PROTEINS: Structure, Function, and Bioinformatics 68:516–529 (2007)

Detection of Pockets on Protein Surfaces Using Small and Large Probe Spheres to Find Putative Ligand Binding Sites

Takeshi Kawabata[1]* **and Nobuhiro Go**[1,2]
[1]*Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, Nara 630-0192, Japan*
[2]*Quantum Beam Science Directorate, Japan Atomic Energy Agency, Souraku, Kyoto 619-0215, Japan*

ABSTRACT **One of the simplest ways to predict ligand binding sites is to identify pocket-shaped regions on the protein surface. Many programs have already been proposed to identify these pocket regions. Examination of their algorithms revealed that a pocket intrinsically has two arbitrary properties, "size" and "depth". We proposed a new definition for pockets using two explicit adjustable parameters that correspond to these two arbitrary properties. A pocket region is defined as a space into which a small probe can enter, but a large probe cannot. The radii of small and large probe spheres are the two parameters that correspond to the "size" and "depth" of the pockets, respectively. These values can be adjusted individual putative ligand molecule. To determine the optimal value of the large probe spheres radius, we generated pockets for thousands of protein structures in the database, using several size of large probe spheres, examined the correspondence of these pockets with known binding site positions. A new measure of shallowness, a minimum inaccessible radius, R_{inaccess}, indicated that binding sites of coenzymes are very deep, while those for adenine/guanine mononucleotide have only medium shallowness and those for short peptides and oligosaccharides are shallow. The optimal radius of large probe spheres was 3–4 Å for the coenzymes, 4 Å for adenine/guanine mononucleotides, and 5 Å or more for peptides/oligosaccharides. Comparison of our program with two other popular pocket-finding programs showed that our program had a higher performance of detecting binding pockets, although it required more computational time. Proteins 2007;68:516–529.** © 2007 Wiley-Liss, Inc.

Key words: binding site; pocket; protein surface; geometry; probe sphere

INTRODUCTION

As the structural genomics projects continue to solve 3D structures of proteins, whose functions are not well characterized, the prediction of ligand binding sites in protein structures is becoming critical for structural bioinformatics.[1–4] Identifying the ligand binding sites is also important for molecular docking and drug design. A simple methodology to achieve this goal is to find pocket (concave, cleft, hole)-shaped regions on the protein surface. Small molecules typically bind to pockets on the protein surface, as binding in a deep pocket is energetically favorable, due to the larger binding surface area. This large binding surface area provides the binding affinity to the specific ligands, as the protein can envelop many ligand atoms using geometric and chemical complementarities. In addition, the pocket excludes water molecules that can disturb enzymatic catalytic activity.[5,6] Many algorithms and programs have been proposed to identify these sites.[7–18] These programs are used for a variety of purposes, including studies about examining binding and active site,[19–21] prediction of active sites,[22,23] functional predictions using local 3D templates,[24] and a preliminary docking calculations.[7,13,17]

Although the concept of binding pocket has been widely accepted by researchers, the geometrical definition of pocket shape is not straightforward. A wide variety of pocket detection algorithms have been proposed, which can be classified into three methodological categories: grid-based, sphere-based, and α-shape-based. Grid-based methods cover the proteins in a 3D grid; empty grid points are then defined as pockets if they satisfy a number of geometric or energetic conditions.[8,9,11,17,18] Based on the theory of mathematical morphology, Delaney[8] and Masuya and Doi[11] defined pockets by *opening* and *closing* operations using probe spheres. Although these models are based on 3D grids, the defined pockets are considered to be the water-accessible space between the molecular surface by large probe spheres[25,26] and the van der Waals (VdW) surface. Levitt and Banaszak[9] and Hendlich et al.[14] developed other grid-based programs, POCKET

Grant sponsors: The Special Coordination Funds Promoting Science and Technology from the MEXT (Ministry of Education, Culture, Sports, Science and Technology of Japan); a Grant-in-Aid for Scientific Research on Priority Area (C), Genome Information Science, from the MEXT.

*Correspondence to: Takeshi Kawabata, Graduate School of Information Science, Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara 630-0192, Japan.
E-mail: takawaba@is.naist.jp

Received 1 June 2006; Revised 31 August 2006; Accepted 22 September 2006

Published online 19 April 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21283

and LIGSITE respectively, which focused on PSP (protein-solvent-protein) events of the grids. When a straight line drawn from a grid point is enclosed on both side by protein atoms, the arrangement of the line for that grid point is termed a PSP event. After examining several directions of lines, grid points having more than a threshold number of PSP events are defined as pockets. If a pocket region has fewer than the minimum threshold number of grids, it is excluded. Instead of the geometric conditions, Laurie and Jackson[18] introduced an energetic condition by placing a methyl group on a grid point and calculating interaction energy with protein atoms. In the sphere-based method, a set of probe spheres are placed on the protein surface.[7,10,13,16] The pocket detection program developed by Kuntz et al., as a part of their docking program, generated probe spheres, which were tangential to two surface points on the molecular surface.[7] Pocket spheres are those generated probe spheres that satisfy a number of geometric conditions, among the generated probe spheres. Laskowski developed the SURFNET program, which places a sphere (called *gap* spheres) between two protein atoms and reduces its radius until it just touches one protein atom.[10] A set of these *gap* spheres are defined as pockets. The concept of *gap* spheres is similar in principle to that underlying the PSP of the LIGSITE program.[14] Brady and Stouten also used probe spheres, placing a probe sphere with a fixed radius tangential to three protein atoms; they then deleted exposed probe spheres using a burial count (BC) value.[16] Ruppert et al.[13] employed an energetic condition by placing three types of probes (hydrohphobic, hydrogen bond donor, and acceptor) and calculating interaction energy with protein atoms. The α-shape-based method was developed using a new theory of the computational geometry.[12,15] An alpha shape is defined as a subset of Delaunay tessellations of protein atoms, omitting edges longer than the sum of the radii of two atoms. Peters et al.[12] and Liang et al.[15] defined a pocket as an empty Delaunay tetrahedron, one of whose edges is omitted in the α-shape. To avoid shallow empty tetrahedrons, both groups also employed additional conditions. Peters et al. removed shallow pockets with a deepness criterion, which is defined as a distance from the atom center to the enveloping surface area.[12] Liang et al. employed a discrete flow method, which extracts only obtuse empty tetrahedrons and their neighbors.[15]

Although these methods vary considerably, we found that all the pocket-finding programs arbitrarily decide on two properties of the pocket: "size" and "depth" (Fig. 1). The "size" property defines the minimum size of the pocket. As a binding pocket should have a space into which a binding ligand enters, we must assume the size of binding ligands to identify the locations of pockets [Fig. 1(a)]. The other important property "depth" signifies how deeply a pocket is buried from the outside. This measurement is equivalent to determining the boundary between the pocket region and the open outer space [Fig. 1(b)], which was called "sea-level" by Laskowski et al.[10,19] The "depth" property depends on the type of binding ligand, and its binding stability and specificity. In the SURFNET program, the "size" is decided from the minimum and maximum radius of the gap spheres.[10] These two parameters also implicitly control the "depth". In the LIGSITE program, the "depth" is decided by the threshold number of PSP events, while the "size" is decided by the minimum threshold number of grid points.[14] In the α-shape methods, the "depth" is decided by the planes generated by Delaunay tessellation, the minimum size is decided by the additional conditions, such as deepness[12] or discrete flow.[15] Algorithms with easily-controllable parameters are preferred, because optimal definitions of these two properties also depend on the putative ligand.

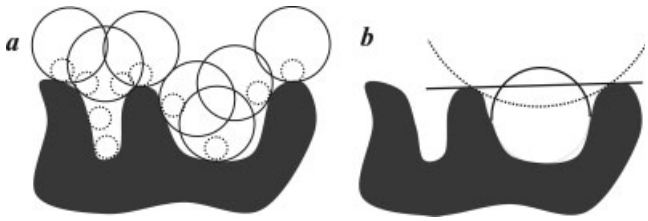


Fig. 1. Two properties for defining pockets. (***a***) "Size" property. The filled black region represents a protein with two different sized pockets, while circles are potential binding ligands. If a binding ligand is the size of the dashed small circle, both the right and left pockets are putative binding sites. If a ligand has a size of the solid large circle, only the right pocket is a putative binding site; the solid large circle can enter into the right pocket, but not into the left pocket. (***b***) "Depth" property. Binding pockets are expected to be deeply buried relative to the outer space. This figure demonstrates the three boundaries between a pocket and the outer space. A solid straight line designates a plane tangential to the entrance atoms of the pockets. This boundary is similar to that used in the LIGSITE program with using the parameter MIN_PSP = 1, and that used in the α-shape method. A dashed curved line represents a largest sphere that can contact the pocket. This boundary is similar to that used in the SURFNET program. A dashed circle designates a tangent sphere with a given radius. This boundary is similar to that used in the method of Masuya and Doi[11] and Peter et al.[12]

Considering the two arbitrary properties used to define a pocket, we proposed a new definition for pockets with two explicit controlling parameters. In our algorithm, small and large probe spheres are placed on the protein VdW surface; pocket regions are defined as a space into which a small probe can enter, but a large probe cannot. As proposed by Brady and Stouten,[16] we employed three-contacting probe spheres in our algorithm. The "size" and "depth" properties can thus be controlled directly by two arbitrary parameters: the small and large radii of the probe spheres.

