International Tables for Crystallography (2006). Vol. F, Section 22.1.1, pp. 531–539.

22. MOLECULAR GEOMETRY AND FEATURES

22.1. Protein surfaces and volumes: measurement and use

BY M. GERSTEIN, F. M. RICHARDS, M. S. CHAPMAN AND M. L. CONNOLLY

22.1.1. Protein geometry: volumes, areas and distances

(M. GERSTEIN AND F. M. RICHARDS)

22.1.1.1. *Introduction*

For geometric analysis, a protein consists of a set of points in three dimensions. This information corresponds to the actual data provided by the experiment, which are fundamentally of a geometric rather than chemical nature. That is, crystallography primarily tells one about the positions of atoms and perhaps an approximate atomic number, but not their charge or number of hydrogen bonds.

For the purposes of geometric calculation, each point has an assigned identification number and a position defined by three coordinates in a right-handed Cartesian system. (These coordinates will be based on the electron density for X-ray derived structures and on nuclear positions for those derived from neutron scattering. Each coordinate is usually assumed to have an accuracy between 0.5 and 1.0 Å.) Normally, only one additional characteristic is associated with each point: its size, usually measured by a van der Waals (VDW) radius. Furthermore, characteristics such as chemical nature and covalent connectivity, if needed, can be obtained from lookup tables keyed on the ID number.

Our model of a protein, thus, is the van der Waals envelope – the set of interlocking spheres drawn around each atomic centre. In brief, the geometric quantities of the model of particular concern in this section are its total surface area, total volume, the division of these totals among the amino-acid residues and individual atoms, and the description of the empty space (cavities) outside the van der Waals envelope. These values are then used in the analysis of protein structure and properties.

All the geometric properties of a protein (*e.g.* surfaces, volumes, distances *etc.*) are obviously interrelated. So the definition of one quantity, *e.g.* area, obviously impacts on how another, *e.g.* volume, can be consistently defined. Here, we will endeavour to present definitions for measuring protein volume, showing how they are related to various definitions of linear distance (VDW parameters) and surface. Further information related to macromolecular geometry, focusing on volumes, is available from http://bioinfo.mbb.yale.edu/geometry.

22.1.1.2. *Definitions of protein volume*

22.1.1.2.1. *Volume in terms of Voronoi polyhedra: overview*

Protein volume can be defined in a straightforward sense through a particular geometric construction called the Voronoi polyhedron. In essence, this construction provides a useful way of partitioning space amongst a collection of atoms. Each atom is surrounded by a single convex polyhedron and allocated the space within it (Fig. 22.1.1.1). The faces of Voronoi polyhedra are formed by constructing dividing planes perpendicular to vectors connecting atoms, and the edges of the polyhedra result from the intersection of these planes.

Voronoi polyhedra were originally developed by Voronoi (1908) nearly a century ago. Bernal & Finney (1967) used them to study the structure of liquids in the 1960s. However, despite the general utility of these polyhedra, their application to proteins was limited by a serious methodological difficulty. While the Voronoi construction is based on partitioning space amongst a collection of 'equal' points, all protein atoms are not equal. Some are clearly larger than others. In 1974, a solution was found to this problem (Richards, 1974), and since then Voronoi polyhedra have been applied to proteins.

22.1.1.2.2. *The basic Voronoi construction*

22.1.1.2.2.1. *Integrating on a grid*

The simplest method for calculating volumes with Voronoi polyhedra is to put all atoms in the system on a fine grid. Then go to each grid point (*i.e.* voxel) and add its infinitesimal volume to the atom centre closest to it. This is prohibitively slow for a real protein structure, but it can be made somewhat faster by randomly sampling grid points. It is, furthermore, a useful approach for high-dimensional integration (Sibbald & Argos, 1990).

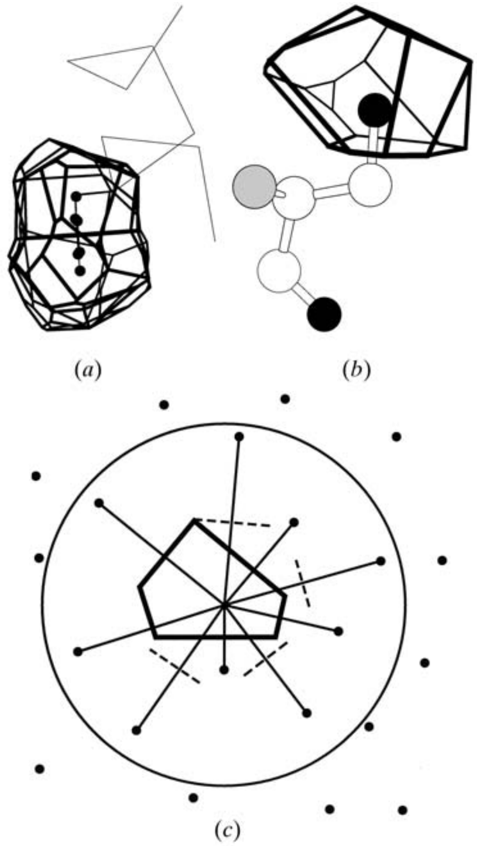

Fig. 22.1.1.1. The Voronoi construction in two and three dimensions. Representative Voronoi polyhedra from 1CSE (subtilisin) are shown. (*a*) Six polyhedra around the atoms in a Phe ring. (*b*) A single polyhedron around the side-chain hydroxyl oxygen (OG) of a serine. (*c*) A schematic showing the construction of a Voronoi polyhedron in two dimensions. The broken lines indicate planes that were initially included in the polyhedron but then removed by the 'chopping-down' procedure (see Fig. 22.1.1.4).

