Structural Bioinformatics

Roll: A new algorithm for the detection of protein pockets and cavities with a rolling probe sphere

Jian Yu[1], Yong Zhou[1], Isao Tanaka[1] and Min Yao[1,*]

[1]Graduate School of Life Science, Hokkaido University, Kita-Ku Kita-10 Nishi-8, Sapporo, 0600810, Japan.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Prediction of ligand binding sites of proteins is significant as it can provide insight into biological functions and reaction mechanisms of proteins. It is also a prerequisite for protein – ligand docking and an important step in structure-based drug design.

Results: We present a new algorithm, Roll, implemented in a program named POCASA, which can predict binding sites by detecting pockets and cavities of proteins with a rolling sphere. To evaluate the performance of POCASA, a test with the same data set as used in several existing methods was carried out. POCASA achieved a high success rate of 77%. In addition, the test results indicated that POCASA can predict good shapes of ligand binding sites.

Availability: A web version of POCASA is freely available at http://altair.sci.hokudai.ac.jp/g6/Research/POCASA_e.html

Contact: yao@castor.sci.hokudai.ac.jp

1 INTRODUCTION

Proteins perform their biological functions by their interactions with other molecules, such as substrates, coenzymes, antigens, nucleic acids, other proteins, *etc*. In general, these interactions occur in concave regions on the protein surface, which are usually called pockets (concavities on the protein surface) or cavities (enveloped by the protein surface). Therefore, it is of interest to identify pockets and cavities in protein structures. Especially, the detection of pockets and cavities is a prerequisite for protein – ligand docking and an important step in structure-based drug design.

In recent years, many computational search methods have been developed. One type of computational search method involves prediction of pockets and cavities by calculating the interaction energy between protein and a probe group. GRID (Goodford, 1985) is a well-known energy calculation method, which calculates interaction energy with a set of empirical energy functions including Lennard-Jones function, electrostatic function and hydrogen bond function, using water, methyl group, or amine nitrogen as the probe group. Further studies of hydrogen bond energy calculation between protein and ligand probe groups were performed (Boobbyer *et al*., 1989; Wade *et al*., 1993) based on GRID. An *et al*. (2004) developed an algorithm called DrugSite based on the transformation of van der Waals potential energy. Another method named Q-SiteFinder (Laurie *et al*., 2005) locates the ligand binding sites by calculating the non-bonded interaction energy. In addition, there is a method (Soga *et al*., 2007) that predicts the binding sites using the ratio of occurrence of each standard amino acid at the binding sites of a training data collection.

In contrast to the search methods mentioned above, one of the most commonly used search methods is the purely geometric search method, which does not require any other non-geometric knowledge but only the protein 3D structure. Based on the algorithm characteristics, the geometric methods can be divided into three primary types: 1) grid system scanning; 2) probe sphere filling; 3) alpha-shape.

In the grid system scanning method, a 3D grid system will first be filled with protein atoms, which are thought of as spheres of van der Waals radii. All grid points are sorted into two types—those that are and are not occupied by protein—and then scanned along several fixed directions. The free grids that are not occupied by protein atoms will be picked out as pocket (or cavity) grids if they satisfy some geometric conditions. Among grid system methods, one early method called POCKET (Levitt *et al*., 1992) scans grid points along the *x*, *y* and *z* axes. Based on POCKET's algorithm, LIGSITE (Hendlich *et al*., 1997) added another four cubic diagonal directions to scan the 3D grid system. Two extensions of LIGSITE, LIGSITE^{cs} using the protein's ***C***onnolly ***S***urface (Connolly, 1983) and LIGSITE^{csc} (Huang *et al*., 2006) ranking pockets with the degree of conservation obtained from the ConSurf-HSSP database (Glaser *et al*., 2005), were developed, but LIGSITE^{csc} is no longer a purely geometric algorithm. Moreover, one function of the program LigandFit (Venkatachalam *et al*., 2003) can also identify pockets and cavities by applying a cubically shaped eraser. For the grid system method, the algorithm is relatively straightforward and the precision is freely adjustable by changing the unit grid size. However, it is sensitive to the scanning direction and the search results are strongly dependent on the orientation of protein structure in the grid system. Therefore,

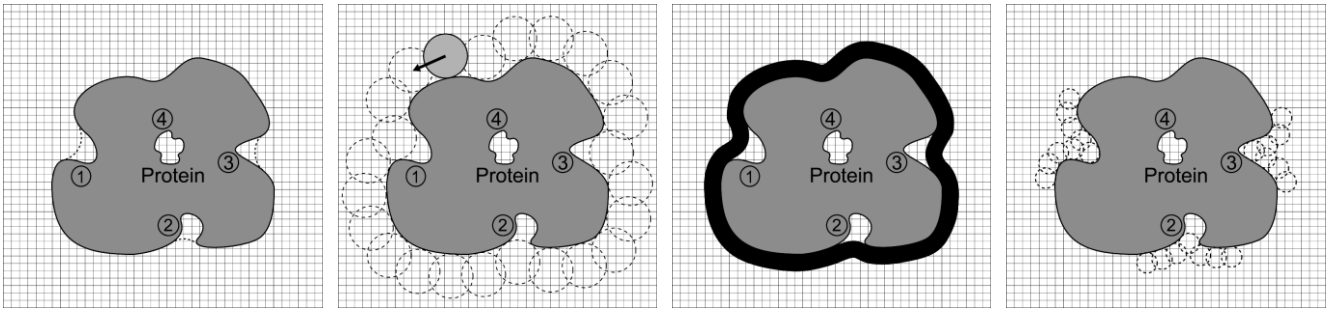
*To whom correspondence should be addressed.

