A new approach to the automatic identification of candidates for ligand receptor sites in proteins: (I) Search for pocket regions

Carlos A. Del Carpio, Yoshimasa Takahashi and Shin-ichi Sasaki

Department of Knowledge-Based Information Engineering, Toyohashi University of Technology, Tempaku-cho, Toyohashi, Japan

The work presented here is aimed at the topographical analysis of localized regions of receptor proteins leading to the identification of pocket areas (superficial depressions or internal cavities), which may play the role of receptor sites. An algorithm is described that yields complete information about the position of each cavity or superficial depression relative to any point of the protein molecules, as well as detailed information on the atoms constituting it.

The applicability of this algorithm to the automatic identification of candidate receptor sites in a receptor protein is also discussed using the typical receptor structure dihydrofolate reductase-methotrexate complex.

Keywords: receptor site, receptor modeling, X-ray data, free surface point, solvent-accessible surface, protein surface depressions

INTRODUCTION

Recent developments in X-ray analysis of the three-dimensional structures of proteins and the consequent enlargement of protein data banks, as well as the evolution of programs for the prediction of proteins' higher structures, have enhanced not only a deeper understanding of the principles involved in the determination of the structures of these macromolecules but also the utilization of that information in a broad spectrum of research fields that require it. One of the active fields of research that requires information on the morphological characteristics of proteins is the design of drugs that act directly on specific regions of the protein.

The present paper is the first of a series of articles aimed at describing the investigations carried out in our laboratories aimed at the extraction and analysis of the constitutional as well as physicochemical properties of localized regions of these macromolecules.

Many studies have undertaken the analysis of drug receptor sites in proteins. These range from direct approaches, such as molecular docking,[1-3] to indirect approaches, such as the analysis of a set of related ligand molecules to map the receptor site.[4,5] Although processes such as those mentioned yield an estimate of the surroundings which a ligand molecule senses during its interaction with the macromolecule, none of them gives a full description of the real conditions to which the molecule is constrained. These conditions include not only the morphological environment of the ligand molecule but also the physical and physicochemical conditions to which the ligand molecule might be exposed. Furthermore, the studies cited above had as *a priori* information a receptor protein whose receptor sites were known, or utilized information about many molecules that interact as ligand molecules with specific receptor sites. Consequently, most of the procedures were directed to supply additional information about existent receptor sites. Such information is certainly valuable in processes involving the design of ligand molecules with shapes and characteristics similar to those of the original ligand molecules.

The approach can be extended to the analysis and design of probable ligand molecules that could act on sites of the macromolecule having morphological and physicochemical characteristics of receptor sites. The first step toward a process of this nature would be the analysis of the structural information of the macromolecule, leading to the identification of surface depressions or internal cavities that could be considered candidate receptor sites. The computation of the physicochemical properties of the constituents of each receptor candidate would then allow a global evaluation of the site, thus allowing definitive conclusions over the real character of the cavity and its consideration as a receptor site relevant to drug design processes. Furthermore, a long-term objective would be the evaluation of the activity of each candidate to find relations between the function of the protein and the interaction of a drug in a candidate receptor site of that nature.

Color Plates for this article are on page 42.

Address reprint requests to Prof. Takahashi at the Dept. of Knowledge-Based Information Engineering, Toyohashi University of Technology, Tempaku-cho, Toyohashi 441, Japan.

Received 28 April 1992; accepted 12 May 1992

This initial study has as its objective the automatic identification and extraction, using geometric considerations, of all the cavities and surface depressions of a protein molecule that have the characteristics of a receptor site. The structural information used is extracted from protein structural data banks.

There are two considerations that led to the conception of a new algorithm for the identification of the surface depressions and inner cavities of protein macromolecules. First, an analysis of the physical as well as physicochemical characteristics of a receptor site, such as the one proposed in the present work, requires the detailed enumeration of all the atomic components of the candidate receptor sites. Docking algorithms aim at reproducing the morphology of the cavity by positioning a probe in the vicinity of a macromolecular region. However, a detailed description of the components of the cavity thus generated would require the analysis of the free surface of the region to assign sets of points to the atomic components. Furthermore, directions for that mapping process would have to be established to achieve a complete map of the region and to find the location of the site within the macromolecule. To perform this, one of the problems one would have to confront is the handling of an enormous number of points to allow the correct mapping of the entire free surface of the site.

