APPENDIX H: IDENTIFYING SURFACE ATOMS

H.1 Definitions of molecular surface

The surface of a molecule has been defined using the concept of a probe sphere, representing a solvent molecule, rolling over the molecule's surface. Atoms that can be touched by the probe sphere are said to be accessible to solvent, and on the ''outside'' of the molecule. Typically, a probe sphere with the radius of a water molecule is used. Definitions of molecular surface have been proposed to allow solvent-accessible areas to be calculated.

Solvent-accessible area was initially defined by Lee and Richards (1971) as the locus of the centre of a probe sphere as it rolled over the surface of a molecule. The main disadvantage of using this definition of a molecular surface is that the surface defined is displaced from the Van der Waals surface.

Richards (1977) defines the molecular surface to consist of those parts of the Van der Waals surface that can be reached by a probe sphere (*contact* surface), with these parts joined by the saddle-shaped regions traced out by the inner-most part of the probe sphere as it rolls touching pairs of atoms, and ''triangular'' concave regions of the surface of the probe sphere as it rests touching triplets of atoms (the vertices of these regions are the points of contact of the probe sphere and the Van der Waals surface when the probe sphere is touching three atoms).

Both the above definitions are equivalent to the Van der Waals surface when a probe sphere of zero radius is specified. Also, the surface proposed by Lee and Richards (1971) is equivalent to the ''Van der Waals surface'' that would be calculated if the Van der Waals radius of each atom is expanded by the radius of a probe sphere.

The definition of molecular surface proposed by Richards (1977) has been implemented by Connolly in the MS program (1983a; 1983b), allowing the solvent-accessible surface to be visualised.

H.2 Distinguishing between interior and exterior atoms

Connolly's MS program does not distinguish between a molecule's exterior surface and the surfaces of internal cavities. The word *cavity* is used here to describe a space in the interior of a molecule where a probe sphere can be accommodated, but from where there is no path by which the probe sphere can reach the protein surface. Therefore, that an atom can be touched by a probe sphere is no guarantee that the atom will be on the outside of the molecule.

A program has been written, in C, to determine which atoms of a globular protein can be touched by a probe sphere of a specified radius. This program also finds which pairs and triplets of atoms that can be touched simultaneously by the probe sphere.

The program identifies adjacent pairs of accessible atoms by calculating which pairs of atoms contribute an arc that bounds a convex region of the molecular surface defined by Lee and Richards (1971). Such a pair is represented within the program by a line, or *edge*, joining the centres of the two atoms. If the arc contributed by two atoms is in fact a circle, then the edge joining those two atoms is marked as a *pair* (situations giving rise to such pairs include *cusps*, as described by Connolly (1985a,b)). If an arc contributed by two atoms is not a complete circle, then the end points correspond to the centre of the probe sphere at positions where it is simultaneously in contact with three atoms. In this case, the three edges between the atoms in this triplet are grouped into a *triangle*, and the normal vector from the plane of the triangle, towards the centre of the probe sphere, indicates which *face* of the triangle is towards the accessible surface. It is possible for both faces of a triangle to be towards the exterior surface.

Not all atoms involved in edges and triangles will necessarily be on the exterior surface, since these atoms could be adjacent to an interior cavity. An algorithm for identifying atoms that contribute to the exterior surface is presented here. The first step is to find an atom that is definitely on the exterior surface. Such an atom is that which is the first touched by a plane parallel to the x- and y-axes, brought towards the molecule from negative infinity. This is the atom for which 'z-coordinate minus Van der Waals radius' has the lowest value (it is the ''lowest atom''). Next, an edge that is definitely on the exterior surface is identified. That identified is the edge from the lowest atom, to a neighbouring one, that gives the smallest angle

with vector antiparallel to the z-axis. Then, the triangular face associated with this edge whose normal is closest to the negative z-axis is identified. This is a face corresponding to a portion of the exterior surface, in that the three atoms whose centres are the vertices of this face can be touched by a probe sphere brought towards the molecule from infinity. It is assumed that, with a suitable probe radius (e.g. 1.4 Angstroms), no atom will be involved in one (or more) edge(s) (i.e. *pairs*), but no triangles.

Having identified a triangular face that is guaranteed to involve three exterior atoms, other exterior faces are identified by a recursive routine that traverses the surface, crossing edges to adjacent exterior faces. It is assumed that all exterior atoms will be involved in at least one triangle and that each of the exterior triangles can be reached by crossing edges, starting from the lowest triangle.

Output from this program takes several forms. A list of edges is written to a file in a format from which Brookhaven format ''CONECT'' records can be written and appended to a Brookhaven file, enabling the mesh of triangles to be displayed using a molecular graphics package. Another file output from this program can be used to generate a colour file, for use with HYDRA, to show interior and exterior atoms in different colours. Information about atoms, edges and triangles is also written to a file in a format suitable for loading into the P/FDM database.

