ELSEVIER

LIGSITE: Automatic and efficient detection of potential small molecule-binding sites in proteins

Manfred Hendlich,* Friedrich Rippmann,† and Gerhard Barnickel†

**Department of Pharmaceutical Chemistry, University of Marburg, Marburg, Germany*
†Preclinical Research, Merck KGaA, Darmstadt, Germany

LIGSITE is a new program for the automatic and time-efficient detection of pockets on the surface of proteins that may act as binding sites for small molecule ligands. Pockets are identified with a series of simple operations on a cubic grid. Using a set of receptor–ligand complexes we show that LIGSITE is able to identify the binding sites of small molecule ligands with high precision. The main advantage of LIGSITE is its speed. Typical search times are in the range of 5 to 20 s for medium-sized proteins. LIGSITE is therefore well suited for identification of pockets in large sets of proteins (e.g., protein families) for comparative studies. For graphical display LIGSITE produces VRML representations of the protein–ligand complex and the binding site for display with a VRML viewer such as WebSpace from SGI. © 1998 by Elsevier Science Inc.

Keywords: binding sites, pockets, surface depressions, receptor–ligand complexes, protein ligands, VRML

INTRODUCTION

With the explosion of high-resolution protein structures deposited in the Brookhaven Protein Databank[1] (PDB), structure-based drug design seems more and more feasible (for reviews of protein-based molecular design, see Refs. 2–5). An impressive example of the successful application of structure-based design techniques is the development of novel human immunodeficiency virus 1 (HIV-1) protease inhibitors.[6–8]

A prerequisite for the docking of small molecule ligands is the determination of the site where the ligand interacts with the protein. Such binding sites of small molecule ligands are generally located in pockets (clefts, grooves) on the surface of proteins. The determination of pockets is therefore an important step toward the rational design and discovery of novel ligands. An in-depth analysis and classification of pockets on the surfaces of the known protein structures might also improve our understanding of the processes involved in ligand binding and selectivity.

A manual definition of binding sites is impractical for several reasons. In most cases it is difficult to define exactly where a pocket ends and free space begins. Proteins may have surface depressions of various sizes and a manual inspection may fail to find all relevant binding sites but the largest one. A manual definition is also impractical when large sets of protein structures must be processed (e.g., for statistical studies of pocket characteristics). Several attempts to automate the identification of surface depressions have been made over the years.[9–15]

A tool that has been widely used for detecting potential binding sites is the program POCKET developed by Levitt and Banaszak.[9] POCKET is well suited for identifying potential binding sites as it does not require any prior knowledge about the location of the binding site on the surface of the protein. Another appealing factor of the POCKET program is its algorithmic simplicity. The identification of pockets is achieved by a series of simple operations on a cubic grid.

The original implementation of the POCKET algorithm suffers from the fact that the correct recognition of pockets and of residues belonging to a pocket strongly depends on the orientation of the protein in the grid. Another drawback is the inefficient implementation of the algorithm, resulting in long execution times if small grid steps are used.

For this reason we decided to develop a new program named LIGSITE, based on an algorithm similar to that of the POCKET program but circumventing most of its drawbacks. Owing to a more rigorous scanning of the protein surface, the dependence of LIGSITE on correct orientation to recognize pockets (a drawback of most grid-based approaches) has been significantly reduced. Furthermore, LIGSITE is sufficiently fast to allow the processing of a large number of proteins,

Color Plate for this article is on page 389.

Address reprint requests to: M. Hendlich, Department of Pharmaceutical Chemistry, University of Marburg, Marbacher Weg 6, D-35032 Marburg, Germany.

Received 30 October 1996; revised 18 December 1997; accepted 6 January 1998.

655 Avenue of the Americas, New York, NY 10010
1093-3263/97/$17.00
PII S1093-3263(98)00002-3

which makes a statistical analysis and classification of pockets on protein surfaces possible.