The second consideration is concerned with the concept of solvent-accessible surface, as established by Lee et al.,[6] and with which the computations carried out here are to some extent related. Thus, though the computation of points on the free surfaces of protein atoms are carried out in an analogous manner to that of the computation of the solvent-accessible surface, fundamental changes, such as that of the radius of the solvent, were necessary to accomplish the purpose of identifying automatically the cavities and surface depressions of the macromolecule. This is done independently of the graphics methods used to display the molecules.

Furthermore, the algorithm presented here yields complete information about the position of each cavity or surface depression relative to any point of the molecule, as well as detailed information on the atoms constituting it. The latter is necessary to perform the corresponding evaluations of the physicochemical characteristics of each of the identified cavities or surface depressions.

THE ALGORITHM

The initial stage in the identification of the surface depressions and cavities within a protein macromolecule requires the evaluation of free points on the surfaces of atoms constituting the molecule. These free points located on atoms of the molecular surface or deeper inside the molecule must first be retrieved. Any free point on the surface of an atom must not be occluded by another atom, and must be able to enter in contact with a probe of 1.30-Å radius, which was found to be the optimal value for the probe size.

Figure 1 shows schematically this definition. An atomic surface, however, is constituted by an infinite number of points, and the enumeration and handling of all of them would be an impossible task. Consequently, an approximate representation of the atomic surface by a limited number of points must be used. The accuracy of the present method,

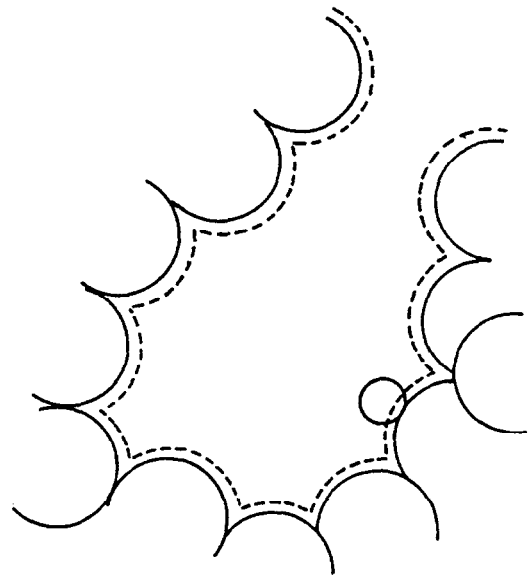

Figure 1. Conception of the protein atoms' free surface points; probe size: 1.30 Å.

however, depends heavily on the number of superficial points and the interpoint distances used in the process, thus requiring a careful selection of the number of points used to represent the atomic surface. One approach involves the assumption that an atom is a spherical body with a radius equivalent to the van der Waals radius of the particular atom. The spherical surface can be approximated by the points obtained by inscribing a polyhedron in the sphere. As already mentioned the number of points used in the representation is of critical importance in the process, so the selection of the appropriate polyhedron is also important. The homogeneity of the distribution of the representative points on the spherical surface limits the number of polyhedrons that can be selected for the purpose. A tetrahedron is the first polyhedron that accomplishes this condition; nevertheless, the reduced number of points representing the surface removes it from consideration. A larger number of superficial representative points is obtained when an icosahedron is inscribed within the sphere, though it is still insufficient for the present process. Figure 2a illustrates the representation of an atom as an icosahedral body. An adequate number of points is obtained when each face of the icosahedron is divided as shown in Figure 2b. The number of points obtained in this way equals those proposed by Lee et al.[6] for the calculation of the solvent-accessible surface. The coordinates of all the superficial points so obtained are directly calculated from the coordinates of the atoms. The radius of each sphere is the van der Waals radius plus the radius of the probe mentioned above. The results are not affected to a large extent when hydrogen atoms are considered to be part of the atom to which they are linked. Consequently, the van der Waals radius of each atom was increased by the addition of a hydrogen atom radius whenever there existed a bond between them. Table 1 lists the van der Waals radii utilized in the process.