To determine the optimal probe radius, we applied our program to a large number of structures, using several different sizes of large probes. The shallowness of ligand-binding pockets was measured by a minimal inaccessible radius R_{inaccess}. We found that each class of ligand has a characteristic deepness and size of their cognate binding pockets. By comparing our program to two proposed pocket-finding programs, we discuss the advantages and disadvantages of our program, suggesting appropriate parameters for the use of our program.

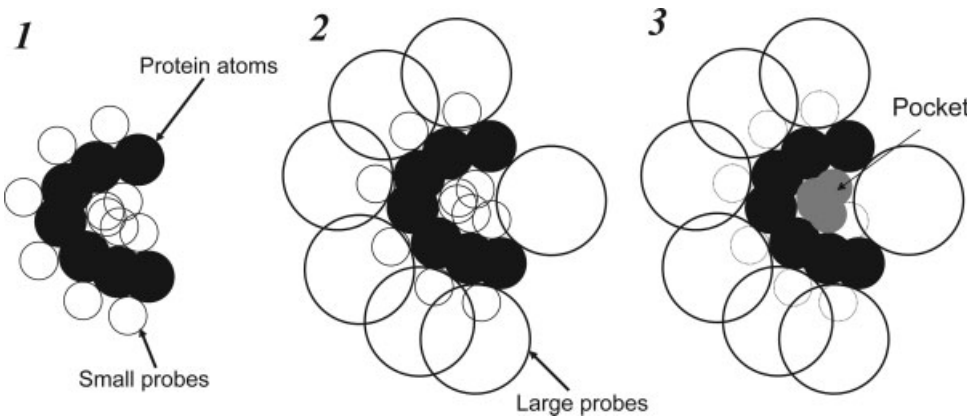


Fig. 2. Basic procedures used in our pocket detection program. (**1**) A small probe, a surface small probe, was placed on the protein VdW surface. (**2**) A large probe was placed on the protein VdW surface. (**3**) Those small probes that overlap with the large probes were removed. The small probes that remain were defined as "pockets".

MATERIALS AND METHODS

Outline of Pocket Detection Processes

Pocket regions are defined as a space into which a small probe sphere can enter, but a large probe sphere cannot. Our algorithm for pocket detection utilizes three steps (Fig. 2). First, small probe spheres are placed on the protein VdW surface. These small probes are also called "surface probes". Then, large probe spheres are placed on the protein VdW surface. Finally, any small probe spheres overlapping the large probe spheres are removed (shaving process). The remaining small probe spheres are defined as "pocket" probes, while the space occupied by the pocket probes is defined as "pocket" regions. To facilitate more rapid calculation, we placed probes in contact with three protein atoms. Although probes did not overlap with any protein atoms, the probes were allowed to overlap with one another. We named the program performing this algorithm, "PHECOM" ("P"robe-based "HECOM"i finder), after the Japanese word "hekomi", which means dent or concave shape.

Our definition of pocket has two explicit arbitrary parameters, the radius of the small probe, R_{small}, and the radius of the large probe, R_{large}. In this study, we fixed the value of R_{small} at 1.87 Å, which is the size of a single methyl group ($-CH_3$). We focused instead on defining the optimal radius for the large probe spheres, R_{large}. Initial attempts to define R_{small} as 1.40 Å, the approximate size of a water molecule, generated too many small pockets for prediction of ligand binding sites. When we employed a much larger R_{small} than 1.87 Å, it was difficult to interpret chemically a probe as a binding atom or molecule.

Figure 3 explains the manner in which different pocket regions can be detected by different sizes of large probes. For a model protein, all the surface small probes are shown in Figure 3(a). Figures 3(b–d) correspond to the small probes remaining after the shaving process using a plane or spheres with 12 and 4 Å radii, respectively. Later, we will discuss the optimal radius of large probe spheres for ligand-binding site detection.

Placing Probes

As proposed by Brady and Stouten,[16] probe spheres were placed so as to be tangential to three protein atoms, without overlapping with any of the protein atoms. We calculated the position of the center of the probe sphere with a given radius and the three tangent protein atoms using an algorithm proposed by Connolly[25,26] and Brady and Stouten.[16] We then confirmed that these probes did not overlap with other protein atoms. As the generated probe spheres often overlapped with other protein atoms, the computational cost necessary to check crashes between probe spheres and protein atoms was large, especially for large probe spheres. To reduce the computational costs, we introduced an efficient heuristic algorithm by modifying the incremental algorithm for convex hull of a set of points.[27] The details of the algorithm are described in the Appendix section. It enabled us to place probe spheres with any radius within a reasonably small computational time.

To calculate two planes tangential to three given spheres, we used an algorithm described by Yeates.[28] Calculation of all of the tangent planes surrounding all of the protein atom spheres was performed using an algorithm similar to that used to generate tangent spheres, which is described in the Appendix.

Radii of Atoms

The value of van der Waals radii, taken from Chothia,[29] were as follow: oxygen 1.65 Å, trigonal nitrogen 1.65 Å, tetrahedral nitrogen 1.50 Å, tetrahedral carbon 1.87 Å, trigonal carbon 1.76 Å, and sulfur 1.40 Å. Hydrogen atoms were not considered.

Clustering of Pockets

Pocket probes were grouped using a single linkage clustering algorithm. If the distance between two probes was smaller than the sum of their radii, these atoms were grouped into one cluster. If a probe belonging to one

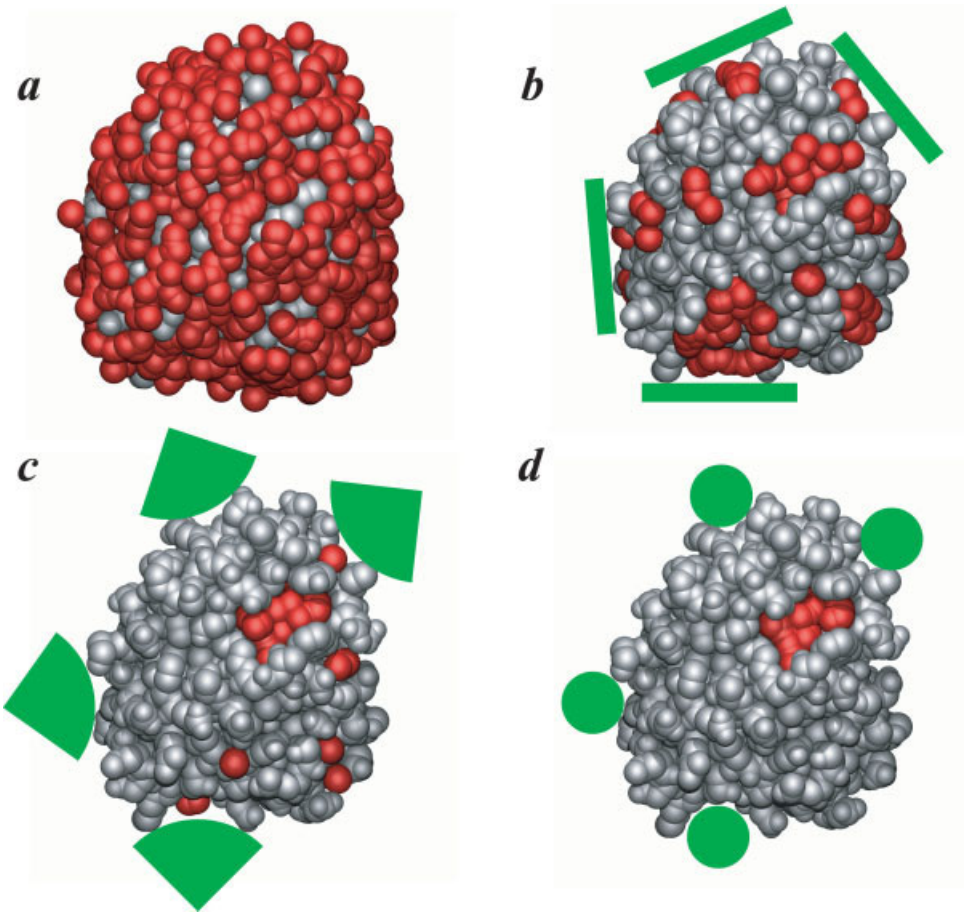


Fig. 3. Generated probes and remaining pocket probes shaved by large probes of differing sizes. The gray spheres represent the atoms of a protein (flavodoxin, PDBcode:1flv), while the red spheres are the small probes with a radius of 1.87 Å. Large spheres or planes are schematically shown in green. (**a**) The 1244 small probes generated on the protein surface before the shaving process. (**b**) The remaining 412 pocket probes after shaving by planes (spheres with infinite radius). (**c**) The remaining 139 pocket probes after shaving by probe spheres with a 12 Å radius. (**d**) The remaining 80 pocket probes after shaving by probe spheres with a 4 Å radius. The largest probe cluster corresponds to the FMN binding site.

cluster was connected to at least one probe of another cluster, these two clusters were grouped into one cluster.

