Finding and filling protein cavities using cellular logic operations

John S. Delaney

ICI Agrochemicals,† Jealott's Hill Research Station, Bracknell, Berkshire, UK

A method for solid-filling protein cavities is presented. The method uses a pattern-recognition technique based on cellular logic operations to distinguish between convex and concave regions of a protein. In doing this it solid fills protein cavities and automatically defines a boundary between cavity and exterior free space. The operations used to fill the cavities also can be used to process the filler to filter out small-scale features. So far the main use of the method has been in visualizing protein active sites for docking. The method can be used to find cavities of a given size range and could be used to find novel protein binding sites.

Keywords: Protein cavities, solid filling, pattern recognition, cellular logic

INTRODUCTION

It is well known that enzyme active sites and other small molecule-binding regions tend to be concave. This basic observation can be explained by the need to exclude solvent and create a highly specific environment for the ligand. The size and shape of these cavities play crucial roles in binding specificity. Thus it is important to be able to characterize them, quantitatively and qualitatively.

This need to analyze cavities creates some problems. It is difficult to discern the limits of a cavity given only the stick diagram of a protein, and this means that aids to cavity visualization are especially important. Analytical measurements of cavity size and shape are hampered by the difficulty of defining exactly where a cavity ends and free space begins. Enzymes contain many hollows of various sizes over their surface and it is possible that some of these may be usefully exploited by the right ligand. When looking for potential binding sites it is easy to be drawn to the biggest concavity while overlooking smaller hollows.

Several attempts have been made to address these problems over the years, particularly the visualization aspect. Molecular surfaces, such as those produced by Connolly's MS program,[1] are probably the most widely used aids for clarifying the boundaries of binding sites. This method uses Richards' definition of a solvent contact/reentrant surface[2] to produce a representation of the solvent-accessible regions of a protein. Connolly also has produced a program that analyzes a solvent-accessible surface in terms of solid angle[3] and provides a quantitative picture of concave regions of a protein. Kuntz et al. have produced a method that systematically searches for binding sites that may sterically match a given ligand.[4] Barford et al. used the lower atom packing density in channels to outline crevices by contouring the density values on a grid.[5] Recently Ho and Marshall have produced a program that solid fills a defined protein pocket.[6] The program is based on a standard two-dimensional (2D) filling routine and requires the user to define a seed point and a tethering constraint to prevent points from "flooding" the surrounding free space.

The aim of the method described below is to solid fill protein cavities in a similar way to Ho and Marshall's cavity search program. The program uses an established pattern recognition technique to distinguish between concave and convex portions of the protein.[7] This fills cavities and solves the difficult problem of defining the boundary between the cavity and free space without the need for the user to intervene. This feature lends itself to the rigorous measurement of the volumes of cavities. The program works on the protein as a whole and finds potentially novel binding sites automatically. Similar techniques can be used to clean up the overall image of the protein cavities by removing small-scale features.

This application uses cellular logic operations which perform transformations on binary images. These operations are usually used in 2D image processing for finding discrete objects in an image. In this case the method has been extended to three dimensions with the aim of finding concave regions of a protein.

METHOD

The program begins by generating a regular Cartesian grid of points that lie within the solvent-accessible surface of the protein. The solvent-accessible surface is defined as the surface described by the center of a spherical solvent probe in van der Waals' contact with the protein (this is the definition of Lee and Richards[8] rather than the subsequent Richards' contact surface as implemented by Connolly). The grid is implemented as a three-dimensional (3D) logical

Color Plates for this article are on page 163.

†ICI Agrochemicals in the UK is part of Imperial Chemical Industries PLC

Address reprint requests to Mr. Delaney at ICI Agrochemicals, Jealott's Hill Research Station, Bracknell, Berkshire, RG12 6EY, UK.

Received 22 October 1991; accepted 19 November 1991

© 1992 Butterworth–Heinemann

array; points within the surface are set true, all others are false. The program can perform two types of basic operation on this grid. The first is known as an expansion or dilation and involves adding a monolayer of points to the surface of the grid. Any grid point that is set false but has one or more neighbors set true is set true by this operation. No other points are affected. The neighborhood can be defined in at least two different ways on a cubic grid. Each point can have either 6 nearest neighbors (neighbors at the cube faces) or 26 nearest neighbors (neighbors at the cube faces, edges and vertices).[9] For the remainder of this discussion it is assumed that the 26 nearest neighbor definition applies, although both definitions have been implemented in the program. The overall effect of one expansion operation is to enlarge the volume of the protein by a layer of points. The second type of operation is a shrinkage or contraction. This is essentially the reverse of an expansion since it removes a monolayer of points from the surface of the grid. Any grid point that is set true but has one or more neighbors set false is also set false by this operation. One shrinkage operation reduces the volume by a layer of points.

