities*. The following steps are only applied on the external component of the protein SES for detecting *pockets*.

(ii) Sampling surface vertices. In order to reduce computational complexity, we simplify the surface triangulation by sampling the surface vertices generated by MSMS. To this end, we consider the surface triangulation as a graph $G(V,E)$, where V denotes the set of all surface vertices, and E denotes the set of all edges between the vertices in V. First, for every vertex v we compute the angle between its normal vector and the normal vector of each adjacent vertex of v. Then, we calculate the average value of all the angles and assign it to v as its *average angle* if all the angles are smaller than 90 degrees. Otherwise the average angle of v is set to infinity. Next, we traverse all the vertices in ascending order of average angle. Our aim here is to reduce the number of vertices in G by choosing a subset of V as representative vertices. During the traverse of the vertices, if we find that a vertex v_c and all its neighbors $N_G(v_c)$ have not been traversed yet, then v_c is sampled as the *representative* vertex for $v_c \cup N_G(v_c)$, and $v_c \cup N_G(v_c)$ are all marked as *traversed*. After the traversal is finished, all remaining non-traversed vertices are also taken as representative vertices. We then construct a new graph $G'(V',E')$, where V' is the set of all representative vertices. Two vertices in G' are adjacent if they are adjacent in G, or one of them is adjacent to any neighbor of the other vertex in G, or their neighbors are adjacent in G. The number of vertices is reduced to approximately 25% of the original number of vertices using this sampling procedure. In the following steps, G' is used for detecting pocket vertices.

(iii) Identifying *pocket vertex pairs* (PktVPs) and *protrusion vertex pairs* (PtrVPs). First of all, two surface vertices are considered to be a *close vertex pair* (ClsVP) if the distance between them is less than or equal to a cutoff d_p. We then identify PktVPs and PtrVPs in ClsVPs (see Fig. 1). A ClsVP (A,B) is moved along the directions of their respective normals for a short distance (0.2 Å). We consider (A,B) to be a PktVP if $d_{AB} - d_{A'B'} > 0.2 \cdot r$, where r is a distance change ratio parameter that reflects how much closer A' and B' have moved toward each other ($-2.0 \le r \le 2.0$). The larger the r value, the more restrictive it is for defining PktVP. The two vertices in a PktVP are regarded as potential pocket vertices, and they are termed the *pocket partners* of each other. For defining PtrVPs, we consider the angle α between the normal $\overrightarrow{N_A}$ of A and the vector $\overrightarrow{AB}$, and the angle β between the normal $\overrightarrow{N_B}$ of B and the vector $\overrightarrow{BA}$. A ClsVP (A,B) is regarded as a PtrVP if $\alpha + \beta > a_p$, where a_p is an angle cutoff parameter.

(iv) Identifying pockets. We cluster pocket vertices based on the adjacency between vertices in G' to define pockets. Here, we construct a graph $G''(V'',E'')$, where V'' is the set of all potential pocket vertices identified in step (iii) and E'' is the set of edges between potential pocket vertices. In G'', two vertices are adjacent if they are adjacent in G' and they are not a PtrVP. Each connected component in G'' is

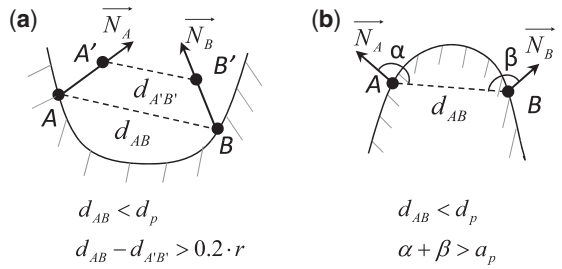

Fig. 1. Geometric methods for detecting PktVPs and PtrVPs. (**a**) Identifying PktVPs. Vertices A and B are a ClsVP if $d_{AB} < d_p$, where d_p is a distance cutoff. We consider (A,B) to be a PktVP if $d_{AB} - d_{A'B'} > 0.2 \cdot r$, where r is a distance change ratio. (**b**) Identifying PtrVPs. A ClsVP (A,B) is considered to be a PtrVP if $\alpha + \beta > a_p$, where a_p is an angle cutoff.

an induced subgraph of G'' and is considered to be a potential surface pocket. Then we prune each potential pocket in order to remove outlier vertices. We iteratively remove vertices that have no more than n_p pocket partners, or have no more than n_b neighbors in G'' until no more vertices are removed from the component. The vertices located at the bottom of pockets have few pocket partners and are often missed in the potential pockets. Therefore, we extend the pruned pockets by adding new vertices to them. A vertex v is added to a pruned pocket p if more than 50% of v's neighbors belong to p.