LIGSITE adopts another four scanning directions compared with POCKET to weaken the influence of this orientation problem. Further, another grid system method named PocketPicker (Weisel *et al*., 2007) scans the protein molecular surroundings along 30 particular directions obtained by octahedron triangulation of a sphere. Increasing scanning directions indeed leads to better search results [PocketPicker showed the highest success rate (Weisel *et al*., 2007)] but was more time consuming.

SURFNET is a well-known program that uses the probe sphere filling method (Laskowski, 1995), which uses a set of probe spheres called "gap spheres" each of which is placed midway between a pair of atoms in a protein molecule. If a gap sphere overlaps any other neighbouring atoms, its radius will be reduced until no overlap occurs. The final gap spheres define the pockets and cavities. Another sophisticated method, PASS (Brady *et al*., 2000), generates a set of probe spheres tangential to all unique triplets of atoms in a protein molecule to coat the protein layer by layer until all pockets and cavities are filled. Then, only the probes with low solvent exposure are retained. Finally, ASPs (active site points) are used to represent potential binding sites. A program named PHECOM (Kawabata *et al*, 2007) generates probe spheres with a radius of 1.87 Å (the same size as -CH_3) in a similar way to PASS. Then, a larger sphere (4 – 12 Å) is used to shave off the probe spheres lying outside pockets. As mentioned above, most probe sphere-based methods directly fill pockets or cavities with a set of specific probe spheres. Therefore, there is no orientation problem similar to that which occurs in grid system methods. However, the algorithms are relatively complicated and sometimes the shapes of pockets and cavities cannot be described properly as regular shaped spheres are used.

CAST (Liang *et al*., 1998) is a search method based on alpha-shape and discrete-flow theory (Edelsbrunner *et al*., 1994, 1995, 1996). First, proteins are triangulated with a series of Delaunay tetrahedra. Then, the alpha shape of the target protein is derived by removing the empty Delaunay tetrahedra, which have some parts outside the protein. Finally, pockets and cavities can be identified according to the alpha shape and discrete-flow method. Another method based on alpha-shape theory is MolAxis (Yaffe,E *et al*., 2008) which can locate cavities inside protein molecules and then find channels emanating from cavities.

Here, we present a new purely geometric search algorithm called ***Roll***. This method uses both the grid system and probe sphere. First, a 3D grid system is filled with atoms in the protein molecule. Second, a probe sphere is adopted to roll along the protein surface to generate a "*probe surface*" based on the inner border tracing algorithm in the image processing field. Then, the regions between the protein and probe surface or those surrounded by the protein surface are defined as pockets and cavities, respectively. To remove noise points, two parameters were designed: Single-Point Flag and Protein-Depth Flag. Moreover, *Roll* can predict pockets of different shape and volume by adjusting the radius of the probe sphere. Finally, for ranking pockets, Volume-Depth was designed to evaluate the possibility that predicted pockets are actual binding sites. This algorithm is implemented in a program named ***POCASA*** (***PO***cket-***CA***vity ***S***earch ***A***pplication). The results of testing and comparison with other methods are discussed in the following sections.



(a) (b) (c) (d)

Fig. 1. Schematic illustration of Roll: (a) Slice of the grid system. The dashed and solid lines show the probe surface and the protein surface, respectively. The grey region is the protein. Regions 1 – 3 are defined as pockets and region 4 is defined as a cavity. (b) The rolling process. The light grey ball indicates the starting position and the dashed balls show the trace of rolling. (c) The black area is the probe surface. (d) The small probe causes pocket 1 to disappear and pockets 2 – 3 to become smaller, while it did not affect cavity 4.

2 METHODS

2.1 Picking out pockets and cavities

The key concept of Roll is to generate a crust-like surface called the *probe surface* enveloping protein to identify the region between the probe surface and protein surface as a "pocket" and the region surrounded by protein surface as a "cavity" (Fig. 1a). Now, the goal is to generate this "probe surface". The first step is to fill the 3D grid system with protein atoms, which are regarded as 3D spheres of different radii. Here, van der Waals radii are adopted: 1.5Å for oxygen atoms, 1.6Å for nitrogen atoms, 1.7Å for carbon atoms and 1.8Å for sulphur and phosphorus atoms. Hydrogen atoms are not taken into account. The grid points occupied by protein atoms are labelled as protein points with a value of 1 and free grid points have a value of 0. Next, Roll uses a rolling probe sphere to generate the "probe surface". In this step, the 3D grid system is divided into a set of slices with the same unit grid thickness. For each slice, the probe starts from the origin point to linearly scan the grid points. Once the probe encounters the first protein point, it begins to roll along the protein surface

without any overlap with protein until returns to the starting position (Fig. 1b). The rolling direction is controlled based on the inner border tracing algorithm (Sonka *et al*., 1998) described in detail in the APPENDIX. The free grids through which the probe rolls are labelled as the "probe surface". The regions consisting of free points between the probe and protein surface or those surrounded by protein surface are then identified as pockets and cavities, respectively (Fig. 1c).