The important thing to note is that these two operations do not commute; i.e., an expansion followed by a shrinkage on a given grid will not recreate necessarily the original grid, although it will produce a grid of approximately the same size. This property allows the program to distinguish concave portions of the protein from convex ones. The procedure for filling the protein cavities is to perform a number of expansions on the grid followed by an equal number of shrinkages. The exact number of operations required depends on the grid spacing being used and the size of cavity being sought. Typical values are between 5 and 10. After performing these operations the original grid can be subtracted logically from the post-processed grid (any true point in the original grid is set false in the processed grid), leaving the extra points that have accumulated in the crevices of the protein. It should be noted that all of these points lie outside the solvent-accessible surface of the protein as defined earlier. The effect of applying these operations on a grid in two dimensions is shown in Figure 1.

This combination of operations finds all cavities smaller than a threshold dimension. The threshold dimension relates to the smallest width of the cavity and can be calculated using

$$W = G(2N - 1) + 2P \quad (1)$$

where W is the width of feature found, N is the number of operations applied to the grid, G is the distance between grid points, and P is the probe radius.

With each expansion operation the opposing walls of a cavity have a layer of points added. When the walls meet, the filled cavity will not be removed by subsequent shrinkage operations, and further expansion operations will have no effect. As each layer is added, the distance between the walls decreases by two layers of points. Thus the distance between added points on opposite sides of a just-filled cavity is $G(2N - 1)$. By adding the probe diameter ($2P$), the separation between the opposite surfaces of the cavity is determined.

The consequence of this is that the program will find many small indentations in the protein surface, particularly thin lines of points linking larger cavities. One way to remove extraneous points is to use a distance constraint on the output so that only points within a certain distance of a user-specified protein atom are kept. This is useful in situations where a particular site is to be studied in depth.

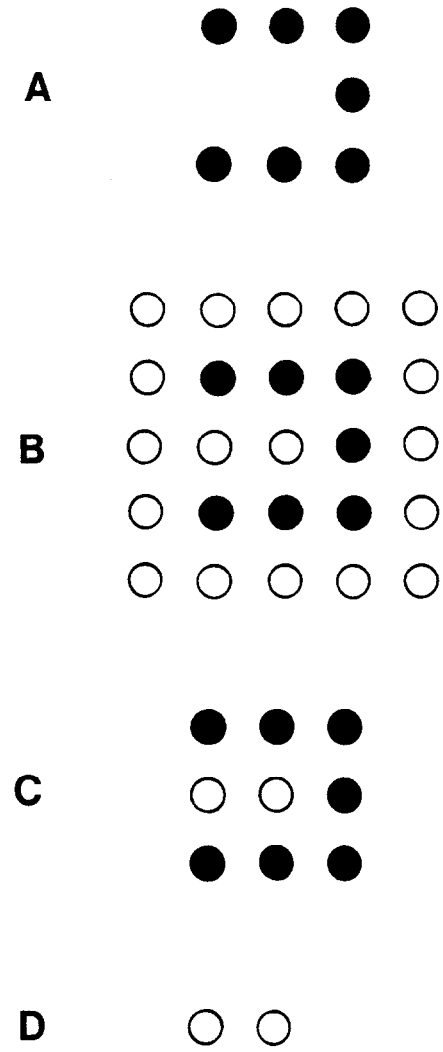

Figure 1. A, Points within protein accessible surface (solid points). B, One expansion operation; hollow points are the points added by expansion. C, One shrinkage operation; residual points left in cleft. D, Subtract original grid (A) leaving filled cavity.

Another way to filter out small scale features is to use a different combination of operations. It is possible to take the processed grid (i.e., the filled cavities after subtraction of the protein volume) and achieve other effects using the same operations. By performing the shrinkage operations before the expansions, rough edges and lines of points connecting larger cavities can be removed to leave discrete clusters. The drawback of this is that it tends to remove surface details from the grid, which may not be desirable if a close study of an active site is required. The advantage is that the volumes of these discrete cavities may be calculated easily and unambiguously. The number of times these operators are applied should be less than the number of operations used to generate the filled volumes, otherwise they will remove all points in the grid! A simple