Till now, surface pockets are defined by representative vertices sampled in step (ii). Here, we replace each representative vertices by all the vertices it represents in G. In the end, each pocket is reported as an induced subgraph of G.

Note that all parameters used in MSPocket are adjustable by users, including distance cutoff d_p, angle cutoff a_p, distance change ratio r, pocket partner number cutoff n_p and neighbor number cutoff n_b. In this work, we set the parameters to $d_p = 8$ Å, $a_p = 200°$, $r = 1.3$, $n_p = 4$, $n_b = 1$.

2.2 Ranking of pockets

In MSPocket, the identified pockets may be ranked according to a variety of measures: the number of pocket vertices, number of pocket atoms associated with pocket vertices, pocket surface area, or pocket volume. Here, we introduce the methods for calculating pocket surface area and estimating the pocket volume.

2.2.1 Calculation of pocket surface area Each pocket identified by MSPocket is defined to be an induced graph of G, which is defined by the triangulation of the protein SES generated by MSMS (see Section 2.1). We calculate the area of each triangle formed by the pocket vertices, and compute the total area of all the triangles as the pocket surface area.

2.2.2 Estimation of pocket volume The pocket volume is estimated using two methods. In the first method, the volume of a pocket is calculated as the sum of the volumes of all tetrahedra formed between C_p, the mass center of all the pocket vertices and all triangles in the triangulation formed by the pocket vertices (see Fig. 2a and b). For each tetrahedron, we calculate the mass center of the triangle (C_t) and the average vector ($\overrightarrow{N_a}$) of the normal vectors at the three vertices in the triangle. The volume of the tetrahedron V_i is defined to be positive if the angle between $\overrightarrow{C_t C_p}$ and $\overrightarrow{N_a}$ is less than 90°, otherwise negative. The pocket volume V_e is defined to be $\sum V_i$.

Pockets detected by MSPocket on the external component of SES are normally not closed. Therefore, the volume V_e estimated using the first method is typically smaller than the pocket volume. To address this problem, we propose the second method for estimating pocket volumes more precisely. We aim to close each pocket by adding a 'cover' on top of the pocket. The

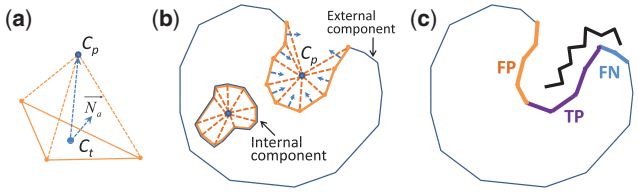

Fig. 2. Measurement of pocket volume (**a** and **b**) and measures for evaluating pocket predictions (**c**). (a) Each tetrahedron is composed of a triangle on the protein surface (triangle in solid line) and the mass center of all pocket vertices (C_p). All three vertices in the triangle must be pocket vertices. (b) The pocket volume V_e is defined to be the total volume of all tetrahedra in the pocket. (c) Overlap between predicted pocket atoms (in orange) and ligand binding site atoms (in blue) is highlighted in purple color. The ligand is represented as a black solid line segment. FP, false positive; TP, true positive; FN, false negative.

method for building a cover for each pocket is described in detail in the Supplementary Material. A more precise estimation of the pocket volume V_p is calculated by considering not only the tetrahedra used for calculating V_e, but also the tetrahedra formed by C_p and all triangles in the pocket covers. Note that covers are only build for pockets. Cavities are always closed and for them $V_e = V_p$.

By default, MSPocket uses V_e to rank pockets and cavities. Pockets and cavities are ranked separately. The computation of V_p is more time-consuming and it is not performed by default in MSPocket.

2.3 Evaluation dataset

We use the dataset collected by Huang and Schroeder (2006) for evaluating the prediction results of MSPocket. There are 48 bound structures and 48 corresponding unbound structures in the dataset. In each bound structure, there is at least one ligand binding to the protein. See Table 1 for the complete list of PDB entry codes and ligands. We refer to this dataset as Huang2006 dataset.

2.4 Evaluation method

There are different criteria proposed for the evaluation of identified pockets. One of the most widely applied criteria is the *single point distance criterion* (SPDc): a predicted pocket is regarded as correct if the *mass center* (MC) of the pocket is within 4 Å of any atom of the ligand binding to the protein (Huang and Schroeder, 2006; Le Guilloux *et al.*, 2009; Weisel *et al.*, 2007). Yu *et al.* (2010) adopted a revised version of SPDc. They measured the distance between the *depth centers* (DCs) of pockets to atoms in the ligand, rather than the commonly used MCs.