Volume of Pockets

The volume of pockets is an important property to characterize the detected pockets. Our pockets were defined by a set of multiple probe spheres that overlapped with one another. As it is difficult to calculate analytically the volume of overlapping spheres, we employed an approximated method using a 3D grid, originally proposed by Laskowski.[10] We generated a 3D grid covering probe spheres, expressing a density of probe spheres at each grid position. The density generated by the i-th probe sphere is represented by the following sigmoid function:

$$f_i(\boldsymbol{x}) = \frac{1}{1 + \exp[k(|\boldsymbol{x} - \boldsymbol{r}_i| - R_i)]}$$

where $\boldsymbol{x}$ is the position on the 3D grids, $\boldsymbol{r}_i$ is the center of the i-th probe sphere, R_i is the radius of the i-th probe sphere, and k is a constant. The density F of the point $\boldsymbol{x}$ on the 3D grids is defined as follows:

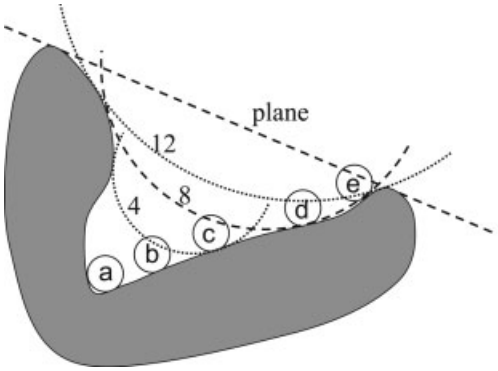
$$F(\boldsymbol{x}) = \arg\max_{i \in atoms} f_i(\boldsymbol{x})$$

The surface of probes is defined as an isocontour on a 3D grids with value $F(\boldsymbol{x}) = 1.0$. The polygon for the isocontour of 3D grids is calculated using the marching cube algorithm. We used 1 Å-width grids and $k = 10$. We introduced a unit M_{probe} to measure the volumes of the probe clusters. M_{probe} is an efficient number of probes composing a probe cluster, defined as follows:

$$\text{Mprobe} = \frac{[\text{Volume of probe clusters}]}{[\text{Volume of one probe sphere}]}$$

Minimum Inaccessible Radius: R_{inaccess}

To measure the shallowness of a small probe p placed on the protein VdW surface, we introduced R_{inaccess}, a



	trying *Rlarge* (Å)				*Rinaccess* (Å)
	4	8	12	plane	
a	-	-	-	-	4
b	T	-	-	-	8
c	T	T	-	-	12
d	T	T	T		12-∞
e	T	T	T	T	plane-accessible (convex)

Fig. 4. Schematic definition of R_{inaccess}. After using three sizes of spheres (4, 8, and 12 Å) and planes as large probe spheres (left figures), we classified the small surface probes into five types: a, b, c, d, and e. Each class has a corresponding R_{inaccess}, as summarized in the table (at right). If none of the four large probes touched a surface small probe (type a), the R_{inaccess} was the smallest radius of the tested large sphere, 4 Å. If a 4 Å sphere can contact the probe, but an 8 Å sphere cannot (type b), its R_{inaccess} value was 8 Å. If only the largest sphere can touch it, but the plane cannot (type d), the R_{inaccess} was defined as "12-∞". If the plane could touch the small surface probe (type e), this probe would not survive the shaving process using any of the large probe spheres. Its R_{inaccess} was then assigned as "plane-accessible (convex)".

minimum inaccessible radius.

$R_{\text{inaccess}}(p)$: the minimum radius R satisfying the condition that all of the probes q in $Q(m,R)$ do not touch the probe p.

$Q(m,R)$: a set of probe spheres with radius R contacting the VdW surface of protein m.

Thus, $R_{\text{inaccess}}(p)$ is the minimum radius of the large probe spheres allowing the small probe p to survive after the shaving of large probes. As it is difficult to estimate this value analytically, we estimated R_{inaccess} by generating several large probes spheres with different radii, for which a schematic example is shown in Figure 4. Using sphere radii (4, 8, and 12 Å) and planes, the small probe can have one of five R_{inaccess} values: 4, 8, 12, "12-∞", or "plane-accessible".

Dataset

A structure dataset for the statistics of the pockets and ligand binding sites was prepared from the SCOP database, version 1.69.[30] The dataset included protein chains with mutual sequence identities of 40% or less. We excluded from the dataset small protein with less than 40 residues and protein chains with domains in class f (membrane proteins), h (coiled-coil proteins), i (low-resolution), j (peptide) and k (designed). As a result, 5405 chains were included in the dataset. We removed the ligands that primarily bound another chain in the same PDB entry. In the final dataset, 4071 chains contacted at least one other molecule other than water molecules. The total number of contacting macromolecules (DNA, RNA, proteins) was 2849, while that of non macromolecular ligands, excluding waters was 4295. The minimum length of DNA and RNA was three, while that for protein is 10. When we selected the specific molecule types appearing frequently in the structural dataset, we excluded small and unnatural molecules, those containing less than seven heavy atoms or those with less than three carbons, and the precipitants: GOL, MPD, TRS, MES, BOG, EPE, and DTT. We checked the covalent bonds between monosacchrides. When the overlap of two atom spheres with VdW radii was more than 1 Å, these two atoms were defined as covalently-bonded. If monosaccharides were covalently bonded to each other, these molecules were grouped as "oligosaccharide" ("osc" in the three letter representation in Fig. 7 and 12). We also identified the covalent bonds between sugars and proteins. When we examined the specific ligand types, sugars covalently bonded to proteins were treated as a different molecule type from sugars that were not bonded to proteins.

RESULTS

Distribution of Shallowness R_{inaccess} for Surface Probes

To determine the relationship between the properties of pockets and their ligand-binding, we applied our pocket detection program to a set of structures in the representative list of SCOP database. First, we focused on the shallowness property of pockets. To determine the radius of large probes spheres optimal for recognition of binding sites, we tested 14 R_{large} values (spheres with radii of integer values between 3 and 15 Å, and a plane) using the structural database. Using these large probes of different sizes, we calculated the shallowness measure R_{inaccess}, a minimum inaccessible radius for each surface small probe. As 14 kinds of large probes were calculated, R_{inaccess} can have 15 kinds of values: 3–15 Å, "15-∞" (15 Å sphere accessible, but plane-inaccessible), and "plane-accessible". Figure 5 displays the distribution of R_{inaccess} values for surface probes in the structural database. The "plane-accessible" category comprised the highest peak for R_{inaccess} values in the distribution, as 35% of surface probes could

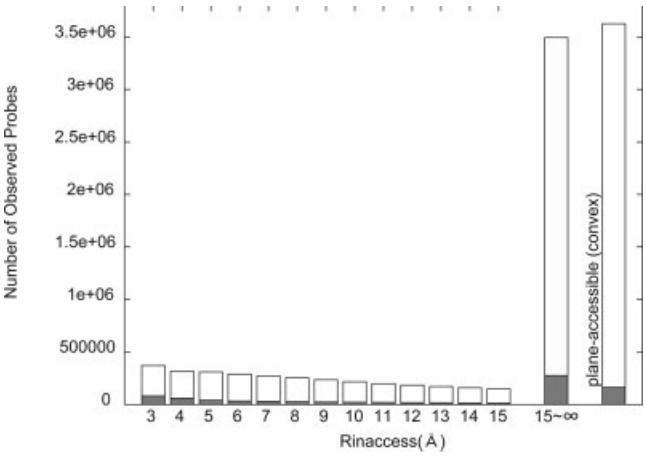


Fig. 5. Distribution of R_{inaccess} values for all of the surface small probes in the database (white bars). Gray bars are the numbers of ligand-overlapping probes for each probe with a specific R_{inaccess} values.

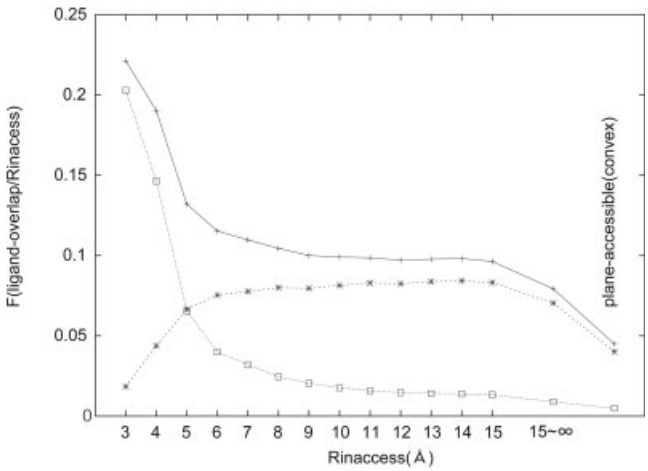


Fig. 6. Ratio of ligand-overlapping probes against probes with a specific R_{inaccess}. Crosses are ligand-overlapping probes for all the ligands excluding waters, stars indicate those for macromolecules (DNA, RNA, proteins with more than 10 residues), while squares represent those for nonmacromolecules.

be shaved by a plane. The fact was also supported by the large difference of number of red spheres between Figures 3(a,b). The second highest peak is "15-∞", which means the pocket probe is accessible to a 15 Å sphere, but inaccessible to a plane. This peak may be due to the fact that we did not try probe spheres larger than 15 Å. If we had evaluated larger probe spheres, the "15-∞" peak would be distributed throughout these new values. In comparison to these two peaks, the bins of the remaining R_{inaccess} values are low; the smaller R_{inaccess} bins, however, contained slightly more probes than the larger ones.