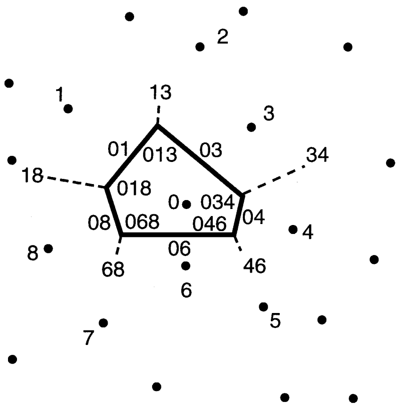

Fig. 22.1.1.2. Labelling parts of Voronoi polyhedra. The central atom is atom 0, and each neighbouring atom has a sequential index number (1, 2, 3. . .). Consequently, in three dimensions, planes are denoted by the indices of the two atoms that form them (*e.g.* 01); lines are denoted by the indices of three atoms (*e.g.* 012); and vertices are denoted by four indices (*e.g.* 0123). In the 2D representation shown here, lines are denoted by two indices, and vertices by three. From a collection of points, a volume can be calculated by a variety of approaches: First of all, the volume of a tetrahedron determined by four points can be calculated by placing one vertex at the origin and evaluating the determinant formed from the remaining three vertices. (The tetrahedron volume is one-sixth of the determinant value.) The determinant can be quickly calculated by a vector triple product, $\mathbf{w} \cdot (\mathbf{u} \times \mathbf{v})$, where **u**, **v** and **w** are vectors between the vertex selected to be the origin and the other three vertices of the tetrahedron. Alternatively, the volume of the pyramid from a central atom to a face can be calculated from the usual formula $Ad/3$, where A is the area of the face and d is the distance to the face.

More realistic approaches to calculating Voronoi volumes have two parts: (1) for each atom find the vertices of the polyhedron around it and (2) systematically collect these vertices to draw the polyhedron and calculate its volume.

22.1.1.2.2.2. *Finding polyhedron vertices*

In the basic Voronoi construction (Fig. 22.1.1.1), each atom is surrounded by a unique limiting polyhedron such that all points within an atom's polyhedron are closer to this atom than all other atoms. Consequently, points equidistant from two atoms lie on a dividing plane; those equidistant from three atoms are on a line, and those equidistant from four atoms form a vertex. One can use this last fact to find all the vertices associated with an atom easily. With the coordinates of four atoms, it is straightforward to solve for possible vertex coordinates using the equation of a sphere. [That is, one uses four sets of coordinates (x, y, z) and the equation $(x-a)^2 + (y-b)^2 + (z-c)^2 = r^2$ to solve for the centre (a, b, c) and radius (r) of the sphere.] One then checks whether this putative vertex is closer to these four atoms than any other atom; if so, it is a real vertex.

Note that this procedure can fail for certain pathological arrangements of atoms that would not normally be encountered in a real protein structure. These occur if there is a centre of symmetry, as in a regular cubic lattice or in a perfect hexagonal ring in a protein (see Procacci & Scateni, 1992). Centres of symmetry can be handled (in a limited way) by randomly perturbing the atoms a small amount and breaking the symmetry. Alternatively, the 'chopping-down' method described below is not affected by symmetry centres – an important advantage to this method of calculation.

22.1.1.2.2.3. *Collecting vertices and calculating volumes*

To collect the vertices associated with an atom systematically, label each one by the indices of the four atoms with which it is associated (Fig. 22.1.1.2). To traverse the vertices on one face of a polyhedron, find all vertices that share two indices and thus have two atoms in common, *e.g.* a central atom (atom 0) and another atom (atom 1). Arbitrarily pick a vertex to start at and walk around the perimeter of the face. One can tell which vertices are connected by edges because they will have a third atom in common (in addition to atom 0 and atom 1). This sequential walking procedure also provides a way of drawing polyhedra on a graphics device. More importantly, with reference to the starting vertex, the face can be divided into triangles, for which it is trivial to calculate areas and volumes (see Fig. 22.1.1.2 for specifics).

22.1.1.2.3. *Adapting Voronoi polyhedra to proteins*

In the procedure outlined above, all atoms are considered equal, and the dividing planes are positioned midway between atoms (Fig. 22.1.1.3). This method of partition, called bisection, is not physically reasonable for proteins, which have atoms of obviously different size (such as oxygen and sulfur). It chemically misallocates volume, giving excess to the smaller atom.

Two principal methods of repositioning the dividing plane have been proposed to make the partition more physically reasonable: method B (Richards, 1974) and the radical-plane method (Gellatly & Finney, 1982). Both methods depend on the radii of the atoms in contact (R for the larger atom and r for the smaller one) and the distance between the atoms (D). As shown in Fig. 22.1.1.3, they position the plane at a distance d from the larger atom. This distance is always set such that the plane is closer to the smaller atom.