To deal with the limitations of SPDc, alternative strategies like *overlap criterion* (OVLc) have been proposed. Hendlich *et al.* (1997) calculated the overlap between the atoms that are in contact with ligands and the atoms predicted by LIGSITE to be part of the binding site. Tuffery and coworkers have proposed a *mutual overlap criterion* (MOc) to assess the performance of the Fpocket program (Le Guilloux *et al.*, 2009). Using MOc, a predicted pocket is considered correct if at least half of the ligand is covered by the pocket, and the pocket is not too big at the same time (at least 20% of the pocket is occupied by the ligand).

We argue that the SPDc has the following limitations compared with OVLc: (i) a single point is an oversimplified representation of a pocket. Distance between the single point and the ligand does not provide a quantitative measure of the overlap between the space occupied by the ligand and the predicted pocket. (ii) the assessment result depends on the definition of the single point. For example, the use of DC leads to slightly different evaluation of POCASA results from using MC (see Table 2). Specifically, for a purine nucleoside phosphorylase (1ULB), the top 1 pocket is considered to be a success if DC is used as the single point to represent pockets. However,

Table 1. Detailed pocket detection results using MSPocket

Bound	Ligand[a]	Rank[b] SPDc	Sens Top1 (%)	Prec[c] Top1 (%)	Sens Top3 (%)	Prec[d] Top3 (%)	Unbound	Rank[b] SPDc	Sens Top1 (%)	Prec[c] Top1 (%)	Sens Top3 (%)	Prec[d] Top3 (%)
1BID	UMP	1	96	24	96	24	3TMS	1	94	24	94	24
1CDO	NAD	3	79	21	79	21	8ADH	1	97	42	97	42
1DWD	MID	1	98	41	98	41	1HXF	1	95	32	95	32
1FBP	AMP/F6P	1	93	34	93	34	2FBP	1	92	25	92	25
1GCA	GAL	1	64	100	64	100	1GCG	1	75	62	75	62
1HEW	NAG	1	70	58	70	58	1HEL	1	76	59	76	59
1HYT	BZS	1	94	38	94	38	1NPC	1	76	35	76	35
1INC	ICL	1	93	50	93	50	1ESA	3	00	00	48	82
1RBP	RTL	1	100	77	100	77	1BRQ	1	81	60	81	60
1ROB	C2P	1	86	61	86	61	8RAT	1	93	41	93	41
1STP	BTN	1	91	91	91	91	1SWB	1	97	77	97	77
1ULB	GUN	0	97	19	97	19	1ULA	0	91	15	91	15
2IFB	PLM	1	97	51	97	51	1IFB	1	93	49	93	49
3PTB	BEN	2	00	00	92	66	3PTN	2	00	00	85	80
2YPI	PGA	2	17	10	86	65	1YPI	2	21	11	95	48
4DFR	MTX	1	96	32	96	32	5DFR	1	63	38	63	38
4PHV	VAC	1	96	78	96	78	3PHV	1	37	29	37	29
5CNA	MMA	7	27	32	27	32	2CTV	3	00	00	86	94
7CPA	FVF	1	81	69	81	69	5CPA	1	86	51	86	51
1A6W	NIP	3	00	00	93	90	1A6U	2	00	00	87	74
1ACJ	THA	0	98	29	98	29	1QIF	0	81	23	81	23
1APU	chainI	1	88	48	88	48	3APP	1	59	20	59	20
1BLH	FOS	1	93	85	93	85	1DJB	1	94	55	94	55
1BYB	GLC	1	93	44	93	44	1BYA	1	88	36	88	36
1HFC	PLH	1	84	81	84	81	1CGE	1	86	81	86	81
1IDA	PY2/QND/ VAL/PRO/PPL	1	91	71	91	71	1HSI	1	29	65	29	65
1IGJ	DGX	1	98	50	98	50	1A4J	2	00	00	91	60
1IMB	LIP	1	91	20	91	20	1IME	1	95	19	95	19
1IVD	ST1	1	100	29	100	29	1NNA	1	99	23	99	23
1MRG	ADN	2	00	00	72	61	1AHC	1	89	43	89	43
1MTW	DX9	2	00	00	64	71	2TGA	0	00	00	10	20
1OKM	SAB	1	77	54	77	54	4CA2	1	81	58	81	58
1PDZ	PGA	1	93	27	93	27	1PDY	1	90	21	90	21
1PHD	HEM/PIM	1	98	66	98	66	1PHC	1	97	65	97	65
1PSO	chainI	1	97	44	97	44	1PSN	1	97	52	97	52
1QPE	PP2	1	89	69	89	69	3LCK	1	71	34	71	34
1RNE	C60	1	98	51	98	51	1BBS	1	93	57	93	57
1SNC	THP	1	98	41	98	41	1STN	1	39	25	39	25
1SRF	MTB	1	80	96	80	96	1PTS	1	75	88	75	88
2CTC	HFA	1	96	47	96	47	2CTB	1	100	35	100	35
2H4N	AZM	1	98	50	98	50	2CBA	1	94	58	94	58
2PK4	ACA	1	88	100	88	100	1KRN	1	72	89	72	89
2SIM	DAN	1	96	45	96	45	2SIL	2	00	00	92	45
2TMN	LEP/NH2	1	96	40	96	40	1L3F	1	73	61	73	61
3GCH	OAC	0	03	06	91	23	1CHG	2	00	00	72	51
3MTH	MPB	0	00	00	00	00	6INS	0	00	00	0	0
5P2P	DHG	1	99	62	99	62	3P2P	1	65	81	65	81
6RSA	UVC	1	83	31	83	31	7RAT	1	57	63	57	63