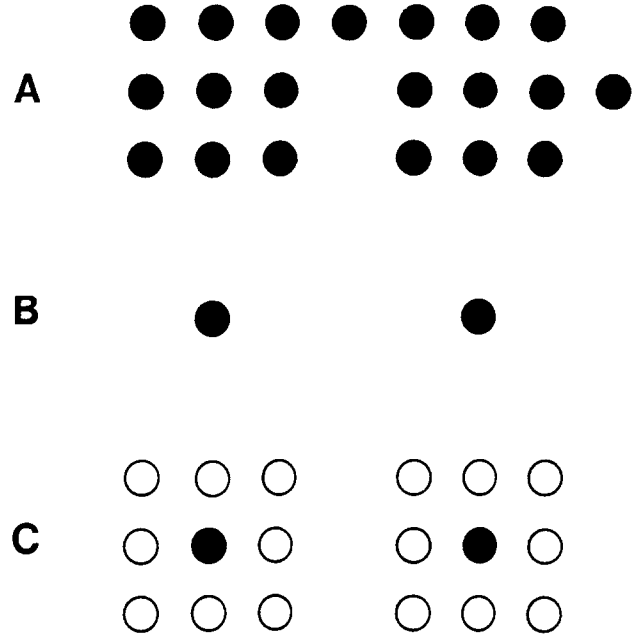

Figure 2. A, Filled cavity points—note bridge between cavities and surface detail on right; B, one shrinkage operation; and C, one expansion operation—hollow points added by expansion. The bridge between the two cavities has been broken and the surface detail removed.

example of a shrinkage followed by an expansion is shown in Figure 2.

The shrinkage/expansion combination of operations excludes cavities smaller than a threshold value. Equation (1) now defines the minimum width that will be kept. By combining these two sets of operations it is possible to tune to a particular size range of cavity. The main problem with doing this is that it can introduce quite severe distortions to the true shape of the cavity.

The processed points can be manipulated in several ways. By assigning each point to its nearest protein atom it is possible to color the solid filled cavities by atom type. This gives some feel for the distribution of hydrophobic and hydrophilic regions of the protein. The points can be treated as solvent molecules (since all points are at or outside the solvent accessible surface) allowing the van der Waals' surface of the assembly to be calculated. This surface corresponds to an approximation of the protein cavities' Connolly surface.

RESULTS

The test molecule used for this program was cytochrome P450 (2CPP in the Brookhaven database). This is a monooxygenase that binds camphor as a substrate. Unlike most of the enzymes studied by crystallography its active site is entirely enclosed by protein. The substrate must enter through an access channel created by the protein's natural movement. This provides an interesting challenge to the program, which has to find the enclosed site and the dip on the outside surface where the access channel forms. Before running the program all of the crystallographic waters and the bound camphor were removed so that the protein could be studied in isolation.

The first run used seven expansion/shrinkage cycles to fill the available cavities at a grid spacing of 0.6 Å and a probe radius of 1.5 Å. This successfully found the binding pocket, which is isolated from the outside of the protein, as well as a myriad of other features at the protein surface. This example illustrates the point that the surface of a protein is covered in small trenches, which tend to link larger depressions. These trenches are a by-product of the way a protein folds, and tend to occur at the interfaces between secondary structural elements. This image contains a lot of detail, but it is rather hard to assimilate (Color Plate 1).

To simplify the image a shrinkage/expansion was applied. This has the effect of breaking thin lines of points and removing isolated points to leave discrete cavities. These cavities tend to be distributed evenly over the surface of the protein (Color Plate 2), except for the active site, which is buried in the heart of the protein. A more extreme filtering was obtained by performing 7 expansion/shrinkage cycles followed by 3 shrinkage/expansion cycles on the protein isolating the binding site (Color Plate 3) and about eight other distinct pockets. It is noticeable that the shape of the binding cavity has been somewhat distorted by these operations. The size of cavity found by this combination of operations is in the range 6–11 Å. The substrate camphor is an approximately spherical molecule, with a diameter of 7–8 Å. Cavities smaller than 6 Å can be found by performing a separate calculation on the protein using 3 expansion/shrinkage cycles and no shrinkage/expansion cycles. This produces a new set of cavities which complement the filtered set. Thus it is possible to separate out the different levels of detail (Color Plate 4).

The access channel to the active site also shows up when the fine detail is restored. Neither the active site nor the access channel appears to be the largest cavity present. This raises the intriguing question as to what the purpose of the larger cavities is. They may be points where P450 interacts with other proteins.

The effect achieved by coloring the filled cavities by nearest protein atom type is shown in Color Plate 5. Hydrophobic pockets show up particularly well in this representation and can be displayed separately from other features.