DETERMINATION OF POTENTIAL BINDING SITES

Scanning for pockets and cavities

The first stage in the identification of pockets is similar to the algorithm of Levitt and Banaszak as used in their POCKET program. The program starts by generating a regular Cartesian grid. At the beginning all grid points are labelled as solvent and set to 0. Grid points that are inaccessible to solvent are assigned a value of -1. As a straightforward method for this steric validation, distance checks are used to see whether a solvent molecule centred at a grid point overlaps with any atom of the protein. To avoid this computationally expensive distance check for each lattice point against each protein atom, we use the "shadow approach" for calculating solvent-accessible grid points. Only grid points within a cube of size $2(r_{atom} + r_{solvent})$ centred at the position of a protein atom need to be checked using distance calculations. The variable r_{atom} represents the van der Waals radius of the corresponding atom; $r_{solvent}$ is the radius of the probe, which is set to 1.4 Å.

The accuracy with which the surface of pockets and cavities is determined depends on the step size that is used. At an infinitely small step size, the surface determined by this algorithm would be equivalent to the contact surface. A large grid step size (Levitt and Banaszak used a step size of 2.0 Å) decreases calculation time but results in a very angular surface. In our program the grid size is freely adjustable. In practice, step sizes between 0.5 and 0.75 Å provide smooth surfaces and acceptable response times.

To determine which grid points lie in a cavity or a pocket the program scans along the *x* axis for regions that are enclosed on both sides by grid points that are inaccessible to solvent (Figure 1). If an area of solvent-accessible grid points is enclosed on both sides by protein atoms, there is a high probability that these grid points are located in a pocket or a cavity. The value of these solvent-accessible grid points on the scanline is increased by 1. This *x*-axis scan is done for all *y* and *z* values. The same procedure is then repeated for *y*- and *z*-axis scans. In the following discussion the detection of such an arrangement (first protein, then solvent, then protein again) will be termed a *PSP event.*

In the original implementation of the POCKET program the correct recognition of the shape of pockets depends very much on the orientation of the protein in the grid. Pockets with an orientation of 45° to the *x, y,* and *z* axes are not recognised correctly, or perhaps not at all, as the scanning for pockets is only done in the *x, y,* and *z* directions (Figure 1). To reduce this orientational dependence LIGSITE also scans along the four cubic diagonals, which doubles the calculation time but enables the precise determination of the extent of pockets independent of the orientation of the protein in the grid.

An additional scanning along the diagonals in the *xy*, *xz*, and *yz* planes was also tested but showed no further improvement.

Recognition of pockets and cavities

After the scanning procedure all grid points accessible to solvent molecules have values between 0 (no PSP event and therefore not located in a pocket or cavity) and 7 (deeply buried with PSP events along the three axes and the four cubic diagonals). Pockets and cavities are now defined as regions of grid points with a minimum number of PSP events (MIN_PSP). This threshold is freely adjustable by the user. In practice a MIN_PSP value of 2 has yielded good results. LIGSITE detects pockets by starting with a grid point with a value ≥MIN_PSP. All nearest neighbours in the grid with values ≥MIN_PSP are added to the region. The newly added points are again checked for nearest neighbours with values ≥MIN_PSP. The process continues until all nearest neighbours with a value ≥MIN_PSP are added to the region. Any grid points left with values ≥MIN_PSP constitute one or more new pockets and for them the process is started again.

To provide more control over the size and extent of the reported pockets and cavities, two strategies are implemented.

1. Often it is difficult to determine exactly where a cavity ends and where free space begins. By adjusting the value of MIN_PSP the user is able to control the extent of the pocket on the surface of the protein. Setting MIN_PSP to low values results in large cavity surfaces. Setting MIN_PSP to high values limits the surface of cavities to more buried areas. A precise definition of pockets is especially important for docking studies, in which it is desirable to have a rather small number of potential interaction sites to reduce the computational complexity.
2. Most protein surfaces are rather irregular with a large number of small bumps. These bumps are not relevant as binding sites for ligands. To avoid the reporting of small pockets and cavities, the number of grid points must be larger than a predefined threshold. This threshold is freely adjustable. Threshold values between 20 and 100 have given good results.