Free points on the atomic surfaces are those that are not occluded by other atoms. The computation process of occluded and free surface points is illustrated in Figure 3. An

occluded point is one whose distance from the center of all the other neighboring atoms is less than their respective radii. In Figure 3, *P1* is occluded while *P2* is a free surface point.

The algorithm begins with the computation of all the free points on the atoms constituting the macromolecule, and the elimination of all other superficial points that are occluded.

A cavity can be thought of as a closed continuous surface of free points in the interior of the molecule, while a surface depression is a continuous concave surface of free points, which may be a channel between an internal cavity and the outside of the molecule.

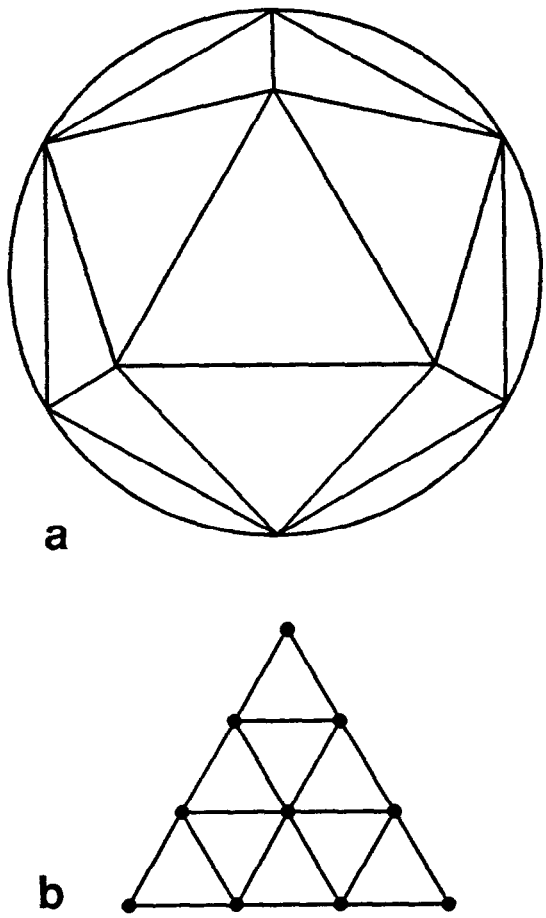

Figure 2. (a) Representation of spherical atoms as icosahedral bodies. (b) Division of each face of the icosahedral into 9 equilateral triangles leads to 7 more free surface points.

The algorithm introduced here identifies both types of sites. The process consists of growing the free surface in a particular region of the macromolecule by connecting free surface points that are separated by a determined distance. This distance is within a range that is explained later. The process to identify any type of site begins at the atomic free surface point that is closest to the center of gravity of the molecule. Therefore, the process is simplified if the coordinates of the points are expressed relative to the center of gravity of the protein molecule.

Figure 4 illustrates schematically the process in both cases, i.e., that of a completely buried cavity and that of the protein surface depression.

The growing process is constrained by three geometrical conditions. The first is the instantaneous distance between points when computing the next free atomic surface point. As shown in Figure 5, the maximal distance allowed between a pair of free surface points is 1.40 Å. This is the distance that, combined with the radius of the probe referred to earlier, was obtained by repetitive experiments carried out to identify existent protein cavities using this method. This pair of constants keeps the surface growing smoothly in the direction of concavity of the region, without sudden deviations to unexpected regions that do not belong to the concave surface.