22.1.1.2.3.1. *Method B and a simplification of it: the ratio method*

Method B is the more chemically reasonable of the two and will be emphasized here. For atoms that are covalently bonded, it divides the distance between the atoms proportionaly according to their covalent-bond radii:

$$d = DR/(R + r). \qquad (22.1.1.1)$$

For atoms that are not covalently bonded, method B splits the remaining distance between them after subtracting their VDW radii:

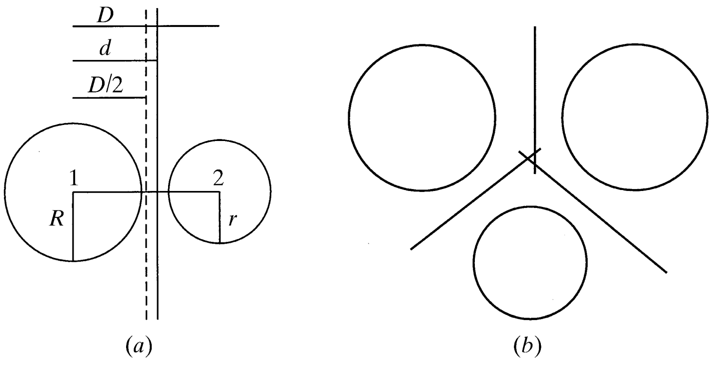

Fig. 22.1.1.3. Positioning of the dividing plane. (*a*) The dividing plane is positioned at a distance d from the larger atom with respect to radii of the larger atom (R) and the smaller atom (r) and the total separation between the atoms (D). (*b*) Vertex error. One problem with using method B is that the calculation does not account for all space, and tiny tetrahedra of unallocated volume are created near the vertices of each polyhedron. Such an error tetrahedron is shown. The radical-plane method does not suffer from vertex error, but it is not as chemically reasonable as method B.

$$d = R + (D - R - r)/2. \qquad (22.1.1.2)$$

For separations that are not very different to the sum of the radii, the two formulae for method B give essentially the same result. Consequently, it is worthwhile to try a slight simplification of method B, which we call the 'ratio method'. Instead of using equation (22.1.1.1) for bonded atoms and equation (22.1.1.2) for non-bonded ones, one can just use equation (22.1.1.2) in both cases with either VDW or covalent radii (Tsai *et al.*, 2001). Doing this gives more consistent reference volumes (manifest in terms of smaller standard deviations about the mean).

22.1.1.2.3.2. *Vertex error*

If bisection is not used to position the dividing plane, it is much more complicated to find the vertices of the polyhedron, since a vertex is no longer equidistant from four atoms. Moreover, it is also necessary to have a reasonable scheme for 'typing' atoms and assigning them radii.

More subtly, when using the plane positioning determined by method B, the allocation of space is no longer mathematically perfect, since the volume in a tiny tetrahedron near each polyhedron vertex is not allocated to any atom (Fig. 22.1.1.3). This is called vertex error. However, calculations on periodic systems have shown that, in practice, vertex error does not amount to more than 1 part in 500 (Gerstein *et al.*, 1995).

22.1.1.2.3.3. *'Chopping-down' method of finding vertices*

Because of vertex error and the complexities in locating vertices, a different algorithm has to be used for volume calculation with method B. (It can also be used with bisection.) First, surround the central atom (for which a volume is being calculated) by a very large, arbitrarily positioned tetrahedron. This is initially the 'current polyhedron'. Next, sort all neighbouring atoms by distance from the central atom and go through them from nearest to farthest. For each neighbour, position a plane perpendicular to the vector connecting it to the central atom according to the predefined proportion (*i.e.* from the method B formulae or bisection). Since a Voronoi polyhedron is always convex, if any vertices of the current polyhedron are on the other side of this plane to the central atom, they cannot be part of the final polyhedron and should be discarded. After this has been done, the current polyhedron is recomputed using the plane to 'chop it down'. This process is shown schematically in Fig. 22.1.1.4. When it is finished, one has a list of vertices that can be traversed to calculate volumes, as in the basic Voronoi procedure.

22.1.1.2.3.4. *Radical-plane method*

The radical-plane method does not suffer from vertex error. In this method, the plane is positioned according to

$$d = (D^2 + R^2 - r^2)/2D. \qquad (22.1.1.3)$$

22.1.1.2.4. *Delaunay triangulation*

Voronoi polyhedra are closely related (*i.e.* dual) to another useful geometric construction called the Delaunay triangulation. This consists of lines, perpendicular to Voronoi faces, connecting each pair of atoms that share a face (Fig. 22.1.1.5).

Delaunay triangulation is described here as a derivative of the Voronoi construction. However, it can be constructed directly from the atom coordinates. In two dimensions, one connects with a triangle any triplet of atoms if a circle through them does not enclose any additional atoms. Likewise, in three dimensions one connects four atoms with a tetrahedron if the sphere through them does not contain any further atoms. Notice how this construction is equivalent to the specification for Voronoi polyhedra and, in a sense, is simpler. One can immediately see the relationship between the triangulation and the Voronoi volume by noting that the volume is the distance between neighbours (as determined by the triangulation) weighted by the area of each polyhedral face. In practice, it is often easier in drawing to construct the triangles first and then build the Voronoi polyhedra from them.