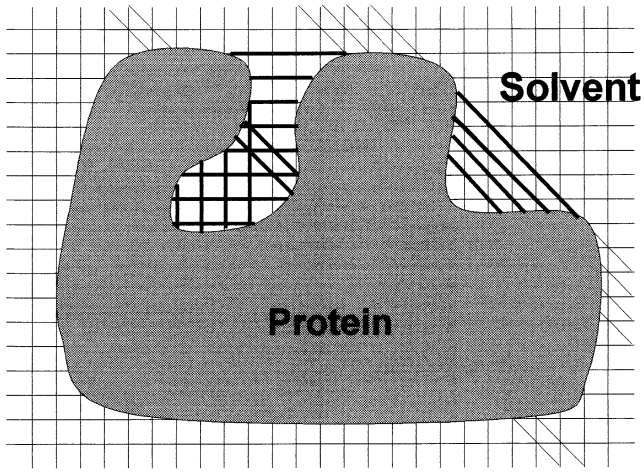

Figure 1. Schematic representation of the pocket-searching algorithm implemented in LIGSITE. Pockets are identified by scanning along the x, y, and z axes and the cubic diagonals (only four diagonals are shown) for areas that are enclosed on both sides by protein (indicated by boldface lines). If the scanning is done only along the x, y, and z axes (as implemented in the POCKET[9] program) the surface depression at upper right cannot be identified.

The distinction between pockets and cavities is straightforward. As cavities do not have connections to the solvent each grid point in a cavity must have a value of 7 (equivalent to the number of scanning directions) and all other nearest neighbour grid points must be inaccessible to solvent.

Determination of the surface of pockets and cavities

Up to this stage all grid points in pockets and cavities in which a water molecule could be placed without overlapping with the protein have been determined. Using these grid points for the determination of the protein–solvent boundary region would approximate a surface that is equivalent to the solvent-accessible surface. To obtain the molecular surface, all grid points within a sphere of radius $r_{solvent}$ centred at a grid point in a cavity that is marked as being accessible to solvent are marked as belonging to the cavity. To keep the calculation times low the preceding shadow algorithm is used again.

Once all grid points that belong to pockets and cavities are identified the determination of the surface is straightforward. The points that represent the surface of a pocket are grid points where at least one of the nearest neighbours in the grid is occupied by a protein atom. For every surface point the nearest neighbour surface points on the grid are stored for the subsequent graphical display. Once the surface of the pocket is determined, the identification of the amino acids and atoms that surround it is straightforward.

Output and display

The program can be run either in an interactive mode with a graphical user interface or in a standalone mode. In the interactive mode proteins and ligands are displayed as solid lines. The surface of pockets and cavities is displayed either as a gridwork of lines or as a solid surface. If charges are specified in the input file the program is able to calculate the electrostatic potential at the surface, using Coulomb's law. The electrostatic potential can be used to colour the surface. Furthermore, the user can selectively display only residues that border the pocket.

In the standalone mode two different output styles can be produced. In the first mode the program reports the number of pockets and the surrounding residues, which can be stored in a PDB-like format. This information can be used by other programs, e.g. docking programs.

The other non-interactive mode produces VRML code for display with a VRML-viewer such as WebSpace from Silicon Graphics. Pockets are drawn as a gridwork of lines. A display of pockets as solid surfaces in VRML is under development. Proteins and ligands can be drawn either as simple lines or in a balls and sticks representation.

Programming details

The program was implemented in ANSI-C on an SGI-Indigo2. The graphical user interface is based on the Tcl/Tk Toolkit. To take advantage of the high speed SGI-GL drawing routines a GLXwin widget is used as the main drawing window. As input the program reads standard PDB and MOL2 files. In the current version of LIGSITE HET-atoms in PDB files are ignored.

PERFORMANCE

In order to assess the ability to recognise binding sites a set of ten receptor/ligand complexes from the PDB was used (Table 1). In each case the program is able to identify the correct location of the binding site with high precision. Using a grid spacing of 0.5 Angstrom all protein atoms which are in contact with the ligand have been recognised (Table 1). This is an absolute prerequisite for the docking of ligands. Otherwise it would not be possible to predict the proper binding mode of a ligand as the correct interaction partners might not be available in the docking process.

As all small bumps and clefts which are not relevant for the binding of ligands are ignored only one pocket is identified in most cases. In all cases (1ADG, 1DR1, 1ASC) where more pockets are found the ligand binds in the largest pocket. This is also true for 1DR1 where both ligands bind in the same pocket which is correctly identified by LIGSITE.

A VRML representation of a complex (PDB-code 121P) between the RAS protein (shown in magenta) and the ligand GTO, a guanosin triphosphate derivate (shown in green) is shown in Color Plate 1a. The surface of the pocket is drawn in blue. The identified pocket comprises the binding site of GTO almost perfectly. All interaction partners on the protein are found. In contrast to the original POCKET program, LIGSITE produces smooth surfaces and the shapes of pockets show no dependence on the orientation of the protein in the grid.