The distance restriction is complemented by a geometrical constraint as illustrated in Figure 6. This constraint establishes that the next point to be added to the growing surface must be at a maximum distance of 1.40 Å from the last point, and also, the normal to the surface of the atom to which it belongs must be within an angle not greater than 90° relative to the normal of the last atom. This is to avoid the selection of

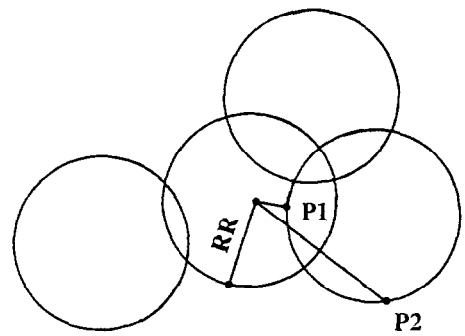

Figure 3. Exclusion of occluded atomic surface points: $\overline{\mathbf{P}_{(x,y,z)}\mathbf{C}} <$ RR ⟶ occluded point; $\overline{\mathbf{P}_{(x,y,z)}\mathbf{C}} >$ RR ⟶ nonoccluded point; $\mathbf{P}_{(1)}$ ⟶ occluded; $\mathbf{P}_{(2)}$ ⟶ nonoccluded.

Table 1. Van der Waals radii for the atoms constituting the amino acids

Atoms	Radius (Å)
All nitrogen: —N—, —NH—, $-NH_2$, $-NH_3^+$	1.50
All oxygen: =O, —O—, —OH	1.40
All sulfur: —S—, —SH	1.85
Nonaromatic carbon: >CH—, $-CH_2-$, $-CH_3$	2.00
Aromatic carbon: =CH—, =CH<	1.85
Carbonyl and all other carbons	1.50

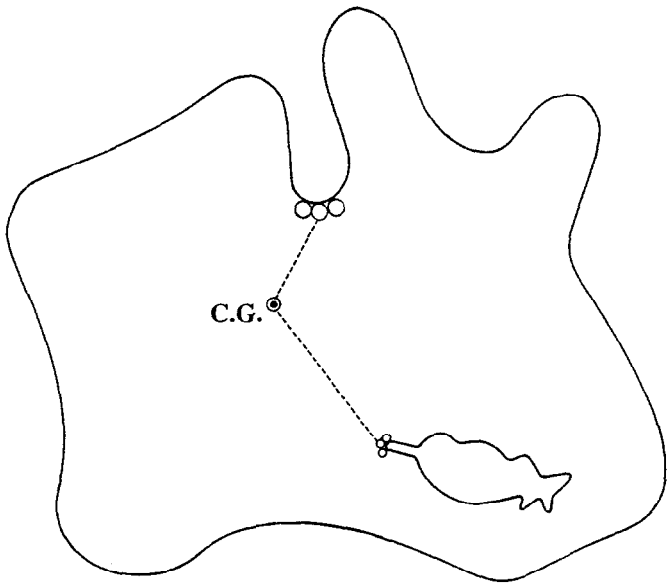

Figure 4. Search for the pocket regions begins at the free surface points nearest the center of gravity of the protein molecule (C.G.).

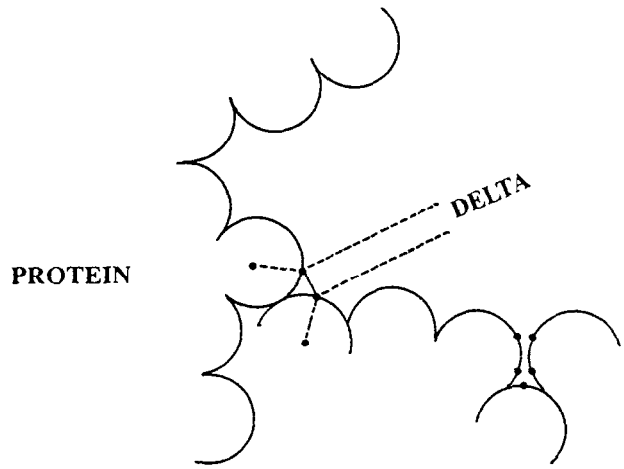

Figure 5. Local restriction to grow or expand the free surface; DELTA = 1.40 Å.