As different probe radii lead to different rolling loci, *i.e.*, probe surfaces, Roll can determine the pockets differing in volume and shape. If the radius of the probe sphere is sufficiently small, the probe can roll into pockets and thus diminish or delete pockets. For example, pocket 1 will disappear and correspondingly pockets 2 and 3 will become smaller with a smaller probe, as shown in Fig. 1 d. In contrast, if larger or shallow pockets exist in the protein, a probe sphere of larger radius is preferable. By adjusting the probe radius, Roll can predict various binding sites and this function is significant for the cases in which the users wish to find specific pockets.

2.2 Removing Noise Points

The grid system has a disadvantage regarding generation of noise points concurrently with pocket (cavity) points. Figure 2 b shows a preliminary search result of protein MinD binding with ADP (PDB 1ION) (Fig. 2a).

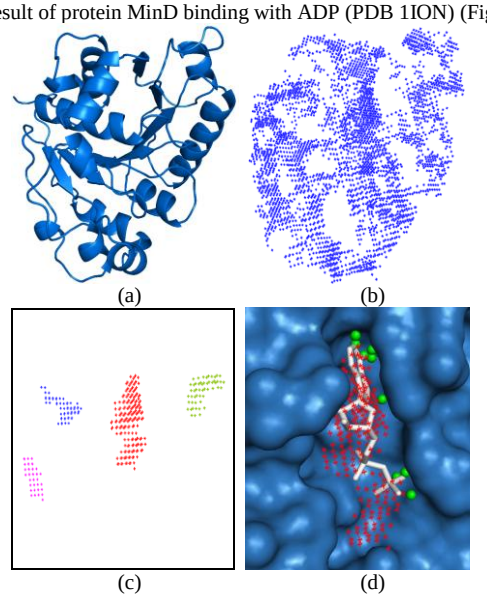


Fig. 2. Use of SPF and PDF to remove noise points. (a) Ribbon model of PDB 1ION. (b) Preliminary search result of 1ION. (c) The top four largest pockets obtained using SPF = 16; the red pocket is the largest. (d) The predicted pocket for the ADP binding site of MinD. The red point pocket was obtained with SPF = 16 and the green ball points were recovered using PDF = 18. (All images generated with PyMOL.)

2.2.1 Single Point Flag (SPF)

To remove noise points, a new parameter, *Single Point Flag* (SPF), was developed to evaluate whether a candidate pocket (cavity) point is an isolated point. SPF_j is the number of pocket points adjacent to the *jth* pocket point. If a neighbouring point is a pocket point, its contribution to SPF_j is 1; otherwise, the contribution is 0. Then, SPF_j can be counted as follows:

$$SPF_j = \sum_{k=1}^{27} f(n_k), \quad f(n_k) = \begin{cases} 1, n_k \in \{ pocket\ point \} \\ 0, n_k \in \{ non-pocket\ point \} \end{cases} \qquad (a)$$

where n_k is the *kth* neighbouring point of the *jth* pocket point and f_k is the contribution of n_k. The minimum SPF of one pocket point is 1, which means it does not have any neighbouring pocket points. The maximum is 27 as in the 3D grid system one pocket point has at most 26 neighbouring pocket points. One pocket point will be regarded as a pocket or cavity point if its SPF value is higher than the SPF threshold, which can be assigned by the user or automatically adjusted by Roll. The default SPF threshold in Roll is set as 16, namely 60% of the maximum SPF value, which means if 60% of its surrounding region is occupied by pockets points, this point belongs to a certain pocket. Figure 2c shows the top four largest pockets of protein MinD after using SPF. Obviously, the SPF results were much clearer than the preliminary results.

2.2.2 Protein-Depth Flag (PDF)

While SPF is useful for deleting noise points, some pocket points on the edges of pockets may also be removed. The edge points between pockets and the protein surface are more meaningful than the edge pocket points near the solvent region. To recover these useful points, we developed another parameter, *Protein-Depth Flag (PDF)*, the definition of which is similar to that of SPF. PDF_j is the number of protein points adjacent to the *jth* pocket point. If PDF_j is larger than the PDF threshold, the *jth* pocket point will not be removed even if SPF_j is below the SPF threshold. Figure 2 (d) shows an example of using PDF. The red pocket is the largest among the predicted pockets in MinD and the green points are the recovered points obtained using PDF. The SPF&PDF pocket (red and green) covered all the atoms of ADP, while SPF pocket (red) only partially covered ADP.

2.3 Ranking

Usually, more than one pocket (cavity) will be found for a target protein. Therefore, the step after the search process is to evaluate which pocket (cavity) has a higher likelihood of being the actual binding site. One method is to rank pockets by volume. Based on this criterion, pockets with larger volume are considered to have higher likelihood of being the binding site. However, ranking pockets only by volume is sometimes inappropriate. For example, as shown in Fig. 3, pocket B has a higher likelihood of being the binding site as it is deeper than pocket A, although they have similar volume. Therefore, a new parameter *Volume Depth (VD)* was designed in Roll for more accurate ranking of pockets.