R_{inaccess} for Ligand-Overlapping Probes

We also identified the number of small surface probes that overlapped other molecules in the structures in the dataset, defined as "ligand-overlapping probes". The gray bars in Figure 5 designate the numbers of the ligand-overlapping probes with specific R_{inaccess} values. To examine the dependence of ligand binding on pocket shallowness, we calculated the ratios of ligand-overlapping probes to probes with a specific R_{inaccess} values (Fig. 6). These results clearly demonstrate that this ratio of ligand-overlapping probes becomes greater for probes with smaller R_{inaccess} values, suggesting that the ligands generally tend to bind in deeper pockets. Twenty two percent of pocket probes with $R_{\text{inaccess}} = 4$ Å in the database overlapped with ligands. We roughly classified all of the ligands, with the exception of water molecules, into two classes, macromolecule and non-macromolecule. DNA, RNA, and proteins with greater than nine residues were classified as macromolecule, while all other ligands were classified as nonmacromolecule. Peptides with less than 10 residues were treated as the non macromolecular ligands. Macromolecules did not exhibit a preference for deep pocket regions, in comparison to nonmacromolecules (Fig. 6).

R_{inaccess} for Probes Overlapping Specific Ligands

To determine the optimal large probe sphere radius for a specific ligand type, we selected several molecules frequently appearing in the structural dataset. We then examined the distribution of R_{inaccess} values for the overlapping surface probes (Fig. 7). We examined the 24 ligands that were observed at least 15 times within the dataset. Metals (such as Ca, Zn, and Mg) and precipitants, such as glycerol (GOL) and 1,2-ehanediol (EDO), were ignored. The majority of molecules binding in deep pockets were coenzymes, including SAH, PLP, HEC, HEM, FAD, SAM, and NDP. Approximately 60% of the probes overlapping with atoms in SAH, PLP, HEC, and HEM displayed a R_{inaccess} value 3 Å. Adenine/guanine mononucleotides, such as AMP, ADP, ANP, ATP, and GTP also bound to pockets; the R_{inaccess} values for these molecules, however, were larger than those for the coenzymes. The macromolecules, DNA, RNA, and proteins, were less likely to bind pocket regions than nonmacromlecules. The R_{inaccess} values for approximately half of probes overlapping these macromolecules were "15-∞" or "plane-accessible". Their distribution did not differ significantly from that of all the surface probes (the right bar "surface"). Peptides, defined as proteins with less than 10 amino acids, were much more likely bind in pockets. Sugars covalently bonded to proteins, such as oligosaccharides [osc(bonded)], *N*-accetyl-D-glucosamine [NAG(bonded)], and mannose [MAN(bonded)] were unlikely to bind pocket regions, whereas those not covalent bound to proteins (osc, NAG, and MAN) were more likely bind in pockets. Glucose (GLC) tended to bind to much deeper pocket than NAG.

Distribution of Pocket Size

Next, we focused on the size of the pocket probe clusters. The small pocket probes were clustered by a single linkage clustering algorithm. The calculated volume of the cluster was then calculated using M_{probe}, which determines the efficient number of small probes within the cluster (see Material and methods). Distribution of pocket sizes depends on the radius of large probes, R_{large}. Larger R_{large}

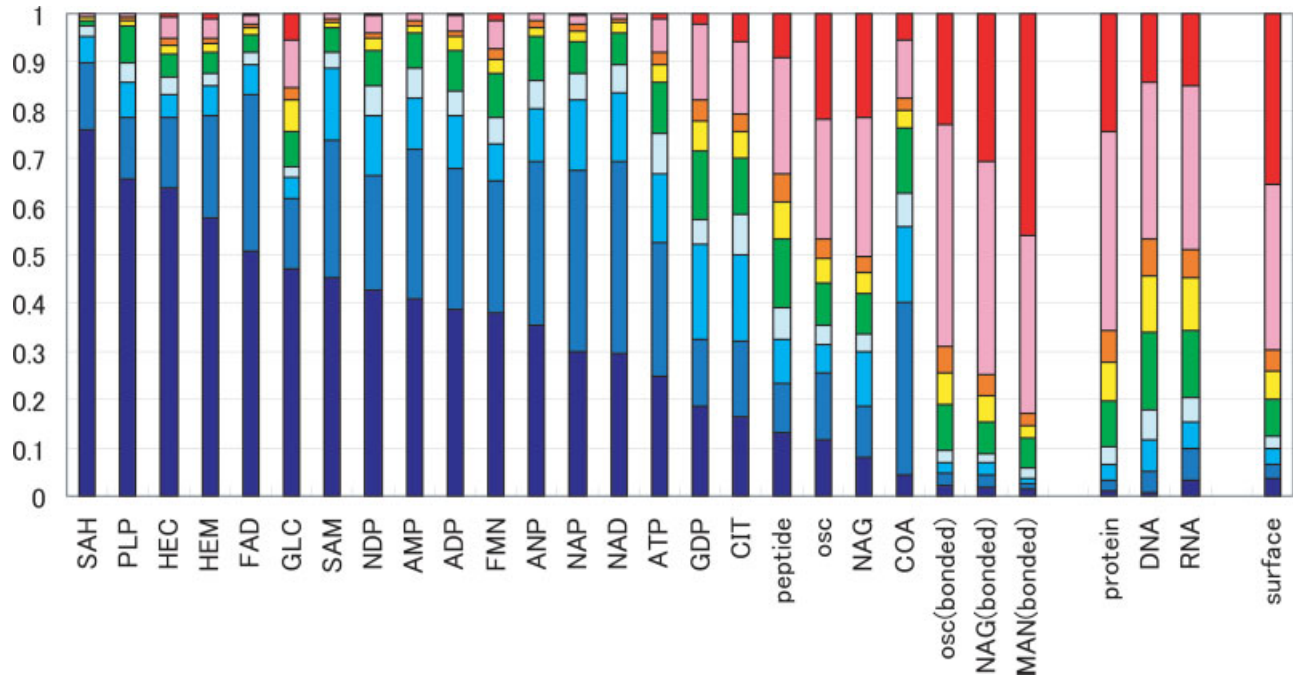


Fig. 7. The proportion of R_{inaccess} values for surface probes overlapping specific ligands. Each color corresponds to a specific R_{inaccess} value. The deep blue, blue, cyan, and light blue correspond to $R_{\text{inaccess}} = 3$, 4, 5, and 6 Å, respectively. The green, yellow, and orange colors correspond to $R_{\text{inaccess}} = 7$–9, 10–12, and 13–15 Å respectively. The pink and red correspond to $R_{\text{inaccess}} =$ "15-∞" and "plane-accessible", respectively. The bar above the "Surface" label indicates the proportion of R_{inaccess} values for all of the surface small probes. osc stands for "oligosaccharides". Molecular names containing "(bonded)" are saccharides covalently bound to proteins.

values produced greater number of pockets of all sizes. Smaller pockets were more frequently observed than larger ones, giving an almost linear relationship between size and observed number of pockets in the log–log plots (data not shown). These results indicate that the pocket incidence is proportional to the power of pocket size.

Figure 8 demonstrates the distribution of pocket size using large sphere of 4 Å. The gray bars indicate numbers of pocket clusters for all kinds of ligands, excluding water molecules. Figure 9 displays the ratio of ligands, including pocket clusters, to all the pocket clusters of different sizes. Larger pockets clusters tended to contain nonmacromolecular ligands. This tendency, however, is not observed macromolecules. As the number of observed tiny pockets is very large (Fig. 8), and the ratio of these pockets including ligands is small (Fig. 9), filtering out the tiny pockets clusters (e.g., $M_{\text{probe}} \leq 2$) could efficiently improve the signal-noise ratio of binding site prediction.

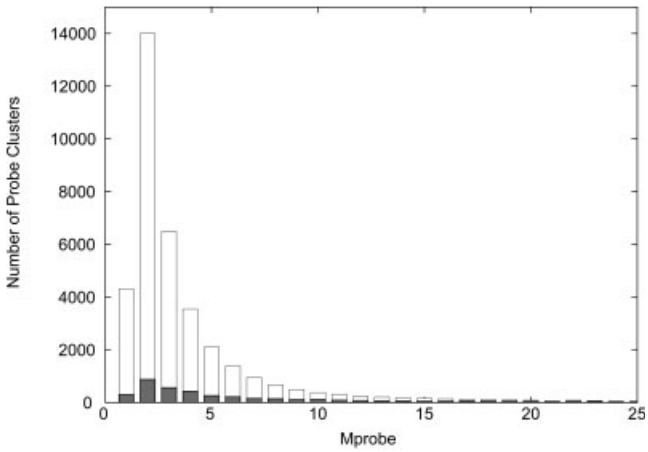


Fig. 8. Size distribution of pocket clusters spheres in the structural database defined by 4 Å large. The gray bars indicates the number of pocket clusters including any ligand type.