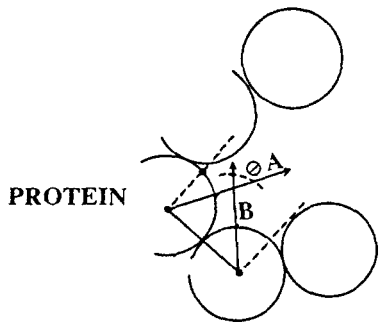

Figure 6. Local restriction to grow or expand the free surface: $\mathbf{A} \cdot \mathbf{B} = AB \cos \Theta > 0$.

points that are on opposite sides of the same atom, which can belong to two distinct free surfaces. Although this is an extreme case, the reduced radius of the probe adopted here allows for its frequent occurrence. Figure 6 illustrates graphically this situation. This restriction is also designed to avoid the selection of a point located on a different atom that meets the distance restriction, but is on the opposite side of the growing surface.

This restriction is expressed mathematically by the relation:

$$\mathbf{A} \cdot \mathbf{B} = AB \cos \Theta > 0 \tag{1}$$

The third is a rather long-range restriction, i.e., for atoms located at large distances from the starting point. It is designed to restrict the broadening of a cavity that is connected to the outside of the molecule. All the atoms constituting the retrieved site must lie on the same semicircle to which the initial point of the cavity belongs.

This restriction can be expressed mathematically by the following relation:

$$\mathbf{A} \cdot \mathbf{B} = AB \cos \phi > 0 \tag{2}$$

and is graphically illustrated in Figure 7.

The algorithm is designed to identify all the cavity-shaped regions of the protein. Consequently, all the free points belonging to a determined site are flagged and stored in memory together with the atoms to which they belong. When the identification of one cavity is finished, the system performs a search for the next unflagged point closest to the center of gravity.

In addition to the third geometrical restriction explained above, a parameter expressing the shallowness of the surface depression is required. In other words, it might be desirable to set the exhaustiveness of the identification process in terms of the shallowest admissible surface depression. The system is given a parameter setting the maximal distance from the center of gravity at which the starting free point must be. Figure 8 illustrates the process. A maximum distance equal to 75% of the distance of the farthest atom from the center of gravity of the molecule yields results within expectations. Moreover, the free surface growing process can also be halted when the distances of the last added free surface points have attained a determined distance from the center of gravity of

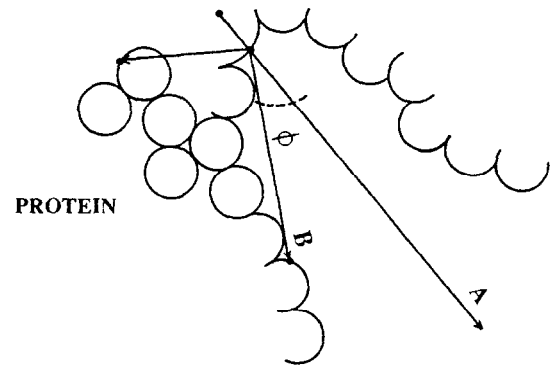

Figure 7. Long range restriction to grow or expand the free surface: $\mathbf{A} \cdot \mathbf{B} = AB \cos \phi > 0$.

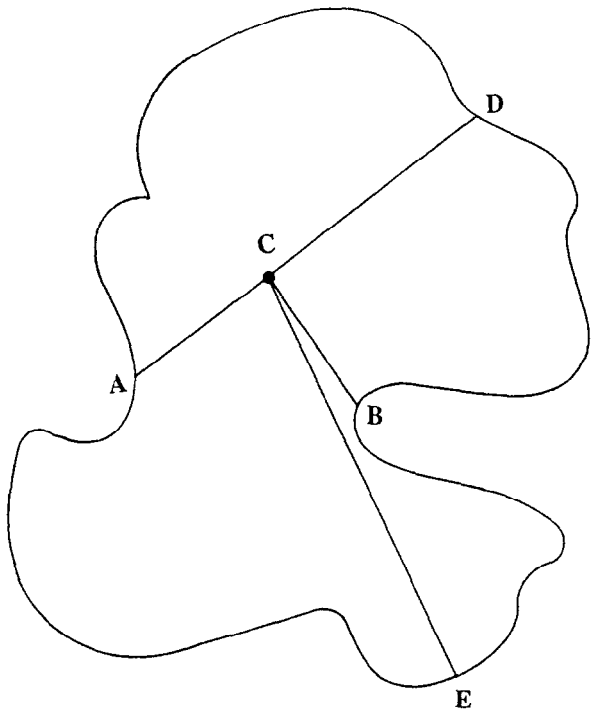

Figure 8. Points (A and B) at superficial depressions (within 75% of distance CF); CF is the distance of the farthest atom to the center of gravity of the molecule.