[a]If there are multiple ligands in a protein structure file, they are considered separately in the evaluation (separated by forward slash). In 1APU, the whole chain I comprises the inhibitor isovaleryl (a pepstatin analogue) for an aspartyl proteinase. Thus, chain I is considered to be the ligand in 1APU. Similarly, in 1PSO, the whole chain I comprises the pepstatin in complex with human pepsin and is considered to be the ligand.
[b]Pockets/cavities are represented by their mass centers.
[c]Precision of the pockets/cavities with the best top 1 sensitivity.
[d]Precision of the pockets/cavities with the best top 3 sensitivity.

Table 2. Success rates in the prediction of pockets for 48 bound/unbound structures using SPDc

Method	Top 1		Top 3	
	Unbound (%)	Bound (%)	Unbound (%)	Bound (%)
MSPocket[a]	75	77	92	90
POCASA[b]	75 (73)	77 (77)	88 (85)	94 (94)
VICE[c]	83	85	90	94
Fpocket[d]	69	83	94	92

[a]MSPocket results are obtained using the same structure files used by POCASA. Pocket MCs are employed in the evaluation.
[b]POCASA results are taken from Yu *et al.* (2010) and are calculated using pocket DCs. As a comparison, an evaluation using pocket MCs (see Section 2.4) is given in parentheses.
[c]VICE results are taken from Tripathi and Kellogg (2010) and are calculated using the 'center-of-gravity' of pockets.
[d]Fpocket results are taken from Le Guilloux *et al.* (2009).

none of the pockets detected by POCASA is regarded correct if MC is used because none of the pocket MCs is within 4 Å of any atoms of the guanine molecule (see Fig. 3a). Similar results are also observed for an unbound structure (1QIF, see Fig. 3b). A counterexample is also obtained. For the top 1 pocket detected on 3GCH, the shortest distance is 4.2 Å between its DC and any ligand atom. This value is 2.8 Å if MC of the pocket is used (Fig. 3c). (iii) The SPDc is biased against the cases in which ligands do not locate in the center of the pockets. This is again illustrated by the two examples shown in Figure 3a and b.

In spite of the limitations of SPDc, we nevertheless evaluated the performance of MSPocket using SPDc and pocket MC. A pocket/cavity prediction was considered to be successful if the MC of the pocket is within 4.0 Å of any atom of the ligand. Both top 1 and top 3 best predictions were considered with respect to their volume V_e. Yu *et al.* (2010) distinguished pockets from cavities in the evaluation. Both pockets and cavities were considered in the evaluation of the Roll algorithm and the best result was reported. In other words, a prediction was regarded to be successful at top 1 if either the top 1 pocket or the top 1 cavity was correct. We followed the same rule in the evaluation of MSPocket results.

We also evaluated the MSPocket program using OVLc. To assess the performance of MSPocket, we examined the overlap between *ligand binding site atoms* and *pocket atoms* as proposed by Hendlich *et al.* (1997). An atom is considered to be part of the ligand binding site if its distance to any ligand atom is less than or equal to the sum of the van der Waals (vdW) radii of the atoms plus 1.0 Å, i.e. $r_{\text{ligand}}+r_{\text{protein}}+1.0$ (Zhu *et al.*, 2008). As for pocket atoms, they are derived from the detected pockets using distance criteria. In MSPocket, predicted pockets are reported as surface vertices. We define all atoms that are in proximity of 3.0 Å to any pocket vertex to be pocket atoms. This cutoff (3.0 Å) was chosen to be close to the upper bound of $r_{\text{protein}}+1.0$, such that the distance is at least $r_{\text{ligand}}+r_{\text{protein}}+1.0$ between the pocket atoms and atoms of the ligand that potentially binds at the pocket. To compare MSPocket results to those of POCASA, we defined atoms associated with POCASA pockets as the set of atoms in proximity of r to any pocket grid point. We evaluated the performance of POCASA using different r values (=3.0, 4.0, …, 10.0 Å). The performance of POCASA with respect to different r values is shown in Supplementary Figures S3 and S4. In the end, we decided to choose $r=5.0$ because POCASA reaches the best balanced performance using this value.