The ability to influence the extent of pockets on the surface is demonstrated in Color Plate 1b, using the same complex between the Ras protein and GTO as before. Increasing MIN_PSP (the minimal value of protein–solvent–protein scans for grid points to be accounted as being located in a cavity) to 3 restricts the extent of the surface to more buried areas. As shallow regions are ignored, LIGSITE identifies two different pockets. The larger pocket matches the binding site of GTO almost completely. Such a restriction of pockets to the core region reduces the degree of freedom in computational docking and therefore improves the calculation speed considerably.

Another significant improvement of LIGSITE is its speed compared with the original POCKET program. Levitt and Banaszak reported an execution time of about 150 s for a protein with 133 residues and a grid spacing of 2.0 Å on an SGI 4D25. Taking into account the difference in CPU power, LIGSITE is, despite the more rigorous scanning of the protein, significantly faster for a protein of similar size and the same grid spacing of 2.0 Å (Table 2). This improvement in speed makes possible a statistical analysis and classification of pockets on the surface of proteins in the PDB. Such a study is currently in progress.

SUMMARY AND OUTLOOK

We have presented a program for the identification and graphical display of pockets and cavities in proteins that does not require any prior knowledge about their position. We have shown that the algorithm is able to detect the binding sites of ligands in 10 receptor–ligand complexes from the PDB with high precision. In each case all residues in contact with a ligand have been successfully identified. The algorithm can be seen as an extension of the POCKET program developed by Levitt and Banaszak.[9] Compared with the original implementation our approach provides several advantages. Owing to a more rigor-

Table 1. Performance of LIGSITE in identifying the correct binding sites of 10 different receptor–ligand complexes from the PDB

Protein (PDB code, number of residues)	Ligand (number of heavy atoms)	Pockets found	Percentage of contacting atoms found[a]	Time (s)
Myoglobin (1MBD, 153)	Heme (43)	1	100	7.36
Cholesterol oxidase (3COX, 507)	Flavin-adenine dinucleotide (53)	1	100	23.70
H-Ras p21 (121P, 166)	Guanosine-5′-[β,γ-methylene] triphosphate (32)	1	100	7.79
Alcohol dehydrogenase (1ADG, 374)	β-Methylene selenazole-4-carboxamide adenine dinucleotide (51)	2	100	18.19
Dihydrofolate reductase (1DR1, 189)	Biopterin (17), $NADP^+$ (48)	2	100	12.17
Ricin (1FMP, 267)	Formycin-monophosphate (23)	1	100	13.15
Aldose reductase (2ACS, 315)	Citric acid (13), $NADP^+$ (48)	1	100	14.29
Guanine nucleotide-binding protein (1GFI, 353)	Guanosine 5′-diphosphate (28)	1	100	17.30
Thymidylate synthase (1TYS, 265)	Dihydrofolate (32), thymidine 5′-monophosphate (21)	1	100	14.78
Aspartate aminotransferase (1ASC, 396)	*N*-Methyl-ca4-deoxyimino-pyridoxal-5-phosphate (17)	3	100	25.14

[a] Percentage of atoms found that are in contact with the ligand. A protein atom and a ligand atom are classified as being in contact if the distance is less than the sum of the van der Waals radii plus a tolerance of 0.5 Å. For all calculations we used a grid step size of 0.5 Å and a MIN_PSP value of 2. VRML files for all complexes can be downloaded from our WWW server (http://www.pharmazie.uni-marburg.de/fbpharmazie/pharmchemie/akklebe/ligsite.html).

ous scanning the dependence on orientation for the identification and shape of pockets, a major disadvantage of most grid-based approaches, is considerably reduced. The second advantage of this more rigorous scanning is the better discrimination of grid points with respect to their position in the pocket. This provides better control in determining the shape and extent of pockets.

Another major advantage of LIGSITE is its speed. A scanning of the complete surface of a medium-sized protein can be done in 5 to 10 s. This allows the use of small grid step sizes, which is a prerequisite for the determination of accurate, smooth surfaces and the processing of a large number of protein structures in a reasonable time. The scanning of all proteins in the PDB for a statistical analysis and classification of pockets is currently in progress. All identified pockets will be stored within the RELIBase[16,17] database system, which is currently being developed by our group for handling and analysing receptor–ligand complexes.