the molecule (Figure 8). A default distance was set in the study presented here that is equal to the distance of the farthest surface atomic point from the center of gravity of the protein.

RESULTS AND DISCUSSION

The algorithm was applied to the identification of cavities of proteins carrying a ligand molecule. This was done to focus the discussion, and to a certain extent, to examine the effectiveness of the algorithm.

The first molecule to which the algorithm was applied was the complex dihydrofolate reductase (E.C.1.5.1.3), containing NADPH and methotrexate. The structural data was extracted from the Protein Data Bank (PDB).

The search starts at the free point of atom 27 that is the closest to the center of gravity, then with an inter-free point distance set at 1.40 Å, points constituting the free surface of atom 27 are added to the starting one. The process grows, taking into account points of atom 219 belonging to the 27th residue, which is leucine. The set of parameters: 1.30 Å for the radius of the probe and 1.40 Å for the interpoint distance yield the results summarized in Table 2. Here, within 234 seconds of processing time on a DG AV6000 minicomputer, 20 cavities and depressions were identified. Two of them, Site 1 and Site 20, are the pockets containing the two ligand molecules, i.e., the NADPH and the methotrexate, respectively.

The data presented here were compared with the results given by a search of the nearest neighbor atoms carried out over all the atoms constituting both ligands. The maximum distance for one atom to be considered a close neighbor of another was set to 6 Å. The results are partially presented in Table 3. The comparison, though biased by the distance parameters established in each process, allows the global appreciation of the exhaustiveness of the search of the algorithm presented here.

Considering a larger value for the radius of the probe forces an increase in the interpoint distance parameter. Experiments were carried out on the same complex to identify the cavities with larger parameters. Tables 4 and 5 show the results of the search for the given parameters values. As shown, increasing

Table 2. Results for dihydrofolate reductase(E.C.1.5.1.3) complex with nadph and methotrexate: probe radius = 1.30 Å, interpoint distance = 1.40 Å, No. of cavities: 20

Atom	Z	At.	Res. No.	Residue
Cavity No. 1 Number of atoms: 105				
27	8	O	4	LEU
783	8	O	97	ALA
33	6	CA	5	TRP
251	6	CE2	30	PHE
46	7	N	6	ALA
367	6	CG2	45	THR
824	6	CE2	103	PHE
790	6	CA	99	GLY
50	6	CB	6	ALA
36	6	CB	5	TRP
103	6	CB	13	ILE
403	6	CE1	49	PHE
34	6	C	5	TRP
210	8	OD1	26	ASP
Cavity No. 6 Number of atoms: 3				
149	6	CD2	19	LEU
102	8	O	13	ILE
49	8	O	6	ALA
Cavity No. 19 Number of atoms: 26				
910	6	CD1	114	LEU
303	6	CE	37	LYS
756	8	OE2	93	GLU
276	8	OE1	33	GLN
304	7	NZ	37	LYS
12	6	CB	2	ALA
Cavity No. 20 Number of atoms: 124				
108	6	CA	14	GLY
996	6	CB	126	THR
109	6	C	14	GLY
128	7	N	17	GLY
997	8	OG1	126	THR
132	7	N	18	HIS
990	8	OD1	125	ASP
989	6	CG	125	ASP
991	8	OD2	125	ASP
987	8	O	125	ASP