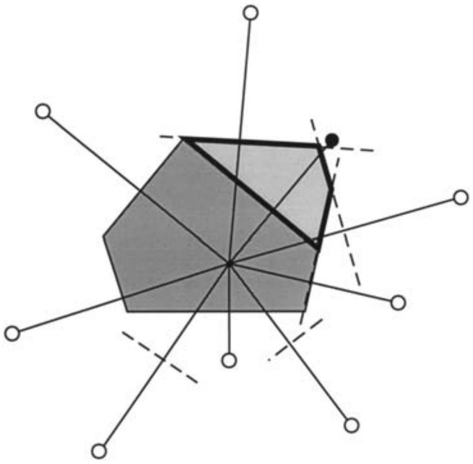

Fig. 22.1.1.4. The 'chopping-down' method of polyhedra construction. This is necessary when using method B for plane positioning, since one can no longer solve for the position of vertices. One starts with a large tetrahedron around the central atom and then 'chops it down' by removing vertices that are outside the plane formed by each neighbour. For instance, say vertex 0214 of the current polyhedron is outside the plane formed by neighbour 6. One needs to delete 0214 from the list of vertices and recompute the polyhedron using the new vertices formed from the intersection of the plane formed by neighbour 6 and the current polyhedron. Using the labelling conventions in Fig. 22.1.1.2, one finds that these new vertices are formed by the intersection of three lines (021, 024 and 014) with plane 06. Therefore one adds the new vertices 0216, 0246 and 0146 to the polyhedron. However, there is a snag: it is necessary to check whether any of the three lines are not also outside of the plane. To do this, when a vertex is deleted, all the lines forming it (*e.g.* 021, 024, 014) are pushed onto a secondary list. Then when another vertex is deleted, one checks whether any of its lines have already been deleted. If so, this line is not used to intersect with the new plane. This process is shown schematically in two dimensions. For the purposes of the calculations, it is useful to define a plane created by a vector **v** from the central atom to the neighbouring atom using a constant K so that for any point **u** on the plane $\mathbf{u} \cdot \mathbf{v} = K$. If $\mathbf{u} \cdot \mathbf{V} > K$, **u** is on the wrong side of the plane, otherwise it is on the right side. A vertex point **w** satisfies the equations of three planes: $\mathbf{w} \cdot \mathbf{v}_1 = K_1$, $\mathbf{w} \cdot \mathbf{v}_2 = K_2$ and $\mathbf{w} \cdot \mathbf{v}_3 = K_3$. These three equations can be solved to give the components of **w**. For example, the x component is given by

$$w_x = \begin{pmatrix} K_1 & v_{1y} & v_{1z} \\ K_2 & v_{2y} & v_{2z} \\ K_3 & v_{3y} & v_{3z} \end{pmatrix} \Bigg/ \begin{pmatrix} v_{1x} & v_{1y} & v_{1z} \\ v_{2x} & v_{2y} & v_{2z} \\ v_{3x} & v_{3y} & v_{3z} \end{pmatrix}.$$

Delaunay triangulation is useful in many 'nearest-neighbour' problems in computational geometry, *e.g.* trying to find the neighbour of a query point or finding the largest empty circle in a collection of points (O'Rourke, 1994). Since this triangulation has the 'fattest' possible triangles, it is the choice for procedures such as finite-element analysis.

In terms of protein structure, Delaunay triangulation is the natural way to determine packing neighbours, either in protein structure or molecular simulation (Singh *et al.*, 1996; Tsai *et al.*, 1996, 1997). Its advantage is that the definition of a neighbour does not depend on distance. The alpha shape is a further generalization of Delaunay triangulation that has proven useful in identifying ligand-binding sites (Edelsbrunner *et al.*, 1996, 1995; Edelsbrunner & Mucke, 1994; Peters *et al.*, 1996).

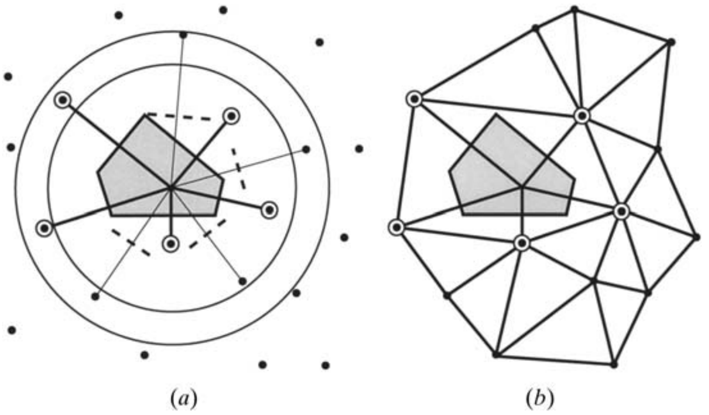

Fig. 22.1.1.5. Delaunay triangulation and its relation to the Voronoi construction. (*a*) A standard schematic of the Voronoi construction. The atoms used to define the Voronoi planes around the central atom are circled. Lines connecting these atoms to the central one are part of the Delaunay triangulation, which is shown in (*b*). Note that atoms included in the triangulation cannot be selected strictly on the basis of a simple distance criterion relative to the central atom. The two circles about the central atoms illustrate this. Some atoms within the outer circle but outside the inner circle are included in the triangulation, but others are not. In the context of protein structure, Delaunay triangulation is useful in identifying true 'packing contacts', in contrast to those contacts found purely by distance threshold. The broken lines in (*a*) indicate planes that were initially included in the polyhedron but then removed by the 'chopping-down' procedure (see Fig. 22.1.1.4).