Pocket Size for Each Ligand

We explored the size distribution of pocket probe clusters for specific ligands. After choosing a specific ligand molecules (such as FAD) from the structural database, we searched for the pocket probe cluster that mostly overlapped with this molecule, The M_{probe} value of this pocket was defined as the pocket size of this ligand molecule. We plotted the distribution of pocket sizes determined defined by 4 Å large probe spheres, which are rather deep pockets (Fig. 10). Red bars correspond to small pockets ($M_{\text{probe}} \leq 2$), while blue bars correspond to large pockets ($M_{\text{probe}} > 40$). Gray bars define those ligands that did not overlap with any pocket clusters. In comparison to our analysis of the distribution of pocket shallowness (Fig. 6), molecules that favors deep pockets, such as HEM, FAD, also favors large pockets. There are some exceptions, however. While PLP binds to deep pockets, their sizes are not large. This result suggests that the size of pocket clusters provides additional information in the prediction of putative binding molecules. The volume of a pocket

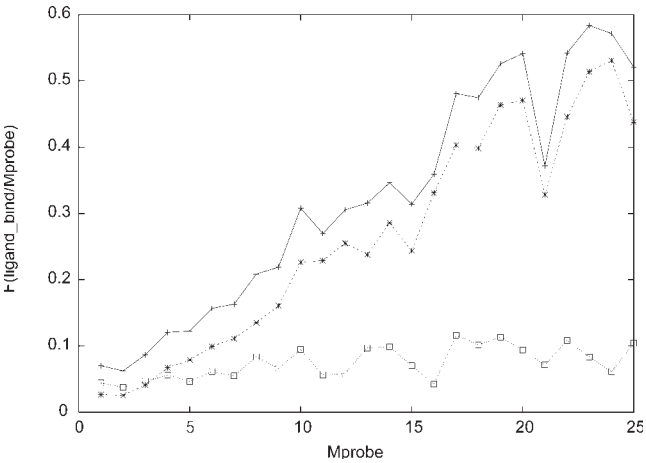


Fig. 9. The ratio of ligands included in pocket clusters to all the pocket clusters with specific size M_{probe}. The line with crosses indicates pockets including ligands. The line with stars indicates pockets including macromolecules (DNA, RNA, and proteins with more than 10 residues). A line with squares indicates pockets including nonmacromolecules.

containing a coenzyme is generally larger than that of the coenzyme itself. For example, the volume of HEM is $\sim 565\ A^3 = 27.5\ M_{probe}$, while that of FAD is $623\ A^3 = 30.3\ M_{probe}$. Pockets with HEM or FAD, however, often have an M_{probe} volume greater than 40 (blue bar in Fig. 10). This is because coenzyme-binding pockets often include spaces for substrate molecules. In addition, several HEM-containing pockets contain more than one HEM molecules.

The Optimal R_{large} Value for the Specific Type of Ligands

Each type of ligands has a specific distribution of pocket shallowness ($R_{inaccess}$) as shown in Figure 7, suggesting that each specific ligands has an optimal R_{large} value for the effective prediction of ligand binding sites. Here we address the determination of optimal R_{large} value for binding site prediction, if the rough properties of putative binding ligands are known. We classified the ligands listed in Figures 7 and 10 into three classes: coenzymes (SAH, PLP, HEC, HEM, FAD, SAM, NDP, FMN, NAP, and NAD), A/G mononucleotides (AMP, ADP, ANP, ATP, and GDP), and peptides/oligosaccharides (peptides and osc). We selected the structures that contained the ligands belonging to one of these classes among the 5405 representative chains. The number of structures for the coenzymes, A/G mononucleotides, and peptide/oligosaccharides are 421, 138, and 109, respectively.

To compare the correspondence between the pocket regions and the binding ligands, we employed grid representations of pockets and ligands. Briefly, grid points are placed around proteins and pockets with 1.0 Å width; each point was checked to determine if it was inside of calculated pocket regions. The same procedure was applied to the ligands. We then calculated the Recall and Precision values to examine the correspondence of pockets with ligands:

$$\text{Recall} = \frac{N_{LP}}{N_L}$$

$$\text{Precision} = \frac{N_{LP}}{N_P}$$

where N_P is the number of grid points in pockets, N_L is the number of grid points overlapping with ligands, and $N_{P\ L}$ is the number of grid points in pockets that overlapped with ligands. Generally speaking, pocket detections using larger R_{large} values tend to have higher Recall values and lower Precision values. To find an R_{large} value that provides a good balance between Precision and Recall, we employed the F-measure defined as follows:

$$\text{F - measure} = 2\left(\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}\right)^{-1}$$

The F-measure is the harmonic mean of Precision and Recall, often used for information retrieval.

F-measure values for the results for PHECOM using several R_{large} values are summarized in Figure 11. The line for the coenzymes class is higher than those for A/G mononucleotides and peptides/oligosaccharides, suggesting that the prediction power for coenzyme binding sites is better than that for the other two classes. Each ligand class has different optimal R_{large} value providing the highest F-measure. The optimal R_{large} value for coenzymes is 3–4 Å, that for A/G mononucleotides is 4 Å, and that for peptides/oligosaccharides is 5 Å or more, respectively. Figure 12 shows three successful examples of detected pockets for the three cases. These optimal values are not strict, as the F-measure is only one of the many proposed measures, however, they serve as reasonable parameter recommendation for users of our program.

Comparison With Pockets Detected by Other Programs

Many pocket detection programs have been developed; comparison of our program to these programs clarifies the advantage and disadvantage of our method. We calculated pockets for our dataset using two freely available programs, SURFNET[10] and PASS.[16] We used two freely available programs, SURFNET and PASS, to calculate pockets. Using the PHECOM program, 14 sets of pockets were calculated using 14 different large probe sphere sizes. To standardize the output of the programs, we employed grid representations of pockets, described in the previous section.

Of the total volume of pockets detected by the three different programs (Fig. 13), that detected by SURFNET was the largest. The pocket volume determined with PASS was similar to that of PHECOM using $R_{large} = 6$ or 8 Å. We characterized these different pockets using the minimum inaccessible radius $R_{inaccess}$, assuming each grid point was a small probe sphere of zero radius. For the grid points in the detected pockets (see Fig. 14), the $R_{inaccess}$ values determined by the SURFNET program were distributed throughout a higher region than those

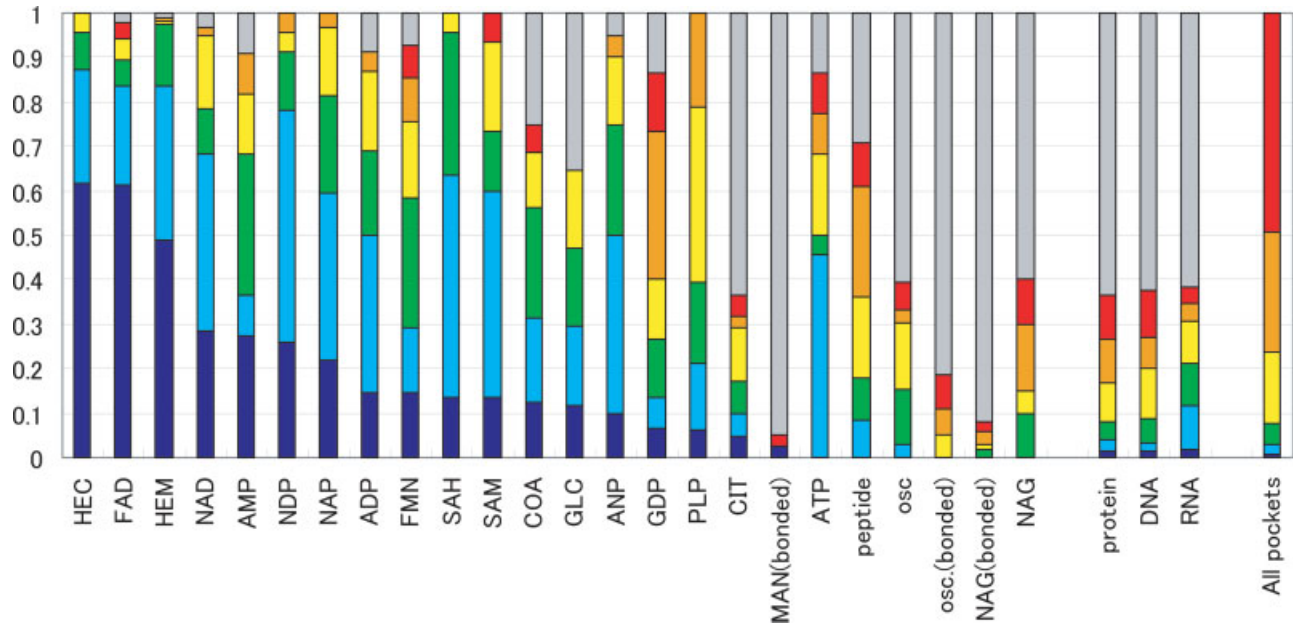


Fig. 10. The proportion of pocket clusters of a defined size that including specific types of ligand molecules. Pockets were determined using 4 Å large probe spheres. Each color corresponds to a value of M_{probe} value. The range of red is $0 < M_{probe} \leq 2$, (0–2), while orange is 2–4, yellow is 4–10, green is 10–20, cyan is 20–40, and blue is greater than 40.

determined by the other two programs. The volume of the smallest $R_{inaccess}$ bin (deepest pocket) was the lowest for the SURFNET program among all three tested. The distribution values from the PASS program was similar to that of PHECOM using a 6 Å large probe; PASS, however, identified a larger number of pockets with high $R_{inaccess}$ values.