2.5 Evaluation measures

We use *LigBSA* to represent the ligand binding site atoms that are identified using the distance criteria in Section 2.4. Similarly, we use *PktBSA* to denote pocket atoms. The overlap between *LigBSA* and *PktBSA* is defined to be *true positive* (TP). The remaining parts of the *LigBSA* and *PktBSA* are defined to be *false negative* (FN) and *false positive* (FP) (see Fig. 2c). We define performance measures *sensitivity* and *precision* as follows for the evaluation of pocket prediction results:

$$\text{sensitivity}=\frac{\text{TP}}{\text{TP}+\text{FN}};\ \text{precision}=\frac{\text{TP}}{\text{TP}+\text{FP}}.$$

A high sensitivity reflects that the predicted pocket has a large overlap with the ligand. But if the high sensitivity is achieved by the excessive size of the pocket, the precision will be low. Similar to the MOc proposed by Le Guilloux *et al.* (2009), we define our OVLc using sensitivity $\geq 50\%$ and precision $\geq 20\%$.

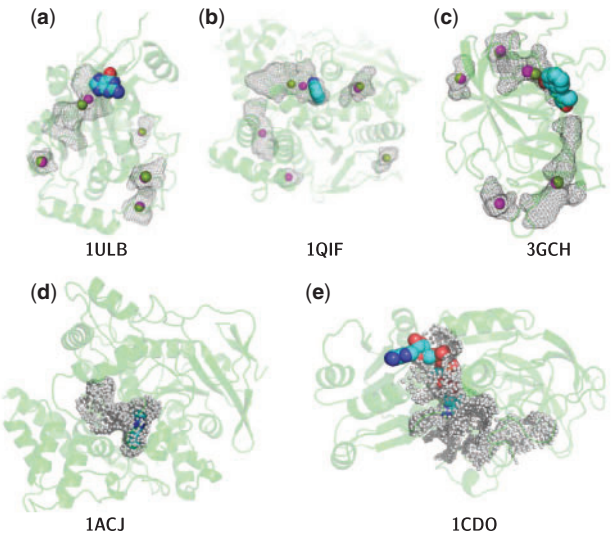

Fig. 3. Overlap between ligands and predicted pockets. Proteins are shown as green cartoons and ligands as spheres (carbon:cyan; nitrogen:blue; oxygen:red). Pockets detected by POCASA are shown as mesh in figure (**a**), (**b**) and (**c**). The DCs of the pockets are shown in magenta and MCs in green spheres. The top 1 pockets detected by MSPocket are shown as gray dots in figure (**d**) and (**e**). Each gray dot stands for a surface vertex in the pocket. (a) Pockets detected by POCASA on the surface of a purine nucleoside phosphorylase (1ULB). The ligand guanine (GUN) is contained in pocket top 1. The distance between the DC of pocket top 1 to the nearest atom in GUN is 3.4 Å, and for MC is 5.5 Å. (b) Pockets detected by POCASA on the surface of an acetylcholinesterase (1QIF). The structure of 1QIF has been superposed to a homologous structure (1ACJ) with a tacrine (THA) bound to it, which is contained in pocket top 1. The distance between the DC of pocket top 1 to the nearest atom in THA is 3.0 Å, and for MC is 6.0 Å. (c) Pockets detected by POCASA on the surface of a chymotrypsin (3GCH). The ligand cinnamate (OAC) is partially contained in pocket top 1. The distance between the DC of pocket top 1 to the nearest atom in OAC is 4.2 Å, and for MC is 2.8 Å. (d) The top 1 pocket detected by MSPocket on the surface of an acetylcholinesterase (1ACJ). (**e**) The top 1 pocket detected by MSPocket on the surface of an alcohol dehydrogenase (1CDO).

3 RESULTS

3.1 Evaluation results using SPDc

The performance of MSPocket using SPDc and pocket MC is reported in Table 2. The details of the MSPocket predictions are shown in Table 1. As a comparison, the performance of POCASA, VICE and Fpocket is also listed in Table 2. MSPocket

and POCASA were run on the same input data. The performance of POCASA using both MC and DC is reported. The performance of VICE and Fpocket is obtained from the respective publications. In this work, we considered POCASA as the reference program and focused on the comparison of MSPocket and POCASA.