22.1.1.3. *Definitions of protein surface*

22.1.1.3.1. *The problem of the protein surface*

When one is carrying out the Voronoi procedure, if a particular atom does not have enough neighbours the 'polyhedron' formed around it will not be closed, but rather will have an open, concave shape. As it is not often possible to place enough water molecules in an X-ray crystal structure to cover all the surface atoms, these 'open polyhedra' occur frequently on the protein surface (Fig. 22.1.1.6). Furthermore, even when it is possible to define a closed polyhedron on the surface, it will often be distended and too large. This is the problem of the protein surface in relation to the Voronoi construction.

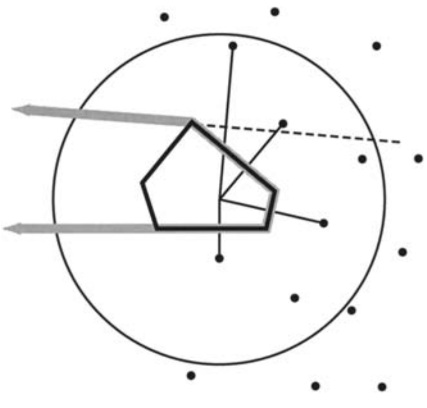

Fig. 22.1.1.6. The problem of the protein surface. This figure shows the difficulty in constructing Voronoi polyhedra for atoms on the protein surface. If all the water molecules near the surface are not resolved in a crystal structure, one often does not have enough neighbours to define a closed polyhedron. This figure should be compared with Fig. 22.1.1.1, illustrating the basic Voronoi construction. The two figures are the same except that in this figure, some of the atoms on the left are missing, giving the central atom an open polyhedron. The broken lines indicate planes that were initially included in the polyhedron but then removed by the 'chopping-down' procedure (see Fig. 22.1.1.4).

There are a number of practical techniques for dealing with this problem. First, one can use very high resolution protein crystal structures, which have many solvent atoms positioned (Gerstein & Chothia, 1996). Alternatively, one can make up the positions of missing solvent molecules. These can be placed either according to a regular grid-like arrangement or, more realistically, according to the results of molecular simulation (Finney *et al.*, 1980; Gerstein *et al.*, 1995; Richards, 1974).

22.1.1.3.2. *Definitions of surface in terms of Voronoi polyhedra (the convex hull)*

More fundamentally, however, the 'problem of the protein surface' indicates how closely linked the definitions of surface and volume are and how the definition of one, in a sense, defines the other. That is, the two-dimensionsl (2D) surface of an object can be defined as the boundary between two 3D volumes. More specifically, the polyhedral faces defining the Voronoi volume of a collection of atoms also define their surface. The surface of a protein consists of the union of (connected) polyhedra faces. Each face in this surface is shared by one solvent atom and one protein atom (Fig. 22.1.1.7).

Another somewhat related definition is the convex hull, the smallest convex polyhedron that encloses all the atom centres (Fig. 22.1.1.7). This is important in computer-graphics applications and as an intermediary in many geometric constructions related to proteins (Connolly, 1991; O'Rourke, 1994). The convex hull is a subset of the Delaunay triangulation of the surface atoms. It is quickly located by the following procedure (Connolly, 1991): Find the atom farthest from the molecular centre. Then choose two of its neighbours (as determined by the Delaunay triangulation) such that a plane through these three atoms has all the remaining atoms of the molecule on one side of it (the 'plane test'). This is the first triangle in the convex hull. Then one can choose a fourth atom connected to at least two of the three in the triangle and repeat the plane test, and by iteratively repeating this procedure, one can 'sweep' across the surface of the molecule and define the whole convex hull.

Other parts of the Delaunay triangulation can define additional surfaces. The part of the triangulation connecting the first layer of water molecules defines a surface, as does the part joining the second layer. The second layer of water molecules, in fact, has been suggested on physical grounds to be the natural boundary for a protein in solution (Gerstein & Lynden-Bell, 1993*c*). Protein surfaces defined in terms of the convex hull or water layers tend to be 'smoother' than those based on Voronoi faces, omitting deep grooves and clefts (see Fig. 22.1.1.7).

22.1.1.3.3. *Definitions of surface in terms of a probe sphere*

In the absence of solvent molecules to define Voronoi polyhedra, one can define the protein surface in terms of the position of a hypothetical solvent, often called the probe sphere, that 'rolls' around the surface (Richards, 1977) (Fig. 22.1.1.7). The surface of the probe is imagined to be maintained at a tangent to the van der Waals surface of the model.