Next, we determined the number of pocket grids that overlap with actual binding ligands. Nonmacromolecules were considered as ligands, excluding DNA, RNA, and proteins. After conversion of the ligand atoms to grid points, we measured the overlap of ligand with pocket grid points, using the Recall and Precision, described in the previous section. The Recall-Precision plot shown in Figure 15 summarizes the correspondences of pocket and ligands for several programs; the curves and points plotted in the upper right have better prediction capacities than those in the lower left. The Recall measure tended to have larger values for methods that output more pocket points, explaining the high Recall values seen for the SURFNET and PHECOM using larger R_{large} values. In contrast, the Precision measure tended to have a larger value for methods that output a smaller number of pocket points. The blue line in Figure 15, which corresponds to the PHECOM results, is plotted over the points for SURFNET and PASS, indicating that the use of the proper R_{large} parameter in the PHECOM program exhibits better both Recall and Precision to detect binding sites than these two popular pocket-finding programs. The green line, which is plotted over the blue line, corresponds to the results by PHECOM program filtering out the tiny pockets clusters ($M_{probe} \leq 2$). This result demonstrates that filtering out the tiny pockets enhance the prediction power of our program. For finding the binding site for the specific type of ligands, F-measure values for the SURFNET and PASS are summarized in Figure 11. This figure also shows that F-measure of PHECOM with the proper R_{large} values is larger than those of SURFNET and PASS, for all the three classes of ligands.

Computational Time

A disadvantage of the PHECOM program is a longer computational time. We examined the computational times needed by the three programs to detect pockets for several protein structures, using a Pentium 4, 3.0 GHz CPU for the calculations. For smaller proteins, the computational times necessary for these three programs did not differ significantly between the three programs. For flavodoxin (PDBcode:1flv, 1326 atoms), SURFNET, PASS, and PHECOM required 4.3, 0.9, and 5.4 s respectively. For the large proteins, however, PHECOM required longer computational times than the other two programs. For dipeptide-binding protein (PDBcode:1dpe, 4066 atoms), SURFNET, PASS, and PHECOM required 8.2, 3.8, and 94.5 s, respectively. For a protein with 8000 atoms, PHECOM finished calculating the pockets, within 13 min. For the PHECOM program, the computational time was approximately proportional to the square of the number of atoms. Considering the high performance of the PHECOM program, however, this additional time may be trivial.

DISCUSSION

We defined a pocket region as a space into which a small probe sphere can enter, but a large probe sphere

cannot. Our definition of pocket is not a completely novel idea. As we explained in Introduction section, all of the proposed pocket detection programs have two similar criteria corresponding to depth (large probe radius) and size (small probe radius) of pockets. Especially, the definition by Masuya and Doi[11] was similar to our definition; although it allowed limited sizes of probe spheres because they used the grid-based methods. The advantage of our definition is its geometric simplicity and explicit definition of the two freely controlling parameters. Our program can generate any size of probe sphere, whose radius ranges from 1.4 A to infinity (plane).

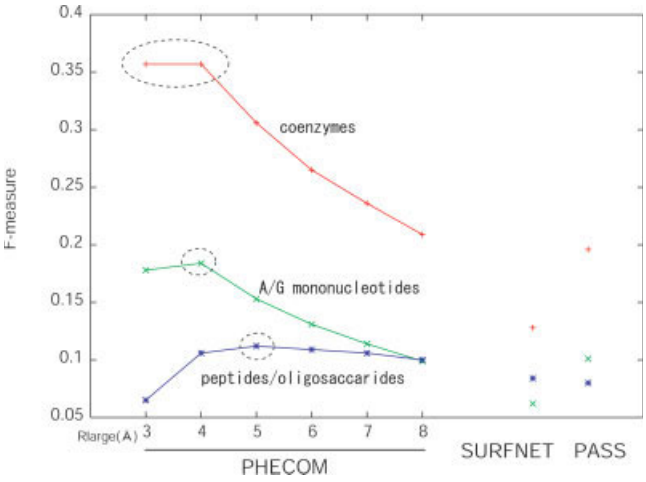


Fig. 11. F-measures for the correspondence of binding ligands with pocket regions detected by several programs. The red line and points corresponds to the binding sites of coenzymes (SAH, PLP, HEC, HEM, FAD, SAM, NDP, FMN, NAP, and NAD). The green line and points correspond to the binding sites of A/G mononucleotides (AMP, ADP, ANP, ATP, and GDP). The blue line and points corresponds to the binding sites of peptides/oligosaccharides (peptides and osc). The best method for each ligand class is emphasized by a dotted circle.

In this study, we fixed the radius of small probe spheres to 1.87 Å, instead focusing on the optimal radius of the large probe spheres. Binding sites for specific ligand in the structural database have a specific distribution of the shallowness (Figures 6, 7, 11, and 12). It suggests that the optimal depth parameter (radius of large probe spheres) depends on binding molecules. We recommend that users choose the R_{large} parameter considering their putative binding molecules. From the F-measure values, the optimal R_{large} value was 3–4 Å for the coenzymes, 4 Å for adenine/guanine mononucleotides, and 5 Å or more for peptides/oligosaccharides. We think that choosing the R_{large} parameter is not too of a much requirement for most of the users, because rough properties of binding molecules are often known in case of binding site prediction and docking calculation. If putative binding molecules are not known, we recommend calculating pockets with several R_{large} parameters, for example, 3, 6, and 8 Å. If deep and large pockets are found, they will be surely binding sites (Figures 6 and 9). If only shallow pockets are found, predic-

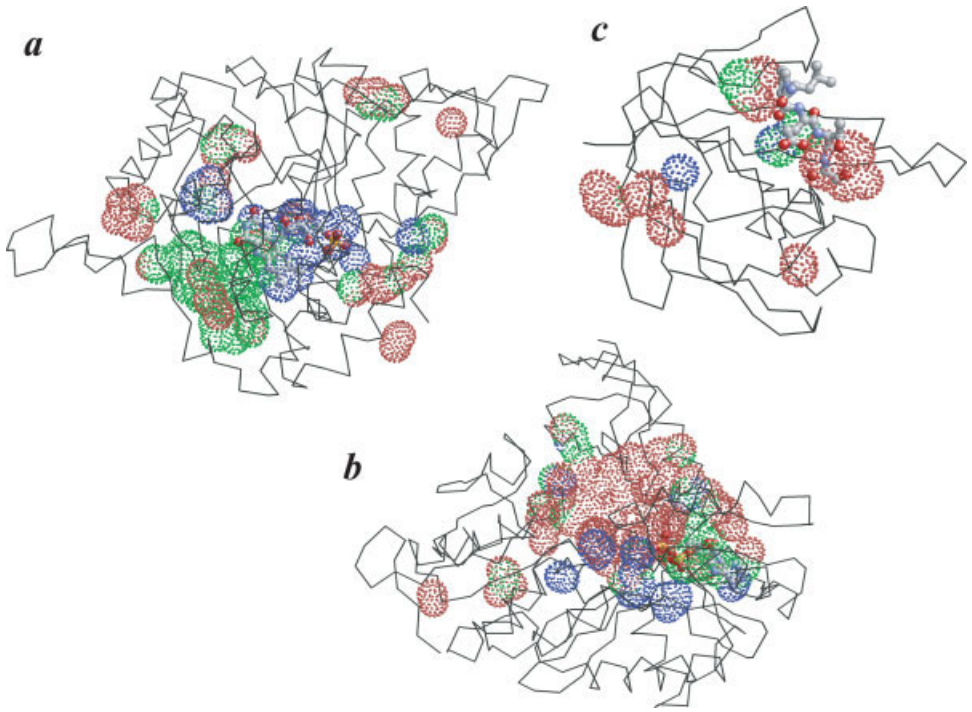


Fig. 12. Examples of detected pockets and binding ligands for the three classes. Dotted spheres are detected pockets by the PHECOM program. Blue, green, red spheres are pockets detected by $R_{\mathrm{large}} = 3$, 4, and 5 Å, respectively. Blue pockets are deep, red pockets are shallow. Binding molecules are drawn by the ball-and-stick model. (**a**) Mandelate dehydrogenase (PDBcode: 1p4c, chain A). A FMN molecule is bound. (**b**) Phosphoribosylglycinamide formyltransferase (PDBcode: 1kjq, chain A). An ADP molecule is bound. (**c**) Sortase A (PDBcode: 1t2w, chain A). A five-length peptide (LPETG) is bound.

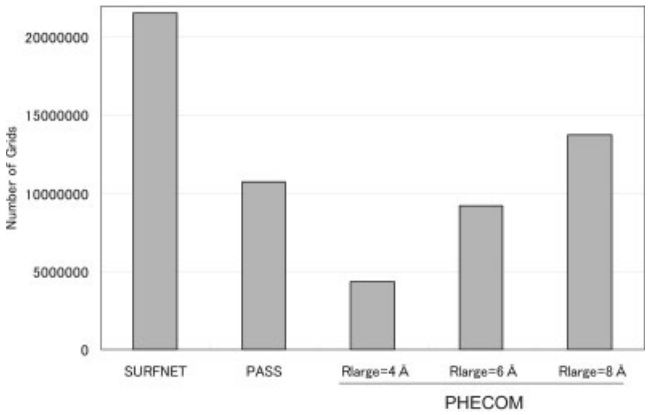


Fig. 13. Volume of pockets detected by several different programs. The vertical axis designates the total number of grid points of the pockets found in the dataset. The width of grid was 1 Å.