In VD calculation, the *depth* of every pocket point, which is defined as the shortest distance from pocket point to probe surface (Fig. 3b), is first calculated. Then, the VD value of the *ith* pocket will be determined by summing the depth of all pocket points by the following expression:

$$VD_i = \sum_{j=1}^{N} d(g_j), \quad g_j \in P_i \qquad (b)$$

where $d(g_j)$ is the depth of pocket point g_j and N is the number of pocket points in pocket P_i. All predicted pockets will be ranked in terms of their VD values. By definition, VD is the criterion not only of the position but also of the volume. VD is used only for pocket ranking, but not for cavity ranks. If some pockets and cavities are found simultaneously from one protein, they will be ranked separately with VD and volume used for ranking pockets and cavities, respectively.

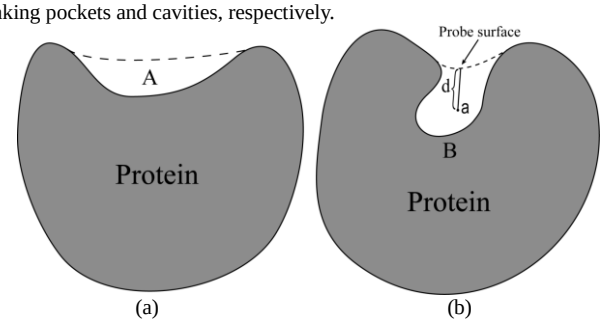


Fig. 3. Volume Depth. (a) Flat pocket A. (b) Deep pocket B with similar volume to pocket A. *d* is the depth of pocket point a. The dashed line is the probe surface.

3 RESULTS

3.1 Evaluation and Comparison

We developed a program to implement Roll named POCASA. To evaluate the binding site prediction performance of POCASA, we adopted an evaluation method similar to that applied in previous studies (Huang *et al*., 2006; Weisel *et al*., 2007). A prediction was regarded as successful if a single point which represents the predicted pocket (cavity) was within 4Å of any atom of the ligand; otherwise, the prediction was regarded as a failure. In this study, a new concept, the depth centre (DC), was defined to represent the predicted pocket (cavity) as follows:

$$DC(P_i) = \frac{\sum_{j=1}^{N} d(g_j) \times c(g_j)}{\sum_{j=1}^{N} d(g_j)}, \quad g_j \in P_i \qquad (c)$$

where $DC(P_i)$ represents the depth centre of pocket P_i. $d(g_i)$ is the depth of pocket point g_j and $c(g_i)$ is the coordinates of it.

The same data set containing 48 bound/unbound structures downloaded from Protein Data Bank (PDB) was used for the test and comparison with the other search methods, and the success rates are shown in Table 1. POCASA showed slightly better performance than the other prediction methods in both Top 1 and Top 3 success rates for bound/unbound structures. Table 2 shows the best search results of each PDB from POCASA. In POCASA, only top 5 pockets (cavities) are considered as output results.

As shown in Table 1, POCASA achieved a high success rate, and was also successful in some difficult cases, for example, 3MTH/6INS, methylparaben insulin. The binding site of methylparaben is too shallow for successful prediction within the top 3 pockets in all the search methods discussed by Huang *et al*., 2006. However, POCASA predicted this binding site as a rank 2 and rank 1 pocket for bound/unbound insulin structure, respectively (Table 2).

Table 1. Success rates of predicting binding sites for 48 bound/unbound structures

Method	Top 1		Top 3	
	Unbound	Bound	Unbound	Bound
POCASA	75%	77%	88%	94%
PocketPicker	69%	72%	85%	85%
LIGSITE^{cs}	60%	69%	77%	87%
LIGSITE	58%	69%	75%	87%
CAST	58%	67%	75%	83%
PASS	60%	63%	71%	81%
SURFNET	52%	54%	75%	78%

As mentioned above, one problem of using grid system is that the search results are dependent on the orientation of molecule in the 3D grid. In Roll, the process of rolling the 3D grid is performed by cutting the 3D grid into 2D slices and then rolling these slices. The orientation problem won't happen in a 2D slice since the inner border tracing algorithm is not affected by the orientation. For the direction of 2D slices in the 3D grid, we tested POCASA by merging the rolling directions as follows: (*x*), (*x*)+(y) and (*x*)+(*y*)+(*z*). While the success rate of rolling (*x*)+(y) greatly increased 15% and 9% for Top 1 and 3 respectively, compared with that of (*x*), the success rate only increased 2% and 4% from (*x*)+(y) to (*x*)+(*y*)+(*z*) (supplementary Fig D). Such result indicated that merging more than three directions hardly affected success rate and the rolling process can reduce the effect of the orientation dependency. Therefore, merging three 3D rolling directions [(*x*)+(*y*)+(*z*)] was adopted in POCASA under a consideration of calculation-complexity and time-consumption.

In POCASA, there are four adjustable parameters: grid size, probe radius, SPF and PDF threshold. For grid size, we tested POCASA with another grid interval of 0.5Å. The success rate for the 48 bound structures is the same as the top 1 (77%) and top 3 (94%) success rate of 1Å. Smaller gird size hardly affects the performance of prediction in POCASA, although it may be helpful to generate pockets (cavities) of good shape. Furthermore, the time consumption correspondingly increased when gird size becomes smaller. Therefore, the unit gird size of 1Å is adopted as the default value in POCASA, and 0.5Å for gird size can be also set by users. The radius of probe sphere is an important parameter for prediction in POCASA. Since the ligand binding sites are of various shape and size, it is difficult to find an all-powerful value. By testing dozens of protein structures (data not shown), we found that 2Å is competent for most cases. The adjustment of the probe radius will be discussed later (see 3.3).