The success rates of MSPocket for the unbound structures are 75 and 92% when the top 1 and top 3 predicted pockets are considered. These performances are better than POCASA, for which the corresponding success rates using pocket MC are 73 and 85%. For bound structures, MSPocket has the same success rate as POCASA if only top 1 of predicted pockets are considered. POCASA produces slightly better results when top 3 predicted pockets are considered. MSPocket results are worse than VICE except for unbound structures when top 3 predictions are considered. The comparison with Fpocket shows that Fpocket reaches better performance except for bound structures when top 1 predictions are considered.

After analyzing the detailed results on the Huang2006 dataset, we notice that some results of MSPocket are different from those of POCASA, although the overall accuracies of the two methods are similar. For the 48 bound structures, MSPocket and POCASA predict the ligand binding site at the same rank for 38 (79%) structures. For the 48 unbound structures, the ligand binding sites of 36 (75%) structures are identified at the same rank. If the top 1 predictions of both methods are combined, the accuracy is improved to 85 and 81% for bound and unbound structures, respectively. Similarly, if the top 3 predictions are combined, the accuracy is improved to 96% for both bound and unbound structures. See supplementary Table S1 for details. We do not find patterns in the different results. Some of the differences in the predictions originate from the limitation of the SPDc (see Section 3.3). The discrepancy in results is mainly due to the different definition of pocket in the two approaches. These results suggest that MSPocket and POCASA may be regarded as complementary methods for the detection of ligand binding pockets. When the predictions of both methods are considered, the success rates are greatly improved.

3.2 Evaluation results using OVLc

We plot the success rates of MSPocket based on different sensitivity and precision cutoffs for unbound structures in Figure 4 (see Supplementary Figure S2 for bound structures). The detailed sensitivity and precision values for MSPocket results are shown in Table 1. The area under the curve (AUC) values are also reported in the same plot. The performance of MSPocket using OVLc is given in Table 3. As a comparison, the performance of POCASA is also shown in Figure 4 and Table 3.

MSPocket successfully identifies 69 and 83% of the ligand binding sites on the unbound structures using MOc. In general, the sensitivities of MSPocket and POCASA are similar to each other. MSPocket performs slightly better than POCASA in identifying ligand binding pockets on bound structures as top 1 pockets/cavities (see AUC values in Supplementary Fig. S2a) and on unbound structures when top 3 pockets/cavities are considered (see AUC values in Fig. 4a). The precision of MSPocket is better than POCASA. The difference in the AUC values is observed to be in the range of 0.06–0.11 (Fig. 4b and Supplementary Fig. S2c).

We again observe that the results of MSPocket are different from those of POCASA. Here, we regard the results to be the same if

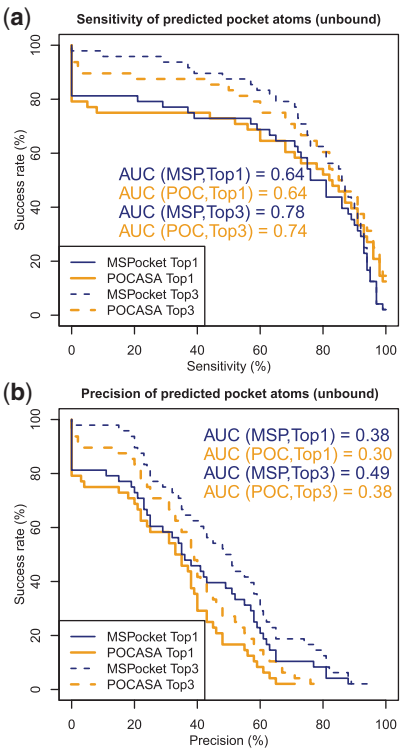

Fig. 4. Performance of MSPocket and POCASA. (**a**) Success rate vs. sensitivity for unbound structures. (**b**) Success rate vs. precision for unbound structures. Plots are obtained using R (R Development Core Team, 2010) and ROCR (Sing *et al.*, 2005). AUC, area under the curve; MSP, MSPocket; POC, POCASA.

Table 3. Success rates in the prediction of pockets for 48 bound/unbound structures using OVLc

Method	Top 1		Top 3	
	Unbound (%)	Bound (%)	Unbound (%)	Bound (%)
MSPocket	69	81	83	94
POCASA	69	77	81	94
Fpocket	69	83	94	92

the rank of the successful predictions are both 1, or both 2 or 3 or both larger than 3. For the 48 bound structures, MSPocket and POCASA results are the same for 41 (85%) structures. For the 48 unbound structures, the number is 31 (65%). If the top 1 predictions of both methods are combined, the accuracy is improved to 85% and 81% for bound and unbound structures, respectively. If the top 3 predictions are combined, the accuracy is 96% for both bound and unbound structures. See Supplementary Table S1 for details. Some of the difference comes from the cutoff values used in the OVLc. Nonetheless, the disagreement in the results is mainly due to the different definition of pocket in the two approaches. These results again illustrate that combining MSPocket and POCASA as

complementary methods allow the user to achieve higher success rate than using any of the methods separately.