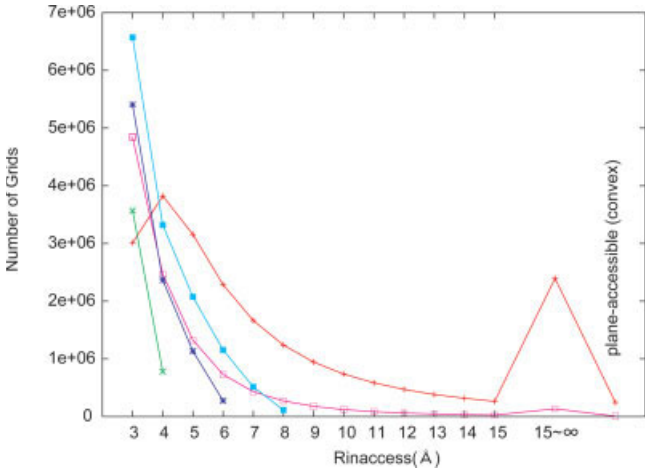


Fig. 14. Distribution of R_{inaccess} values for several pocket detection programs. The vertical axis designates the number of grid points in each of the pockets found in the dataset. The width of grid was 1 Å. The red and magenta lines are the distribution of pockets defined by the SURFNET and PASS programs, respectively. The green, magenta, and cyan lines correspond to the distributions of pockets determined by our program using 4, 6, and 8 Å large probe spheres, respectively.

tions of nongeometric approaches, such as the sequence conservation analysis and the template search for binding sites[1–4] have to be considered. Because our method only considers geometric properties of molecules, it has its own limitation. Geometric properties may be able to identify the location of binding sites on a protein, physical and chemical properties of protein surface is necessary for prediction of binding molecule type.

Selection of the appropriate R_{large} parameters also depends on users' required prediction accuracies. If users seek a high Precision of binding site prediction, we recommend using PHECOM with a lower R_{large} value such as 4 Å. If users seek a larger number of predicted binding site (a high Recall), we recommend using PHECOM program with a higher R_{large} value such as 8 Å.

The minimum radius of inaccessible surface R_{inaccess} was introduced to measure the shallowness of pockets.

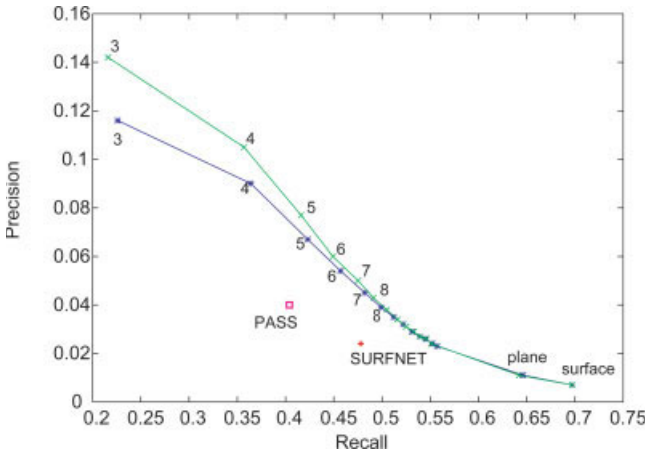


Fig. 15. A Recall-Precision plot for the recognition of ligands binding the pocket regions detected by the several programs. The red cross and the magenta square corresponds to the result of the SURFNET and the PASS programs, respectively. The blue line corresponds to the results of the PHECOM program. Figures, such as "3" and "4", designate the R_{large} parameter. The green lines corresponds to the result of the PHECOM program after removing pocket clusters whose M_{probe} values were less or equal to two.

Our concept of the minimum inaccessible radius R_{inaccess} is similar to the maximum contact radius proposed by Yeates.[28] The main difference between these concepts is that our calculation sought to characterize the probe spheres within pockets, while Yeates's method sought to characterize the protruding protein atoms. Another difference is that Yeates analytically calculate the maximum contact radius using the geometric inversion technique, whereas we estimate the approximate inaccessible radius by generating several sized spheres and employing the lower bound of inaccessible sphere. Thus, our minimum inaccessible radius should be slightly larger than Yeates's maximum contact radius. The R_{inaccess} measure can be used in various ways because it can be calculated for any binding object, including pocket probes, grid points, or ligand atoms that have been determined by experiments or docking calculation. It may be useful to compare environments of binding sites, and estimate stabilities of binding molecules.

Comparison of PHECOM to two other pocket finding programs demonstrated that the PHECOM program had a higher performance to identify binding sites, although its computational timer is a little longer. The PHECOM program can provide useful information for function prediction of protein with unknown function and finding candidate binding sites for docking calculation. We now plan to distribute our PHECOM program as a freeware.

ACKNOWLEDGMENTS

We thank Dr. Gautam Basu for his helpful comments. We also thank Mr. Hiromichi Yamagiwa, who used the beta version of the PHECOM program for his Masters thesis concerning the active sites of enzymes, for reporting prob-

lems and bugs in the developing version. We also thank Drs. Clara Shionyu-Mitsuyama and Kengo Kinoshita for their testing of the PHECOM program in their studies of sugar binding sites and providing us with helpful advices.

REFERENCES

1. Thornton JM, Todd AE, Milburn D, Borkakoti N, Orengo CA. From structure to function: approaches and limitations. Nat Struct Biol 2000;7 Suppl:991–994.
2. Yakunin AF, Yee AA, Savchenko A, Edwards AM, Arrowsmith CH. Structural proteomics: a tool for genome annotation. Curr Opin Chem Biol 2003;8:42–48.
3. Kim SH, Shin DH, Choi IG, Schulze-Gahmen U, Chen S, Kim R. Structural-based functional inference in structural genomics. J Struct Funct Genomics 2003;4:129–135.
4. Laskowski RA, Watson D, Thornton JM. From protein structure to biochemical function? J Struct Funct Genomics 2003;4:167–177.
5. Voet D, Voet JG. Biochemistry, 3rd ed. New York: Wiley; 2004. 460p.
6. Petsko GA, Ringe D. Protein structure and function. London: New Science Press; 2004. 56p.
7. Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE. A geometric approach to macromolecule-ligand interactions. J Mol Biol 1982;161:269–288.
8. Delaney JS. Finding and filling protein cavities using cellular logic operations. J Mol Graphics 1992;10:174–177.
9. Levitt DG, Banaszak LJ. POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. J Mol Graph 1992;10:229–234.
10. Laskowski RA. SURFNET: a program for visualising molecular surfaces, cavities, and intermolecular interactions. J Mol Graph 1995;13:323–330.
11. Masuya M, Doi J. Detection and geometric modeling of molecular surfaces and cavities using digital mathematical morphological operations. J Mol Graph 1995;13:331–336.
12. Peters KP, Fauck J, Frommel C. The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. J Mol Biol 1996;256:201–213.
13. Ruppert J, Welch W, Jain AN. Automatic identification and representation of protein binding sites for molecular docking. Protein Sci 1997;6:524–533.
14. Hendlich M, Rippmann F, Barnickel G. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. J Mol Graph Model 1997;15:359–363.
15. Liang J, Edelsbrunner H, Woodward C. Anatomy of protein pockets and cavities: measurement of binding site geometry and implication for ligand design. Protein Sci 1998;7:1884–1897.
16. Brady P, Stouten PFW. Fast prediction and visualization of protein binding pockets with PASS. J Comput Aided Mol Des 2000; 14:383–401.
17. Venkatachalam CM, Jiang X, Oldfield T, Waldman M. LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. J Mol Graph Model 2003;21:289–307.
18. Laurie ATR, Jackson RM. Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. Bioinformatics 2005;21:1908–1916.
19. Laskowski RA, Luscombe NM, Swindells MB, Thornton JM. Protein clefts in molecular recognition and function. Protein Sci 1996; 5:2438–2452.
20. Binkowski TA, Adamian L, Liang J. Inferring functional relationships of proteins from local sequence and spatial surface patterns. J Mol Biol 2003;332:505–526.
21. Bartlett GJ, Porter CT, Borkakoti N, Thornton JM. Analysis of catalytic residues in enzyme active sites. J Mol Biol 2002;324: 105–121.
22. Gutteridge A, Bartlett GJ, Thornton JM. Using a neural network and spatial clustering to predict the location of active sites in enzymes. J Mol Biol 2003;330:719–734.
23. Ota M, Kinoshita K, Nishikawa K. Prediction of catalytic residues in enzymes based on known tertiary structure, stability profile, and sequence conservation. J Mol Biol 2003;327:1053–1064.
24. Schmitt S, Kuhn D, Klebe G. A new method to detect related function among proteins independent of sequence and fold homology. J Mol Biol 2002;323:387–406.
25. Connolly ML. Analytical molecular surface calculation. J Appl Crystallogr 1983;16:548–558.
26. Connolly ML. Solvent-accessible surfaces of proteins and nucleic acids. Science 1983;221:709–713.
27. O'Rourke . Computational geometry in C. Cambridge: Cambridge University Press; 1994. 101p.
28. Yeates TO. Algorithms for evaluating the long-range accessibility of protein surfaces. J Mol Biol 1995;249:804–815.
29. Chothia C. The Nature of the accessible and buried surfaces in proteins. J Mol Biol 1976;105:1–14.
30. Andreeva A, Howorth D, Brenner SE, Hubbard TJP, Chothia C, Murzin AG. SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res 2004;32 (Database Issue):D226–D229.