Various algorithms are used to cause the probe to visit all possible points of contact with the model. The locus of either the centre of the probe or the tangent point to the model is recorded. Either through exact analytical functions or numerical approximations of adjustable accuracy, the algorithms provide an estimate of the area of the resulting surface. (See Section 22.1.2 for a more extensive discussion of the definition, calculation and use of areas.)

Depending on the probe size and whether its centre or point of tangency is used to define the surface, one arrives at a number of

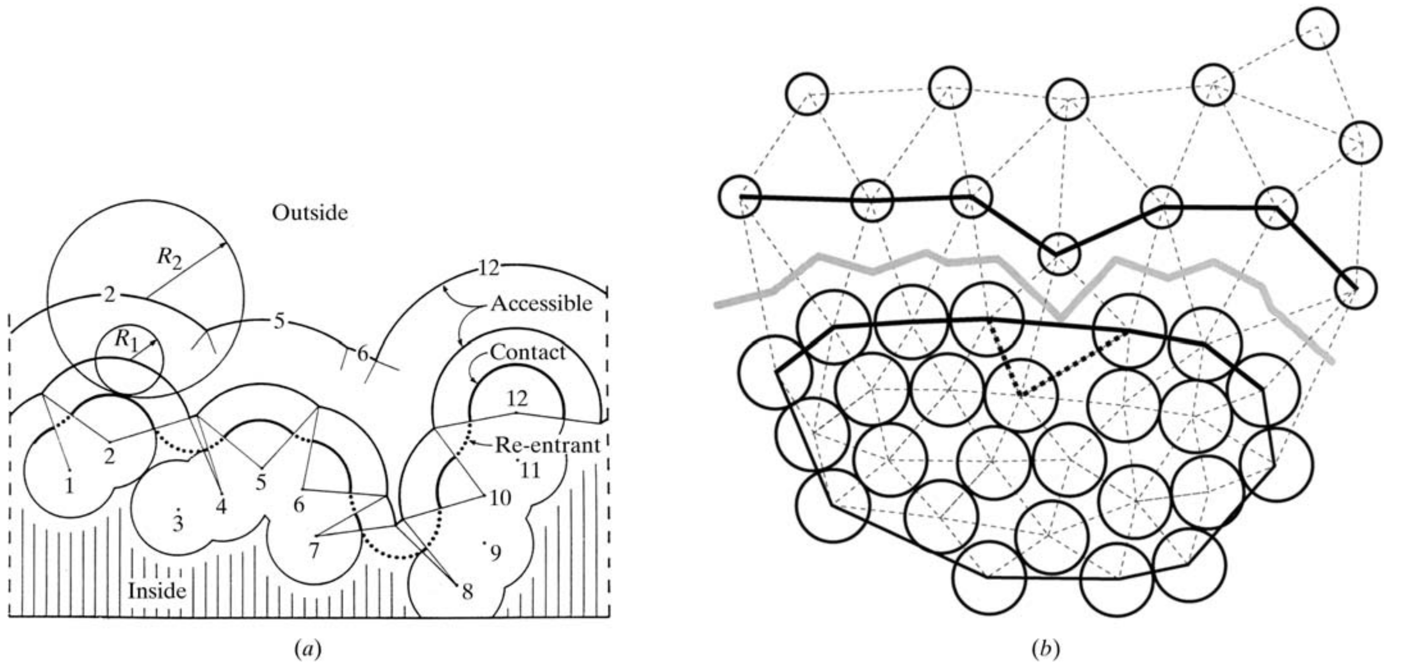

Fig. 22.1.1.7. Definitions of the protein surface. (*a*) The classic definitions of protein surface in terms of the probe sphere, the accessible surface and the molecular surface. (This figure is adapted from Richards, 1977). (*b*) Voronoi polyhedra and Delaunay triangulation can also be used to define a protein surface. In this schematic, the large spheres represent closely packed protein atoms and the smaller spheres represent the small loosely packed water molecules. The Delaunay triangulation is shown by dotted lines. Some parts of the triangulation can be used to define surfaces. The outermost part of the triangulation of *just* the protein atoms forms the convex hull. This is indicated by the thick line around the protein atoms. For the convex-hull construction, one imagines that the water is not present. This is highlighted by the thick dotted line, which shows how Delaunay triangulation of the surface atoms in the presence of the water diverges from the convex hull near a deep cleft. Another part of the triangulation, also indicated by thick black lines, connects the first layer of water molecules (those that touch protein atoms). A time-averaged version of this line approximates the accessible surface. Finally, the light thick lines show the Voronoi faces separating the protein surface atoms from the first layer of water molecules. Note how this corresponds approximately to the molecular surface (considering the water positions to be time-averaged). These correspondences between the accessible and molecular surfaces and time-averaged parts of the Voronoi construction are understandable in terms of which part of the probe sphere (centre or point of tangency) is used for the surface definition. The accessible surface is based on the position of the centre of the probe sphere while the molecular surface is based on the points of tangency between the probe sphere and the protein atoms, and these tangent points are similarly positioned to Voronoi faces, which bisect interatomic vectors between solvent and protein atoms.

commonly used definitions, summarized in Table 22.1.1.2 and Fig. 22.1.1.7.