Another parameter which can influence the results is SPF. As mentioned in Section 2.2.1, SPF value of a point is calculated by counting its neighbouring pocket points within a cube of a 2Å side. SPF value of a point ranges from 1 to 27 when 1Å is used for grid size. The default value of SPF threshold is 16 which is determined by testing dozens of protein structures. The last parameter, PDF, is designed to recover the points on the edges of pockets so it does not affect the search results so much and an empirical value for PDF threshold is around 18. For most cases, 16 for $SPF_{threshold}$ and 18 for $PDF_{threshold}$ is adequate to predictions. If there are still a few noise points remained, the users can slightly increase the threshold of SPF and PDF to obtain better shapes for pockets.

3.2 Shape prediction of ligand binding site

As the entire region of pockets can be enveloped by the probe surface and SPF & PDF is effective for removing noise points, POCASA can obtain good shapes for pockets (cavities) coinciding well with the bound ligand. We next examined this ability using flavin-adenine dinucleotide (FAD), for which abundant data are available regarding protein – ligand complexes (494 complexes containing FAD in Protein Data Bank)(Shin *et al*., 2005). Fifty-four FAD bound/unbound complexes, of which fifty-two complexes were randomly chosen, were used for the analysis (Table 3). We divided these complexes into three groups according to FAD binding conformation: flat shape, partial interaction, and U-shape. In the third group, the U-shape group, FAD bound to the protein in a U-shaped conformation so only a block-shaped pocket could be detected by POCASA. In this case, the shape of predicted pockets can not provide so much meaningful information. In the other two groups, POCASA provided meaningful information about the shape of the FAD binding site as follows:

1) *Flat shape*. FAD stretched itself in a linear shape. In this group, most atoms of FAD could be enveloped by the probe surface and so POCASA could generate pockets that coincided well with FAD. Figure 4 shows the predicted pockets of FAD bound/unbound

Table 2 Search results for 48 bound/unbound structures

Ligand ID	bound	Rank	Distance (Å)	Total	unbound	Rank	Distance (Å)	Total
UMP	1BID	1	2.66	3	3TMS	1	3.07	3
NAD	1CDO	1	2.88	5	8ADH	1	0.80	5
MID	1DWD	2	1.78	5	1HXF	1	3.51	5
F6P	1FBP	1	3.36	5	2FBP	1	2.68	5
GAL	1GCA	1	0.93	5	1GCG	1	0.75	5
NAG	1HEW	1	1.91	2	1HEL	1	1.09	3
BZS	1HYT	1	1.24	5	1NPC	–	–	5
ICL	1INC	3	0.84	5	1ESA	3	0.92	5
RTL	1RBP	1	0.73	5	1BRQ	1	0.93	4
C2P	1ROB	1	1.08	3	8RAT	1	1.27	3
BTN	1STP	1	0.31	4	1SWB	1	0.92	4
GUN	1ULB	1	3.43	5	1ULA	1	2.91	5
PLM	2IFB	1	2.37	2	1IFB	1	2.12	4
BEN	3PTB	2	0.28	5	3PTN	2	0.70	5
PGA	2YPI	2	0.84	5	1YPI	1	2.39	5
MTX	4DFR	1	2.61	2	5DFR	1	1.51	5
VAC	4PHV	1	1.72	1	3PHV	1	1.55	1
MMA	5CNA	–	–	5	2CTV	5	0.89	5
FVF	7CPA	1	0.86	4	5CPA	1	3.52	5
NIP	1A6W	3	1.65	5	1A6U	4	0.67	5
THA	1ACJ	1	1.73	5	1QIF	1	3.03	5
STA	1APU	1	1.59	3	3APP	1	3.18	5
FOS	1BLH	1	0.67	5	1DJB	3	0.72	5
GLC	1BYB	1	1.36	5	1BYA	1	3.21	5
PLH	1HFC	1	1.31	5	1CGE	2	1.16	4
PPL	1IDA	1	2.72	3	1HSI	1	3.04	2
DGX	1IGJ	3	1.01	5	1A4J	5	0.94	5
LIP	1IMB	1	1.37	3	1IME	1	2.02	5
ST1	1IVD	1	0.82	5	1NNA	1	1.31	5
ADN	1MRG	3	0.69	5	1AHC	1	1.76	5
DX9	1MTW	1	3.03	5	2TGA	–	–	5
SAB	1OKM	1	0.93	5	4CA2	1	1.72	5
PGA	1PDZ	1	3.76	5	1PDY	1	2.44	5
PIM	1PHD	1	1.46	5	1PHC	1	1.56	5
STA	1PSO	1	0.76	5	1PSN	1	1.64	5
PP2	1QPE	1	0.66	5	3LCK	1	0.75	5
C60	1RNE	1	0.94	3	1BBS	1	0.67	5
PTP	1SNC	1	3.19	4	1STN	1	1.9	5
MTB	1SRF	1	0.74	2	1PTS	1	0.87	2
LOF	2CTC	1	1.00	5	2CTB	1	1.51	5
AZM	2H4N	1	0.95	5	2CBA	1	1.04	5
ACA	2PK4	1	1.48	2	1KRN	2	0.72	3
DAN	2SIM	–	–	5	2SIL	3	1.58	3
LEP	2TMN	1	1.19	5	1L3F	1	0.71	5
CIN	3GCH	(1)	4.23	5	1CHG	5	2.42	5
MPB	3MTH	2	0.70	3	6INS	1	3.42	3
DHG	5P2P	1	1.46	4	3P2P	1	0.53	2
UVC	6RSA	1	1.61	5	7RAT	1	1.02	5