APPENDIX

Heuristic Algorithm for Calculating the Position of Probe-Spheres Tangential to Three Protein Atoms, So as Not to Be Overlapped With Other Atoms

A critical part of our algorithm is the placement of a sphere to be tangential to three protein atoms, without overlapping with any other protein atoms. We call this problem the "three-contacting sphere" problem. Connolly[25] and Brady and Stouten[16] provided a mathematical solution to this problem by calculating the position of a sphere of a given radius that was tangential to three spheres whose radii are not necessarily equal (see Appendix A in Brady and Stouten[16]). The simply implemented algorithm, however, requires $O(N^4)$ computational time, where N is the number of protein atoms. There are $O(N^3)$ combinations of three protein atoms, each of which takes $O(N)$ time to check for crashes with other protein atoms. If the radius of a probe sphere is sufficiently small (about 1.4 to 1.8 Å), the number of three atoms combinations can be restricted by considering only the neighboring atoms, making the number of combinations that need to be examined $O(N)$. This assumption makes the actual computational time for placing a small sphere $O(N^2)$. This allows the Connolly surface and probes of PASS to be calculated in a reasonably short time. Placement of much larger probe spheres (6 Å or more), however, are necessary for our program. Unfortunately, combinations of three protein atoms for large probe spheres cannot be efficiently reduced by considering only neighboring atoms. The simple algorithm for large probe spheres requires too much computational time; the computational cost remains in the range of $O(N^4)$. Therefore, we invented a more efficient heuristic algorithm for the three-contacting sphere problem in much shorter time, based on the algorithms of the convex hull of points.

The three-contacting sphere problem is addressed by calculating only a subset of the molecular surface,[25,26] which is composed of three types of surfaces: spherical triangles, saddle-shaped rectangles, and spherical regions. The three-contacting sphere problem can be addressed by calculating only the spherical triangles, which correspond to probe spheres tangential to three protein atoms. If the

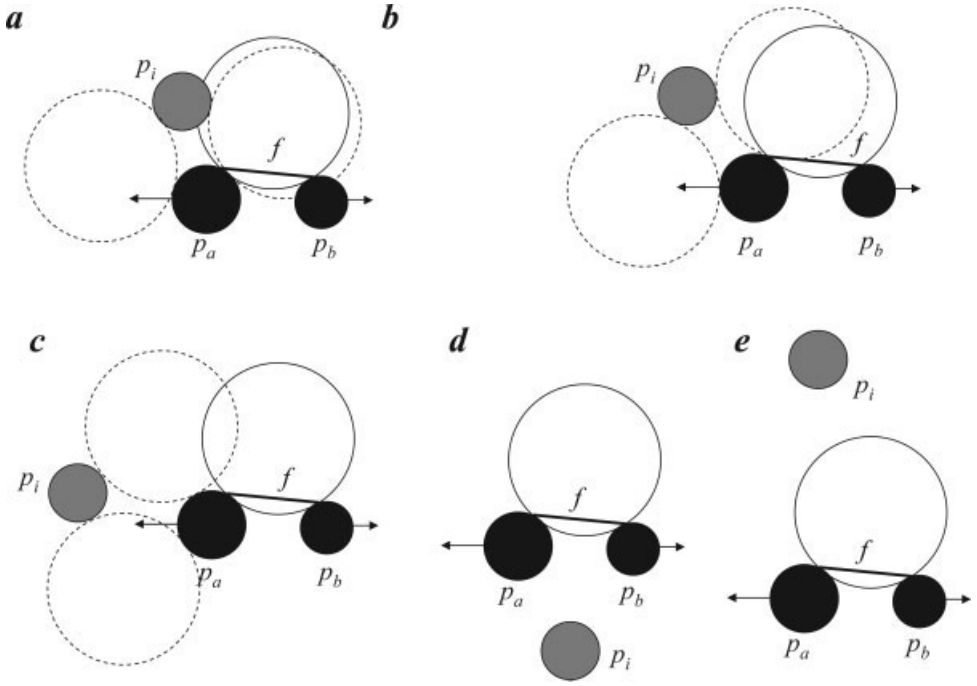


Fig. A1. Five cases demonstrating the addition of new protein atoms to the previous hull, in our incremental algorithm for three-contacting spheres. The arrangements of the protein atoms and probe spheres are schematically described in the two-dimensional space. A gray-filled circle is an atom to be added (p_i), while the black-filled circles are the atoms of previous hull (p_a, p_b). An empty circle surrounded by a solid line is the probe of the previous hull that is tangential to the atom p_a and p_b. An empty circle surrounded by a dotted line is the new probe that is tangential to the atom p_i and p_a. The line between p_a and p_b is called the face *f*. (**a**) *Collided*: The new atom p_i collided with the probe of the previous hull, allowing a new probe tangential to p_i and p_a to be generated. (**b**) *Visible*: The new atom p_i is above the spherical face *f*, allowing a new probe tangential to p_i and p_a to be generated. (***c***) *Outside*: The new atom p_i is outside of the normal vector of atom p_a, allowing the generation of a new probe tangential to p_i and p_a. (**d**) *Inside:* The new atom p_i is inside of the face *f*. (***e***) *Isolated:* The new atom p_i is far from the atoms p_a and p_b; thus, a new probe tangential to *pi* and atoms p_a or p_b cannot be generated.

radius of the probes placed on the protein surface is increased, the molecular surface finally converges to a convex hull, which is the smallest convex shape containing the given spheres. A convex hull of a set of spheres is composed of three parts: triangles, convex cylindrical rectangles, and convex spherical regions. If the radii of the protein atom spheres are sufficiently small, the convex hull and molecular surface can be approximated as a set of triangles or spherical triangles, ignoring the other parts. An algorithm for a convex hull of a set of points is much simpler than that of spheres.[27] Avoiding unnecessary programming complexity, we developed an algorithm to calculate the spheres tangential to three protein atoms, by modifying the incremental algorithm for convex hull of a set of points.[27] While this algorithm does not explicitly calculate the nontriangle parts of the molecular surface, the normal vector of protein atoms were introduced for to consider the convex spherical parts. As this is an approximating algorithm, we empirically confirmed that the results did not differ significantly from a slower, rigorous algorithm. The outline of the algorithm is described in the following text.

Algorithm: INCREMENTAL ALGORITHM FOR THREE-CONTACT SPHERES

INPUT: A set P of atoms in 3D space, whose radii are not necessarily equal. $P = (p_1, p_2 \ldots, p_n)$. The order of the atoms is given by their distance from their center of gravity.

OUTPUT: A list H_n containing the spherical triangle face *f,* which corresponds to a probe sphere with a given radius of R that is tangential to three atoms in P, and does not overlapped with any other atoms in P. Each face is uniquely identified by three set of atoms (p_i,p_j,p_k) in P, whose order is decided so as to be counterclockwise from outside of the face. S_n is the set of atoms for which the faces in H_n are tangential.

Initialize H_3 to two faces $f_1 = (p_1,p_2,p_3)$ and $f_2 = (p_1,p_3,p_2)$.

Initialize S_3 to (p_1, p_2, p_3).

for $i = 3, \ldots. n$ do

Initialize *NewFace* to empty.

for each spherical triangle face f of H_{i-1} do

Suppose a spherical triangle face f is tangent to three atoms (p_a,p_b,p_c).

Check collision *col* between the new atom p_i and the probe sphere of face f

Check visibility *vis* of the spherical triangle face f from the new atom p_i.

Check outside *out* of the new atom p_i for normal vectors of the atoms p_a or p_b or p_c.

If (*col* is *true*) or (*visible* is *true*) or (*out* is *true*),

then add six new spherical triangle faces (p_i,p_a,p_b), (p_i,p_b,p_a), (p_i,p_b,p_c), (p_i,p_c,p_b), (p_i,p_c,p_a), and (p_i,p_a,p_c) to the list *NewFace*.

If (col is true)

then delete f from H_{i-1}.

If (*col* is *false*) and (*visible* is *false*) and (*out* is *false*),

then move the atom p_i to the end of the list P.

$H_i = H_{i-1}$

for each spherical triangle face f of *NewFace* do

If the probe sphere for f does not collide with any atoms stored in S_{i-1},

then add f to H_i.

Update S as to contain all of the contacted atoms of f in H_i.

Update normal vectors for atoms using $p_1, p_2, \ldots, p_i$.

Check crashes of the probe spheres in H_n with all of the atoms in P. Crashed probe spheres are then removed from H_n.

When a new protein atom is added, we can classify the new situation as *collision*, *visible*, *outside*, or *inside*, as shown in Figure A1. A normal vector for each atom is defined such that for each atom p_i, a neighboring atom set Q_i is defined.

$$\boldsymbol{n}(p_i) = \boldsymbol{c}(p_i) - \frac{1}{|Q_i|}\sum_{q \in Q_i} \boldsymbol{c}(q)$$

The distance between the center of neighboring atom q and atom p_i should be less than the sum of both the radius and the diameter of probes.

$$|\boldsymbol{c}(p_i) - \boldsymbol{c}(q)| \leq r(p_i) + r(q) + 2R$$

This algorithm is considered to be as an approximated incremental algorithm of the molecular surface, as it explicitly considers only spherical triangles, not the remaining two surface parts, saddle-shaped rectangles and spherical regions. These two parts, however, are implicitly considered in the procedure for checking visibility and accessibility from the outside. The final procedure for checking crashes of generated probe spheres in H_n with every atom in P is still required to correct errors due to these approximations. We confirmed that the result of this algorithm was consistent with that of the naïve $O(N^4)$ algorithm; both results were identical for several small proteins using a probe sphere with a radius larger than 1.87 Å. From the numerical experiment, the computational cost of this algorithm is estimated at $O(N^2)$, which is significantly improved from the naïve algorithm.