Information about the accessibility of particular residues and atoms can be useful in the evaluation of modelled structures. Baumann, Frommel and Sander (1989) propose some tests, e.g. tests concerning the accessibility of polar groups. Also, users may want to enquire about the accessibility of individual atoms, e.g. whether particular atoms expected to be in the functional site of the model are accessible.

This code for identifying interior and exterior atoms can be called from the Prolog environment, and has been made available as a database method.

H.3 Identifying cavities

Triangular faces that are not included in the exterior surface will join the centres of triplets of atoms that are involved in bounding an interior cavity. Those atoms bounding a particular cavity can be identified by selecting a triangular face corresponding to a portion of the surface of that cavity, and then traversing the rest of this surface using the recursive routine mentioned in the previous section.

Information on internal cavities may be useful in evaluating modelled protein structures, since these may suggest poor packing in parts of the molecule, indicating regions that should be remodelled.

H.4 Molecular shape measurement

Information about the geometry of the surface in the vicinity of particular atoms may be of interest in evaluating protein models. For example, it may be useful to know whether a particular atom is at the end of a protruding side chain, or whether another is accessible, but at the bottom of a pocket.

Connolly (1986a) describes a method for measuring surface shape using solid angles. This representation of molecular shape has been used in attempting to predict protein-protein interactions, based on matching complementary patterns of 'knobs' and 'holes' (Connolly 1986b). Connolly's method for detecting knobs and holes consists of first triangulating the molecular surface (Connolly, 1985b), and then evaluating a surface function for each of the vertices of the resulting polyhedron. The value of the surface function at each vertex is then compared with the values at each of the adjacent vertices to find local minima and maxima, corresponding to knobs and holes respectively. The surface function used consists of placing a ball of a specified radius centred on each surface point, filling the balls with cubes, and then for each ball counting how many cubes lie inside the molecular surface.

H.5 Finding knobs and holes using P/FDM

A procedure for identifying atoms at ''critical points'' (i.e. knobs and holes) on the molecular surface is presented here. The word *knob* is used here to describe an exterior atom that is more exposed than those exterior atoms to which it is adjacent. Similarly, the word *hole* is used here to describe an exterior atom that is less exposed than all its neighbouring surface atoms.

A computationally cheaper, but less accurate, alternative to Connolly's surface function is to apply a very simple exposure function to each exterior atom. The exposure function used here consists of placing

balls, of a specified radius, at the centre of each exterior atom, and, for each ball, counting how many atoms have their centres inside that ball. A low value for this function indicates that there are few nearby atoms, therefore it is likely that the atom under investigation will be highly exposed. The value of the exposure function for each exterior atom is compared with the values at each of the adjacent exterior atoms to find local minima and maxima, corresponding to knobs and holes respectively.

It is important to choose a suitable value for the radius of the ball to be placed on each of the exterior atoms. The radius has to be sufficiently large to ensure that the exposure function will give a large range of values. On the other hand, the radius has to be sufficiently small, otherwise the shape of the molecular surface far from the atom of interest will affect the value of the exposure function for that atom. Empirically, 8 Angstroms has been found to be a suitable ball radius for finding intuitively acceptable knobs and holes.

The exposure function values are calculated by a program, written in C, that takes as input a file of interior and exterior atoms, and produces a file containing the exposure function values in a form that can be loaded into the P/FDM database. A Prolog query is then used to find local minima and maxima, and create corresponding knob and hole entity instances (knobs and holes are declared to be subclasses of the class *atom*). When knobs and holes have been identified, a list of these can be written as a colour file for use with HYDRA.

This exposure function also enables protruding loops to be identified, since atoms of consecutive residues in these loops will have very low values for the exposure function.

Extract from Kemp, G.J.L. (1991) "Protein Modelling Using an Object-Oriented Database", PhD Thesis, University of Aberdeen.

Baumann, G., Frommel, C. and Sander, C. (1989) Polarity as a criterion in protein design. *Protein Engineering*, **2**, 329-334.

Connolly, M.L. (1983a) Analytical Molecular Surface Calculation. *J. Appl. Cryst.*, **16**, 548-558.

Connolly, M.L. (1983b) Solvent-Accessible Surfaces of Proteins and Nucleic Acids. *Science*, **221**, 709-712.

Connolly, M.L. (1985a) Computation of Molecular Volume. *J. Am. Chem. Soc.*, **107**, 1118-1124.

Connolly, M.L. (1985b) Molecular Surface Triangulation. *J. Appl. Cryst.*, **18**, 499-505.

Connolly, M.L. (1986a) Measurement of protein surface shape by solid angles. *J. Mol. Graphics*, **4**, 3-7.

Connolly, M.L. (1986b) Shape Complementarity at the Hemoglobin $\alpha 1 \beta 1$ Subunit Interface. *Biopolymers*, **25**, 1229-1247.

Lee, B. and Richards, F.M. (1971) Interpretation of Protein Structures — Estimation of Static Accessibility. *J. Mol. Biol.*, **55**, 379-400.

Richards, F.M. (1977) Areas, Volumes, Packing and Protein-Structure. *Annu. Rev. Biophys. Bioeng.*, **6**, 151-176.