*Default Parameters in POCASA of Grid Size = 1Å, Probe Radius = 2Å, SPF = 16, PDF = 18 were used for calculation. Distance column represents the distance from pocket (cavity) depth centre to the nearest ligand atom. Total column means the number of predicted pockets and cavities for each protein structure. The number in round bracket means the binding site is predicted in the pocket (cavity) whose geometric centre is beyond 4 Å range of the nearest ligand atom.

Table 3. FAD-bound complexes obtained from PDB

FAD binding status	PDB ID
Flat shape	1an9, 1b4v, 1b8s, 1b37, 1bwc, 1bzl, 1c0i, 1cbo, 1coy, 1e8f*, 1e8g, 1grt, 1gfx, 1h81, 1ksu, 1l9e, 1lj1, 1nek, 1ng3, 1pbd, 1rsg, 1s2q, 1tdk, 1typ, 1ve9, 1zov, 1zp0, 2b7r, 2b9w, 2du8, 2fja, 2q6u, 2r4e
Partial interaction (flavin)	1a8p, 1bx1, 1b2r, 1bjk, 1bqe, 1bxo, 1fnb, 1frn, 1oqc, 1quf, 1r2j, 1rp4, 1siq, 1wgb*, 1yoa, 2b3d, 2b5o
U-shape	1foh, 1np7, 1tj0, 1u3c

*unbound structures

structures. The shapes of the four pockets matched flat shaped FAD very well, indicating that POCASA can generate good shapes for ligand binding sites. This capability may provide insight for ligand identification and drug prediction requiring only the 3D protein structure.

2) *Partial interaction.* Only the flavin of FAD bound to the protein, while the other parts mostly extended to the solvent area so it was impossible to envelope the whole FAD with the probe surface. However, POCASA could still determine the flavin part of FAD. In this group, the flavin parts of all 12 FADs participated in the

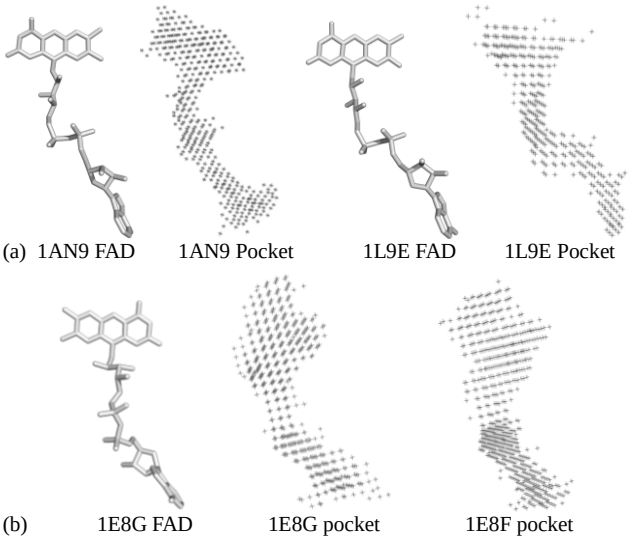


Fig. 4. Pockets of flat-shaped FAD. (a) Two bound conformations of FAD and predicted pockets in 1AN9 and 1L9E. (b) Predicted pockets in 1E8G (bound) and 1E8F (unbound). (R=2Å, $SPF_{threshold}$ = 16, $PDF_{threshold}$ = 18)

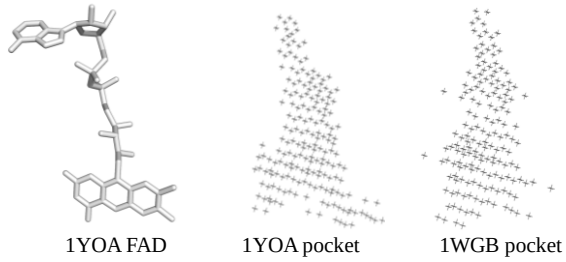


Fig. 5. Predicted binding sites of 1YOA and 1WGB, which are represented by the point regions. (R = 3Å, $SPF_{threshold}$ = 16, $PDF_{threshold}$ = 18)

interaction with proteins, which could be observed from the pockets predicted by POCASA (Fig. 5). Compared with the flat shape, POCASA calculated the rough shapes for this partial interaction but the flavin parts were still completely covered by the predicted pockets (Fig. 5).

3.3 Different Probe Radii

One function of POCASA is to use different probe radii to generate pockets of different geometric features, such as volume and shape. Here, two examples are now presented to show how to change probe radii for more reasonable results. The first example is glycosyltransferase GtfA (PDB 1PNV), which binds with the very large ligand vancomycin (VAN, $C_{66}H_{77}Cl_2N_9O_{24}$, molecular weight = 1450.62) in a large flat pocket (Fig 6). A small probe sphere can easily enter such a large pocket. While the binding site was divided into three parts using a probe sphere of radius 2 Å (Fig. 6a), the predicted pocket detected with the probe sphere of radius 8 Å described the binding site quite well and almost covered the whole ligand (Fig. 6b).