Table 3. List of atoms that belong to the set of methotrexate nearest neighbors

Atom	Z	At.	Res. No.	Residue	Dist.*
27	8	O	4	LEU	5.79
783	8	O	97	ALA	5.58
33	6	CA	5	TRP	5.23
251	6	CE2	30	PHE	5.08
46	7	N	6	ALA	5.99
367	6	CG2	45	THR	4.85
824	6	CE2	103	PHE	6.00
790	6	CA	99	GLY	3.99
50	6	CB	6	ALA	5.92
36	6	CB	5	TRP	4.92
103	6	CB	13	ILE	5.10
403	6	CE1	49	PHE	5.37
34	6	C	5	TRP	5.80
210	8	OD1	26	ASP	5.99
789	7	N	99	GLY	3.48
336	6	CA	42	GLY	3.96

* Distance between a cavity component atom and the closest methotrexate atom (nearest neighbors within 6 Å).

the probe parameter has the effect of splitting the surface of a site into smaller regions (Table 4). This is due to the fact that an increase in the probe parameter leads to larger interpoint distances. Consequently the process of growing the free surface comes to a halt when the distance restriction is not met. This happens with distances between points that, before the increase, were separated by a distance that was within the allowed range of variation. The effects of the increase of the probe radius parameter could be balanced with an increase of the allowed interpoint distance, however, distances that counteract an increment of the radius of the probe lead to interpoint distances similar to the van der Waals radii of the atoms. Interpoint distances of this magnitude interfere with the orientation of the growing process, leading to the incorrect addition of points, and thus to atoms located in positions that do not correspond to the original free surface (Table 5).

The experiments performed using diverse complexes to establish a pair of parameters (the radius of the probe and the interpoint distance) that allowed a smooth growing process, and in the correct orientation, led consequently to the adoption of the parameters established at the beginning of the work.

The results output by the program are points of the free surface of each candidate receptor, the atoms to which they belong, and consequently, the residues to which they belong. However, a visual appreciation of each site is considered to be of the utmost importance. Computer graphics are the best method to obtain a physical appreciation of this situation. Points constituting the free surfaces of the cavities identified with the algorithm can be plotted on the screen of a computer terminal. The coordinates of the points forming the cavity however, are only approximately points of the surface of the atoms constituting the molecule because of the approximation of the atomic morphology. The display of the cavities' physical appearance by plotting the coordinates of the points computed by the algorithm are, however, hampered by this

Table 4. Results for dihydrofolate reductase(E.C.1.5.1.3) complex with nadph and methotrexate: probe radius = 1.40 Å, interpoint distance = 1.00 Å, No. of cavities: 78

Atom	Z	At.	Res. No.	Residue
Cavity No. 1 Number of atoms: 4				
783	8	O	97	ALA
148	6	CD1	19	LEU
33	6	CA	5	TRP
251	6	CE2	30	PHE
Cavity No. 2 Number of atoms: 2				
783	8	O	97	ALA
367	6	CG2	45	THR
Cavity No. 4 Number of atoms: 3				
783	8	O	97	ALA
148	6	CD1	19	LEU
403	6	CE1	49	PHE
Cavity No. 8 Number of atoms: 2				
36	6	CB	5	TRP
366	8	OG1	45	THR
Cavity No. 10 Number of atoms: 6				
46	7	N	6	ALA
108	6	CA	14	GLY
210	8	OD1	26	ASP
246	6	CB	30	PHE
219	6	CD2	27	LEU
148	6	CD1	19	LEU
Cavity No. 14 Number of atoms: 2				
34	6	C	5	TRP
219	6	CD2	27	LEU
Cavity No. 74 Number of atoms: 5				
790	6	CA	99	GLY
793	7	N	100	ALA
798	7	N	101	GLN
803	6	CG	101	GLN
359	7	NH1	44	ARG
Cavity No. 78 Number of atoms: 2				
910	6	CD1	114	LEU
303	6	CE	37	LYS

assumption. A reflection calculation of each point on the surface of the atom (when it is assumed to be a sphere of radius equal to the van der Waals radius) would lead to better results. However, besides the computer processing time spent in the calculation over all the points, the uncertainty would still remain.