3.3 SPDc versus OVLc

We have noticed that the success rate of MSPocket is different using SPDc and OVLc. Apart from the examples of 1ULB and 1QIF we introduced in Section 2.4, the predicted pockets/cavities are also assessed differently in a few other cases. In the case of 1ACJ, the ligand tacrine (THA) locates almost completely within the top 1 pocket (sensitivity = 98%). However, the top 1 pocket is large and THA locates at one corner of the pocket (precision = 29%) (see Fig. 3d). Similar results are also observed for 1CDO (see Fig. 3e). In these two examples, the top 1 pockets are not regarded as successful predictions according to SPDc because the distance between the MCs of the pockets and all atoms in the ligand exceeds 4.0 Å. But, in all the four cases, such spatial relationship between pockets and ligands is clearly reflected by higher sensitivity and low precision values. This demonstrates that OVLc provides a more comprehensive assessment of the predicted pockets than SPDc.

3.4 Pockets and protein orientation

MSPocket does not use any cubic grid representation of proteins. Its results are thus independent of protein orientations (given consistent generation of SES). On the contrary, the pockets detected by methods based on grid representations of proteins are not necessarily the same if proteins are transformed. Consequently, this may lead to inconsistent prediction results of pockets. For instance, in the unbound structure of α-momorcharin (PDB:1AHC), when the original coordinates of the protein in PDB are used, MSPocket identifies the ligand binding site at rank 1, while POCASA only detects the binding site at rank 2. Only when the unbound structure of α-momorcharin is superposed using PyMOL to its corresponding bound structure in the dataset (PDB:1MRG), is POCASA able to detect the binding site at rank 1. In other words, POCASA produces different results if the structure of α-momorcharin is transformed in 3D space. In another example, POCASA fails to identify ligand binding pockets in the top 5 pockets on the unbound structure of a neutral protease (PDB:1NPC) if the neutral protease is first aligned to its bound structure (PDB:1HYT). However, the ligand binding pocket is successfully identified by POCASA at rank 1 if the original coordinates of the neutral protease 1NPC is used. MSPocket identifies the ligand binding site of the neutral protease at rank 1 in both cases.

3.5 Implementation

MSPocket is implemented using Python. We also used Python packages BioPython (Cock *et al.*, 2009) and NetworkX (Hagberg *et al.*, 2008). The source code of MSPocket is publicly available at http://appserver.biotec.tu-dresden.de/MSPocket/. We have also created a MSPocket plugin for PyMOL (DeLano, 2002). Figure 5 shows a snapshot of the MSPocket PyMOL plugin. The runtime of MSPocket is given in the Supplementary Material.

4 CONCLUSIONS

Here we present MSPocket, a program for detecting pockets using geometric criteria on protein surfaces calculated by MSMS. MSPocket can be used for the prediction of potential ligand binding site on proteins. The comparative work presented here shows that in an evaluation using the same dataset, the sensitivity of MSPocket results is comparable to that of POCASA, which is one of the currently best performing methods and used as the reference in this work. The precision of the pockets predicted by MSPocket is higher than that of POCASA. Furthermore, we illustrated that the two methods successfully predict ligand binding pockets for different subsets of structures in the test dataset. The discrepancy in the results appears mainly because of the different definition of pocket used in the two approaches. The results obtained indicate that MSPocket and POCASA may be used as complementary programs for identifying ligand binding pockets. The combination of the results obtained on the prediction of pockets/cavities using each of these two methods may increase the probability of discovering biologically meaningful ligand binding sites. We foresee that a method combining the two algorithms could potentially be used to generate higher accuracy predictions.

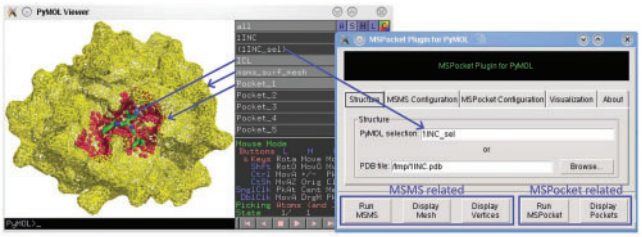

Fig. 5. MSPocket plugin for PyMOL. SES vertices and normal vectors can be rendered using MSMS plugin, which is a part of the MSPocket plugin. Here, the SES is shown as a yellow mesh. Pockets detected by MSPocket are shown as red spheres and ligand in green sticks. Users can configure the parameters for MSMS and MSPocket, as well as for the visualization of SES and pockets.