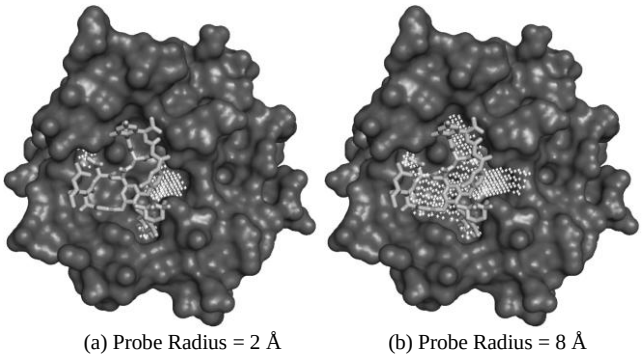


Fig. 6. Search result of 1PNV with different probe radii. (a) The binding domain of 1PNV with vancomycin. The predicted pocket obtained with Probe Radius = 2 Å. (b) The predicted pocket obtained with Probe Radius = 8 Å. ($SPF_{threshold}$ = 16 and $PDF_{threshold}$ = 18)

The probe radius was also increased to search for pockets in 1ION. As the first phosphate was positioned in a flat region of the protein surface, the predicted pocket was divided into two parts by the probe of radius 2 Å (Fig. 7 a). However, when a probe of radius 3Å was adopted, the two separate parts were connected together and formed one large pocket (Fig. 7 b), which is more appropriate for ADP. Moreover, as the shape of this predicted pocket indicated, ATP may also be bound in this binding site. The two examples show that it is sometimes better to change the probe radius, although a probe of radius 2Å is adequate in most cases. In principle, large probe spheres are used for large ligands or large flat pockets, while small probe spheres for small ligands.

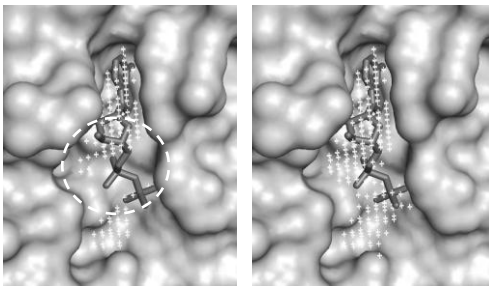


Fig. 7. Search results of the ADP binding site in 1ION. (a) Results with a probe of radius 2 Å. The pocket was disconnected at the first phosphate, as shown in the region marked by the circle. (b) The whole pocket obtained with a probe of radius 3 Å. ($SPF_{threshold}$ = 14 and $PDF_{threshold}$ = 18)

3.4 Ranking with Volume Depth

For binding site prediction, one crucial step is to rank the predicted pockets. Several measures were used for this purpose, such as volume (Weisel *et al*., 2007), degree of conservation (Huang *et al*., 2006), the distance from molecular centroid (Yaffe,E *et al.*, 2008) and so on. In POCASA, a new concept, volume depth was presented to directly and quantitatively describe the position and volume together of a pocket. Here, we discuss PDB 1GCA containing the ligand β-D-galactose (GAL) as an example to show how VD functions. Figure 8 shows the top two largest predicted

pockets of 1GCA. The largest pocket of 82 pocket points had VD of 223, and the second largest pocket of 58 pocket points had VD of 589. If we rank pockets only by volume, the largest pocket will be considered the rank 1 pocket even if the small pocket is buried very deep within the protein. However, if the position information is taken into account using the VD value, the small pocket with VD value of 589 should be taken as the binding site rather than the largest pocket. In fact, ligand GAL was bound with 1GCA at the small pocket.

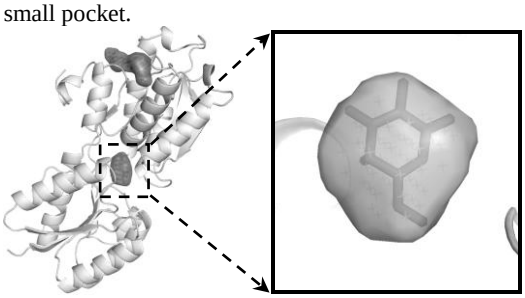


Fig. 8. The top two largest pockets of 1GCA. The pocket at top left is the largest pocket. The central pocket is the second largest pocket; however, it has the largest VD value among all predicted pockets of 1GCA. The right figure is a higher magnification view of the binding pocket surface and ligand GAL. ($SPF_{threshold}$ = 16 and $PDF_{threshold}$ = 22)

By ranking with VD values, among the 48 complexes shown in Table 2, the ranks of the binding pockets in 1ACJ, 1GCA and 1HFC were changed from rank 2 to rank 1; and the ranks of the binding pocket in 1MRG and 2YPI were changed from rank 4 to rank 2. The success rates of Top 1 and Top 3 were greatly improved from 71% (obtained by ranking with volume) to 77% and from 90% to 94%. There were no contrary cases in which the rank orders of binding pockets among the 48 complexes decreased by VD. Therefore, we concluded that VD is a good descriptor for estimating the possibility that pockets may be binding sites.