The approach adopted here was an averaging of the distances of the surface points belonging to a determined orien-

Table 5. Results for dihydrofolate reductase(E.C.1.5.1.3) complex with nadph and methotrexate: probe radius = 1.4 Å, interpoint distance = 1.7 Å, No. of cavities: 6

Cavity No. 1 Number of atoms: 630				
Atom	Z	At.	Res. No.	Residue
783	8	O	97	ALA
27	8	O	4	LEU
33	6	CA	5	TRP
251	6	CE2	30	PHE
249	6	CD2	30	PHE
34	6	C	5	TRP
367	6	CG2	45	THR
403	6	CE1	49	PHE
50	6	CB	6	ALA
211	8	OD2	26	ASP
46	7	N	6	ALA
210	8	OD1	26	ASP
149	6	CD2	19	LEU
219	6	CD2	27	LEU
246	6	CB	30	PHE
824	6	CE2	103	PHE
247	6	CG	30	PHE
148	6	CD1	19	LEU
366	8	OG1	45	THR
36	6	CB	5	TRP
790	6	CA	99	GLY
250	6	CE1	30	PHE
252	6	CZ	30	PHE
133	6	CA	18	HIS
134	6	C	18	HIS
Cavity No. 2 Number of atoms: 1				
251	6	CE2	30	PHE
Cavity No. 5 Number of atoms: 1				
30	6	CD1	4	LEU
Cavity No. 6 Number of atoms: 1				
108	6	CA	14	GLY

tation with respect to the normal vector of the starting point of each cavity. The points are first divided in groups corresponding to zones, located at a determined distance from the starting point. Each zone is characterized by the distribution of points within that distance.

The first cavity obtained by the algorithm described here, and displayed by this procedure, is shown in Color Plate 1.

CONCLUSIONS

A rational approach to drug design requires sources of information on the nature of the environment where the object molecule might act. When the molecular environment belongs to a protein, then information on the particular region where it will probably bind is of the utmost importance. Furthermore it is known that the topographical characteristics of the protein molecular surface are intimately related with its function. Consequently, information on this topography and the process to obtain it led us to introduce a new algorithm for the identification of surface depressions and internal cavities that could, from the physical point of view, behave as receptor sites to possible ligand molecules. The process is geometrical, and no chemical knowledge has been utilized.

The consideration of a probe of a particular radius, as discussed in the previous section, was necessary to establish an interpoint distance that could maintain the growing process until a determined distance from the center of the molecule was achieved. The splitting of the cavity in the middle of the process of growing the free surface was a problem that arose from a probe radius that was set to values equivalent to those of solvent radii.

The situation is critical when the interpoint distance is less than 150% of the radius of the probe. As a consequence, probe radii equivalent to solvent radii would mean interpoint distances of the order of atomic van der Waals radii. Interpoint distances of this magnitude would lead to growing processes in unexpected directions. Thus the relevance of the radius of the probe and the interpoint distance is critical for the process.

As shown in Table 1, cavities of different sizes can be identified by the algorithm, and this is predictable because of the probe radius. The elimination of trivial pockets is, however, a simple process. The number of atomic components of each site are representative of the importance of the site.

Finally, the present algorithm is able to yield a relative orientation of the site within the macromolecule. It yields the components and the residues to which they belong. This is relevant when the respective physicochemical analysis of the site is to be performed. Therefore, the algorithm constitutes the departure point towards a global process of analysis of the characteristics of the topology of the macromolecule concerning its function and the features that are necessary for a process of automatic drug design.

REFERENCES

1 Ghose, A.K. and Crippen, G.M. *J. Med. Chem.* 1983, **26**, 996–1010
2 Crippen, G.M. *Quant. Struct. Act. Relat.* 1983, **2**, 95–100
3 Donne-Op den Keldler, G.M. *J. Computer-Aided Molecular Design* 1987, **1**, 257–264
4 Tintelot, M. and Andrews, P. *J. Computer-Aided Molecular Design* 1989, **3**, 67–84
5 Wendoloski, J.J., Wasserman, Z.R., and Salemme, F.R. *J. Computer-Aided Molecular Design* 1987, **1**, 313–322
6 Lee, L. and Richards, F.M. *J. Mol. Biol.* 1971, **55**, 379–400