MSPocket does not use any cubic grid system to represent proteins, which makes the core algorithm of the MSPocket approach independent of protein orientations. MSPocket yields consistent results when proteins are transformed. This is a feature that grid-based methods like POCASA program lack. MSPocket is also implemented as a PyMOL plugin, which provides a flexible and interactive tool for the graphical investigation of protein surface pockets.

We also demonstrated that OVLc provides a more comprehensive assessment of pocket prediction results. Using sensitivity and precision measures, the spatial relationship between the predicted pockets and ligands is clearly reflected. This feature of OVLc is also important in the comparison of the geometry of pockets and ligands. For instance, Yu *et al.* (2010) reported that the shape of the predicted pockets agrees well with that of bound ligands. The comparison of the shapes was performed by visually inspecting the geometry of the pockets and ligands. Using OVLc, the assessment work may be carried out automatically, because a predicted pocket of both high sensitivity and precision must have a very similar shape to that of the ligand. This again illustrates that OVLc is a better strategy than SPDc for the assessment of predicted pockets.

ACKNOWLEDGEMENTS

We are grateful to Zhao Dong, Dr John Hawkins and Dr Joan Teyra for helpful discussions, and Dr Jian Yu for kindly providing the

POCASA program and dataset for the comparative work. We thank the reviewers for their valuable suggestions to improve the work.

Funding: This work has been partially funded by the German Research Council SFB-TRR 67 and the Klaus Tschira Stiftung gGmbH.

Conflict of Interest: none declared.

REFERENCES

Brady,G.P. and Stouten,P.F. (2000) Fast prediction and visualization of protein binding pockets with pass. *J. Comput. Aided Mol. Des.*, **14**, 383–401.

Campbell,S.J. *et al.* (2003) Ligand binding: functional site location, similarity and docking. *Curr. Opin. Struct. Biol.*, **13**, 389–395.

Chen,B.Y. and Honig,B. (2010) VASP: a volumetric analysis of surface properties yields insights into protein-ligand binding specificity. *PLoS Comput. Biol*., **6**, e1000881.

Cock,P.J.A. *et al.* (2009) Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.

DeLano,W.L. (2002) The PyMOL Molecular Graphics System. DeLano Scientific LLC.

Hagberg,A.A. *et al.* (2008) Exploring network structure, dynamics, and function using NetworkX. In Varoquaux,G. *et al.* (eds), *Proceedings of the 7th Python in Science Conference (SciPy2008).* Pasadena, CA USA, pp. 11–15.

Hendlich,M. *et al.* (1997) LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol. Graph. Model*, **15**, 359–363.

Huang,B. and Schroeder,M. (2006) Ligsite^{csc}: predicting ligand binding sites using the connolly surface and degree of conservation. *BMC Struct. Biol.*, **6**, 1–11.

Kellogg,G.E. *et al.* (2005) New application design for a 3d hydropathic map-based search for potential water molecules bridging between protein and ligand. *Internet Electron. J. Mol. Des.*, **4**, 194–209.

Laskowski,R.A. (1995) SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graph.*, **13**, 323–330.

Le Guilloux,V. *et al.* (2009) Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics*, **10**, 1–11.

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

Liang,J. *et al.* (1998) Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci.*, **7**, 1884–1897.

London,N. *et al.* (2010) The structural basis of peptide-protein binding strategies. *Structure*, **18**, 188–199.

Nayal,M. and Honig,B. (2006) On the nature of cavities on protein surfaces: application to the identification of drug-binding sites. *Proteins*, **63**, 892–906.

Petsalaki,E. *et al.* (2009) Accurate prediction of peptide binding sites on protein surfaces. *PLoS Comput. Biol.*, **5**, e1000335.

R Development Core Team (2010) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sanner,M.F. *et al.* (1996) Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers*, **38**, 305–320.

Sing,T. *et al.* (2005) ROCR: visualizing classifier performance in r. *Bioinformatics*, **21**, 3940–3941.

Sotriffer,C. and Klebe,G. (2002) Identification and mapping of small-molecule binding sites in proteins: computational tools for structure-based drug design. *Farmaco*, **57**, 243–251.

Tripathi,A. and Kellogg,G.E. (2010) A novel and efficient tool for locating and characterizing protein cavities and binding sites. *Proteins*, **78**, 825–842.

Weisel,M. *et al.* (2007) Pocketpicker: analysis of ligand binding-sites with shape descriptors. *Chem. Cent. J.*, **1**, 1–17.

Yu,J. *et al.* (2010) Roll: a new algorithm for the detection of protein pockets and cavities with a rolling probe sphere. *Bioinformatics*, **26**, 46–52.

Zhu,H. *et al.* (2008) Alignment of non-covalent interactions at protein-protein interfaces. *PLoS One*, **3**, e1926.