22.1.1.3.3.1. *van der Waals surface (VDWS)*

The area of the van der Waals surface will be calculated by the various area algorithms (see Section 22.1.2.2) when the probe radius is set to zero. This is a mathematical calculation only. There is no physical procedure that will measure van der Waals surface area directly. From a mathematical point of view, it is just the first of a set of solvent-accessible surfaces calculated with differing probe radii.

22.1.1.3.3.2. *Solvent-accessible surface (SAS)*

The solvent-accessible surface is convex and closed, with defined areas assignable to each individual atom (Lee & Richards, 1971). However, the individual calculated values vary in a complex fashion with variations in the radii of the probe and protein atoms. This radius is frequently, but not always, set at a value considered to represent a water molecule (1.4 Å). The total SAS area increases without bound as the size of the probe increases.

22.1.1.3.3.3. *Molecular surface as the sum of the contact and re-entrant surfaces (MS = CS + RS)*

Like the solvent-accessible surface, the molecular surface is also closed, but it contains a mixture of convex and concave patches, the sum of the contact and re-entrant surfaces. The ratio of these two surfaces varies with probe radius. In the limit of infinite probe radius, the molecular surface becomes convex and attains a limiting minimum value (*i.e.* it becomes a convex hull, similar to the one described above). The molecular surface cannot be divided up and assigned unambiguously to individual atoms.

The contact surface is not closed. Instead, it is a series of convex patches on individual atoms, simply related to the solvent-accessible surface of the same atoms. In complementary fashion, the re-entrant surface is also not closed but is a series of concave patches that is part of the probe surface where it contacts two or three atoms simultaneously. At infinite probe radius, the re-entrant areas are plane surfaces, at which point the molecular surface becomes a convex surface. The re-entrant surface cannot be divided up and assigned unambiguously to individual atoms. Note that the molecular surface is simply the union of the contact and re-entrant surfaces, so in terms of area MS = CS + RS.

22.1.1.3.3.4. *Further points*

The detail provided by these surfaces will depend on the radius of the probe used for their construction.

One may argue that the behaviour of the rolling probe sphere does not accurately model real hydrogen-bonded water. Instead, its 'rolling' more closely mimics the behaviour of a nonpolar solvent. An attempt has been made to incorporate more realistic hydrogen-bonding behavior into the probe sphere, allowing for the definition of a hydration surface more closely linked to the behaviour of real water (Gerstein & Lynden-Bell, 1993*c*).

The definitions of accessible surface and molecular surface can be related back to the Voronoi construction. The molecular surface is similar to 'time-averaging' the surface formed from the faces of

Voronoi polyhedra (the Voronoi surface) over many water configurations, and the accessible surface is similar to averaging the Delaunay triangulation of the first layer of water molecules over many configurations.

There are a number of other definitions of protein surfaces that are unrelated to either the probe-sphere method or Voronoi polyhedra and provide complementary information (Kuhn *et al.*, 1992; Leicester *et al.*, 1988; Pattabiraman *et al.*, 1995).

22.1.1.4. *Definitions of atomic radii*

The definition of protein surfaces and volumes depends greatly on the values chosen for various parameters of linear dimension – in particular, van der Waals and probe-sphere radii.

22.1.1.4.1. *van der Waals radii*

For all the calculations outlined above, the hard-sphere approximation is used for the atoms. (One must remember that in reality atoms are neither hard nor spherical, but this approximation has a long history of demonstrated utility.) There are many lists of the radii of such spheres prepared by different laboratories, both for single atoms and for unified atoms, where the radii are adjusted to approximate the joint size of the heavy atom and its bonded hydrogen atoms (clearly not an actual spherical unit).

Some of these lists are reproduced in Table 22.1.1.1. They are derived from a variety of approaches, *e.g.* looking for the distances of closest approach between atoms (the Bondi set) and energy calculations (the *CHARMM* set). The differences between the sets often come down to how one decides to truncate the Lennard–Jones potential function. Further differences arise from the parameterization of water and other hydrogen-bonding molecules, as these substances really should be represented with two radii, one for their hydrogen-bonding interactions and one for their VDW interactions.

Perhaps because of the complexities in defining VDW parameters, there are some great differences in Table 22.1.1.1. For instance, the radius for an aliphatic CH (>CH=) ranges from 1.7 to 2.38 Å, and the radius for carboxyl oxygen ranges from 1.34 to 1.89 Å. Both of these represent at least a 40% variation. Moreover, such differences are practically quite significant, since many geometrical and energetic calculations are very sensitive to the choice of VDW parameters, particularly the relative values within a single list. (Repulsive core interactions, in fact, vary almost