IMPLEMENTATION

The program was written in DEC FORTRAN on a VAX 11/750 and moved to a Silicon Graphics 4D/220 with minimal alteration. Little regard has been paid to program efficiency in this implementation, and as a consequence the program uses a lot of CPU and memory resource. Also no effort has been made to utilize the multiprocessing facility on the Silicon Graphics. Memory and CPU requirements limit the resolution of the grid, so it would be useful to work on these aspects. The program reads and writes SYBYL MOL2 files[10] and uses enhanced van der Waals' radii for the protein atoms unless hydrogens have been added to the structure. The cavities are written out as MOL2 files, each point represented by an atom record.

The timings for these runs turned out to depend substantially on the time taken to decide whether points lie within the solvent-accessible surface of the protein. All of the runs

took about 45 minutes on one processor of the SGI 4D/220 at a grid spacing of 0.6 Å.

There are several problems with the program as it stands. Apart from the efficiency snags mentioned earlier, there is the problem of the level of detail. As it stands it is impossible to separate the major cavities without losing fine detail. The size of the output has been a significant limiting factor on the resolution, the finer the grid used the more points generated. Currently the program has a limit of 20000 points, which still presents our display hardware (an Evans & Sutherland PS390) with considerable difficulties in manipulating smoothly. The problem has been partially solved by removing points from the centres of the solid filled cavities leaving only the surface layer of points. This has little effect on the visual aspect of the cavities but in favorable cases it can significantly reduce the size of the output.

CONCLUSIONS

The initial motivation for writing this program was a need for an improved method for visualizing an active site cleft. Its applicability to other problems became apparent only after it had been written. The most pleasing thing about this program is that it addresses the problem of defining the limits of a cavity, especially its boundary with free space. It has been mentioned before[3] that it is extremely difficult to define protein topographical features in terms of a natural "sea level," since the size of features being sought are not sufficiently small in comparison to the whole object. The program fills on a layer by layer basis without imposing a coordinate system or defining "up."

The program has been used successfully to study the detailed structure of known binding sites. The level of detail is sufficient to enable molecules to be visually docked and the fact that the cavity is solid filled seems to help with docking. This process could be automated by using some form of volume overlay method, perhaps including critical distance constraints.

The program has proved capable of finding and filling protein cavities in a thorough and systematic fashion, as well as isolating features of a given size. The points are useful as cavity visualization aids and as starting points for other calculations.

FUTURE DEVELOPMENTS

Post processing the raw, filled cavities seems to offer considerable scope for development. It seems to be possible to filter out cavities of different sizes which can then be analyzed individually. This was apparent from the P450 work as the program could be successfully tuned to find the active site and little else. The problem is that in separating out cavities so much fine detail is lost. The solution may be to use the filtered, discrete cavities as seeds which can be applied to the prefiltered structure and allow all points to be assigned to one or other parent seed. This could be done by an iterative propagation algorithm that successively assigned nearest neighbors to seed groups until all points had been assigned. It may then be possible to recombine different components to generate hierarchies of structures. This might be a means of separating different major cavities from each other without losing any of the fine detail of their structure. One possible application of discrete cavities would be to answer the question "how frequently is the active site of an enzyme the largest pocket?" The general answer at the moment appears to be "usually,"[3,4] but there does not appear to be a definitive, quantitative answer in the literature.

ACKNOWLEDGEMENTS

The author would like to thank Keith Heritage for posing the original problem that led to this work being done and for his help and encouragement in its development. Additional thanks are due to Graham Sexton and Robin Taylor for their critical reading of this manuscript.

REFERENCES

1 Connolly, M.L. *Science* 1983, **221**, 709–713
2 Richards, F.M. *Ann. Rev. Biophys. Bioeng.* 1977, **6**, 151–176
3 Connolly, M.L. *J. Mol. Graph.* 1986, **4**, 3–6
4 Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R. and Ferrin, T.E. *J. Mol. Biol.* 1982, **161**, 269–288
5 Barford, D., Schwabe, J.W.R., Oikonomakos, N.G., Acharya, K.R., Hajdu, J., Papageorgiou, A.C., Martin, J.L., Knott, J.C.A., Vasella, A., and Johnson, L.N. *Biochemistry* 1988, **27**, 6733–6741
6 Ho, M.W. and Marshall, G.R. *J. Comp.-Aided Mol. Design* 1990, **4**, 337–354
7 James, M. *Pattern Recognition,* BSP Professional Books, Oxford, 1987, pp 113–116
8 Lee, B., Richards, F.M. *J. Mol. Biol.* 1971, **55**, 379–400
9 Kovalevsky, V.A. *Pattern Recognition Letters* 1984, **2**, 281–288
10 Tripos Associates Inc., St. Louis, Missouri, USA