4 CONCLUSIONS

We presented a new algorithm called Roll to identify pockets and cavities, and developed a program called POCASA that can automatically predict the ligand binding sites of proteins. POCASA achieved a higher success rate of 77%/75% for the 48 bound/unbound structures than previous search methods, which substantiates that POCASA can provide believable results for ligand binding site prediction and analysis. The test results indicated that POCASA can generate the shapes of pockets and cavities that coincide well with bound ligand. Meanwhile, the novel function in that the pockets could be determined by different probe spheres makes POCASA versatile for various ligands and proteins. In addition, volume depth was confirmed to be a very useful factor for POCASA to correctly rank predicted pockets.

ACKNOWLEDGEMENTS

We would like to thank Mr. Yamashita for his help with the web interface of POCASA.

Funding: Hokkaido University President's Fellowship.

REFERENCES

An,J.H. *et al.* (2004) Comprehensive Identification of "Druggable" Protein Ligand Binding Sites. *Genome Informatics*, **15(2)**, 31–41.

Ben-Shimon ,A and Eisenstein, M. (2005) Looking at enzymes from the inside out: the proximity of catalytic residues to the molecular centroid can be used for detection of active sites and enzyme-ligand interface. *J.Mol.Biol.*, **351**,309-326.

Boobbyer,D.N.A. et al. (1989) New hydrogen-bond potentials for use in determining energetically favorable binding sites on molecules of known structure. J. Med. Chem., 32 (5), 1083-1094.

Brady,G.P and Stouten,P.F.W. et al. (2000) Fast prediction and visualization of protein binding pockets with PASS. J.Comput.Aid.Mol.Des., 14, 383-401.

Connolly,M.L. (1983) Analytical molecular surface calculation. *J.Appl.Cryst*., **16**, 548-558.

Delano,W. (2009) PyMOL Molecular Graphics System. http://www.pymol.org/.

Edelsbrunner.H and Mucke,E.P. (1994) Three-dimensional alpha shapes. *ACM T Graphic*, **13**, 43-72.

Edelsbrunner,H. (1995) The union of balls and its dual shape. *Discrete Comput Geom*, **3**, 415-440.

Edelsbrunner,H. *et al*. (1995) Measuring proteins and voids in proteins. *In: Proc. 28th annual Hawaii international conference system sciences*. Los Alamitos, California: IEEE Computer Society Press, 256-264.

Edelsbrunner,H. *et al*. (1996) On the definition and the construction of pockets in macromolecules. In: Hunter L, Klein T, eds. *Biocomputing: Proceedings of the 1996 Pacific symposium.* Singapore: World Scientific Publishing, 272-281

Edelsbrunner,H. and Shah,N.R. 1996. Incremental topological flipping works for regular triangulations. *Algorithmica*, **15**, 223-241.

Glaser.F. *et al*. (2005) The ConSurf-HSSP database: the mapping of evolutionary conservation among homologs onto PDB Structures. *Proteins*, **58**, 610-617.

Goodford,P.J. (1985) A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.*, **28**, 849–857.

Hendlich,M. *et al.* (1997) LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J.Mol.Graph.Model.*, **15(6)**, 359-363.

Huang,B.D and Schroder M. (2006) LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct.Biol.*, **6**,19-29.

Kawabata.T and Go.N. (2007) Detection of Pockets on Protein Surfaces Using Small and Large Probe Spheres to Find Putative Ligand Binding Sites. *Proteins*, **68**, 516–529.

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

Laurie,A.T.R and Jackson,R.M. (2005) Q-SiteFinder: an energy-based method for the prediction of protein–ligand binding sites. *Bioinformatics*, **21(9)**, 1908–1916.

Levitt,D.J and Banaszak,L.G. (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.

Shin,J.M and Cho,D.H. (2005) PDB-Ligand: a ligand database based on PDB for the automated and customized classification of ligand-binding structures. *Nucleic Acids Res*, **33**, 238-241.

Soga,S. *et al*. (2007) Use of amino acid composition to predict ligand-binding sites. *J. Chem. Inf. Model.*, **47**, 400-406.

Sonka,M. *et al*. (1998) Image Processing, Analysis, and Machine vision (Second Edition). Pws Pub Co. Chapter 5, Border tracing, 142.

Venkatachalam,C.M. *et al*. (2003) LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. *J.Mol.Graph.Model.*, **21**, 289–307.

Wade,R.C. *et al*. (1993). Further development of hydrogen bond functions for use in determining energetically favorable binding sites on molecules of known structure. 1. Ligand probe groups with the ability to form two hydrogen bonds. *J. Med. Chem.*, **36(1)**, 140–147.

Wade,R.C. and Goodford,P.J. (1993) Further development of hydrogen bond functions for use in determining energetically favorable binding sites on molecules of known structure. 2. Ligand probe groups with the ability to form more than two hydrogen bonds. *J. Med. Chem.*, **36**, 148–156.

Weisel,M *et al*. (2007) PocketPicker: analysis of ligand binding-sites with shape descriptors. *Chemistry Central Journal*, **1**,7-23.

Yaffe,E *et al*. (2008) MolAxis: a server for indentification of channels in macromolecules. *Nucl. Acids Res.*, **36**, 210-215.