Table 2. Execution times for identifying pockets on the surface of the H-Ras p21 protein (PDB code 121P, 166 residues), using different grid sizes[a]

Grid size (Å)	Execution time (s)	Number of pockets found
2.0	0.11	1
1.0	0.89	1
0.75	2.07	1
0.66	3.15	1
0.5	7.79	1
0.25	95.82	1

[a] All calculations were performed on an SGI Indigo2 workstation with an R4400 processor running at 150 MHz.

ACKNOWLEDGMENTS

This work was supported by funds from the Bundesminister für Bildung und Forschung, Germany.

REFERENCES

1 Bernstein, F.C., Koetzle, T.F., Williams, G.J.D., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanochi, T., and Tasumi, M. The Protein Data Bank: A computer based archival file for macromolecular structures. *J. Mol. Biol.* 1977, **112,** 535–542

2 Kuntz, I.D., Meng, E.C., and Shoichet, B.K. Structure based molecular design. *Accounts Chem. Res.* 1994, **27,** 117–123

3 Whittle, P.J., and Blundell, T.L. Protein structure based drug design. *Annu. Rev. Biophys. Struct.* 1994, **23,** 349–375

4 Lybrand, T.P. Ligand–protein docking and rational drug design. *Curr. Opin. Struct. Biol.* 995, **5,** 224–228

5 Bamborough, P., and Cohen, F.E. Modelling protein–ligand complexes. *Curr. Opin. Struct. Biol.* 996, **6,** 236–241

6 Rutenber, E., Fauman, E.B., Keenan, R.J., Fong, S.,

Furth, P.S., de Montellano, P.R.O., Meng, E., Kuntz, I.D., DeCamp, D.L., Salto, R., Rose, J.R., Craik, C.S., and Stroud, R.M. Structure of non-peptide inhibitor complexed with HIV-1 protease. *J. Biol. Chem.* 1993, **268,** 15343–15346

7 Ghosh, A.K., Thompson, W.J., Fitzgerald, P.M.D., Culberson, J.C., Axel, M.G., McKee, S.P., Huff, J.R., and Anderson, P.S. Structure-based design of HIV-1 protease inhibitors: Replacements of two amides and a 10π-aromatic system by a fused bis-tetrahydrofuran. *J. Med. Chem.* 1994, **37,** 2506–2508

8 Lam, P.Y.S., Jadhev, P.K., Eyermann, C.J., Hodge, C.N., Ru, Y., Bacheler, L.T., Meek, J.L., Otto, M.J., Rayner, M.M., Wong, Y.N., Chang, C.H., Weber, P.C., Jackson, D.A., Sharpe, T.R., and Erickson-Viitanen, S. Rational design of potent, bioavailable, nonpeptide cyclic ureas as HIV protease inhibitors. *Science* 1994, **263,** 380–384

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

10 Voorinthold, R., Kosters, M.T., Vegter, G., Vriend, G., and Hol, W.G.T. A very fast program for visualising protein surfaces, channels and cavities. *J. Mol. Graphics* 1989, **7,** 243–245

11 Delaney, J.S. Finding and filling protein cavities using cellular logic operations. *J. Mol. Graphics* 1992, **10,** 174–177

12 Del Carpio, C.A., Takahashi, Y., and Sasaki, S. A new approach to the automatic identification of candidates for ligand receptor sites in proteins. 1. Search for pocket regions. *J. Mol. Graphics* 1993, **11,** 23–29

13 Kisljuk, O.S., Kachalova, G.S., and Lanina, N.P. An algorithm to find channels and cavities within protein crystals. *J. Mol. Graphics* 1994, **12,** 305–307

14 Masuya, M., and Doi, J. Detection and geometrical modelling of molecular surfaces and cavities using digital mathematical morphological operations. *J. Mol. Graphics* 1995, **13,** 331–336

15 Laskowski, R.A. SURFNET: A program for visualising molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graphics* 1995, **13,** 323–330

16 Hemm, K., Aberer, K., and Hendlich, M. Constituting a receptor–ligand information base from quality enriched data. In: *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology, 1995.* Robinson College, Cambridge, 1995, pp. 170–178

17 Hendlich, M., Rippmann, F., and Barnickel, G. In preparation (1998)