Table 22.1.1.1. *Standard atomic radii (Å)*

For '*' see following notes on specific sets of values. *Bondi*: Values assigned on the basis of observed packing in condensed phases (Bondi, 1968). *Lee & Richards*: Values adapted from Bondi (1964) and used in Lee & Richards (1971). *Shrake & Rupley*: Values taken from Pauling (1960) and used in Shrake & Rupley (1973). >C= value can be either 1.5 or 1.85. *Richards*: Minor modification of the original Bondi set in Richards (1974). (Rationale not given.) See original paper for discussion of aromatic carbon value. *Chothia*: From packing in amino-acid crystal structures. Used in Chothia (1975). *Richmond & Richards*: No rationale given for values used in Richmond & Richards (1978). *Gelin & Karplus*: Origin of values not specified. Used in Gelin & Karplus (1979). *Dunfield et al.*: Detailed description of deconvolution of molecular crystal energies. Values represent one-half of the heavy-atom separation at the minimum of the Lennard–Jones 6–12 potential functions for symmetrical interactions. Used in Nemethy *et al.* (1983) and Dunfield *et al.* (1979). *ENCAD*: A set of radii, derived in Gerstein *et al.* (1995), based solely on the *ENCAD* molecular dynamics potential function in Levitt *et al.* (1995). To determine these radii, the separation at which the 6–12 Lennard–Jones interaction energy between equivalent atoms was 0.25 k_BT was determined (0.15 kcal mol^{-1}; 1 kcal = 4.184 kJ). *CHARMM*: Determined in the same way as the *ENCAD* set, but for the *CHARMM* potential (Brooks *et al.*, 1983) (parameter set 19). *Tsai et al.*: Values derived from a new analysis (Tsai *et al.*, 1999) of the most common distances of approach of atoms in the Cambridge Structural Database.

Atom type and symbol		Bondi (1968)	Lee & Richards (1971)	Shrake & Rupley (1973)	Richards (1974)	Chothia (1975)	Richmond & Richards (1978)	Gelin & Karplus (1979)	Dunfield *et al.* (1979)	*ENCAD* derived (1995)	*CHARMM* derived (1995)	Tsai *et al.* (1999)
$-CH_3$	Aliphatic, methyl	2.00	1.80	2.00	2.00	1.87	1.90	1.95	2.13	1.82	1.88	1.88
$-CH_2-$	Aliphatic, methyl	2.00	1.80	2.00	2.00	1.87	1.90	1.90	2.23	1.82	1.88	1.88
$>CH-$	Aliphatic, CH	—	1.70	2.00	2.00	1.87	1.90	1.85	2.38	1.82	1.88	1.88
$\geqslant CH=$	Aromatic, CH	—	1.80	1.85	*	1.76	1.70	1.90	2.10	1.74	1.80	1.76
$>C=$	Trigonal, aromatic	1.74	1.80	*	1.70	1.76	1.70	1.80	1.85	1.74	1.80	1.61
$-NH_3^+$	Amino, protonated	—	1.80	1.50	2.00	1.50	0.70	1.75	—	1.68	1.40	1.64
$-NH_2$	Amino or amide	1.75	1.80	1.50	—	1.65	1.70	1.70	—	1.68	1.40	1.64
$>NH$	Peptide, NH or N	1.65	1.52	1.40	1.70	1.65	1.70	1.65	1.75	1.68	1.40	1.64
$=O$	Carbonyl oxygen	1.50	1.80	1.40	1.40	1.40	1.40	1.60	1.56	1.34	1.38	1.42
$-OH$	Alcoholic hydroxyl	—	1.80	1.40	1.60	1.40	1.40	1.70	—	1.54	1.53	1.46
$-OM$	Carboxyl oxygen	—	1.80	1.89	1.50	1.40	1.40	1.60	1.62	1.34	1.41	1.42
$-SH$	Sulfhydryl	—	1.80	1.85	—	1.85	1.80	1.90	—	1.82	1.56	1.77
$-S-$	Thioether or –S–S–	1.80	—	—	1.80	1.85	1.80	1.90	2.08	1.82	1.56	1.77

Table 22.1.1.2. *Probe radii and their relation to surface definition*

The values of 1.4 and, especially, 10 Å are only approximate. One could, of course, use 1.5 Å for a water radius or 15 Å for a ligand radius, depending on the specific application.

Probe radius (Å)	Part of probe sphere	Type of surface
0	Centre (or tangent)	van der Waals surface (VDWS)
1.4	Centre	Solvent-accessible surface (SAS)
1.4	Tangent (one atom)	Contact surface (CS, from parts of atoms)
1.4	Tangent (two or three atoms)	Re-entrant surface (RS, from parts of probe)
1.4	Tangent (one, two, or three atoms)	Molecular surface (MS = CS + RS)
10	Centre	A ligand- or reagent-accessible surface
∞	Tangent	Minimum limit of MS (related to convex hull)
∞	Centre	Undefined

exponentially.) Consequently, proper volume and surface comparisons can only be based on numbers derived through use of the same list of radii.

In the last column of the table we give a recent set of VDW radii that has been carefully optimized for use in volume and packing calculations. It is derived from analysis of the most common distances between atoms in small-molecule crystal structures in the Cambridge Structural Database (Rowland & Taylor, 1996; Tsai *et al.*, 1999).

22.1.1.4.2. *The probe radius*

A series of surfaces can be described by using a probe sphere with a specified radius. Since this is to be a convenient mathematical construct in calculation, any numerical value may be chosen with no necessary relation to physical reality. Some commonly used examples are listed in Table 22.1.1.2.

The solvent-accessible surface is intended to be a close approximation to what a water molecule as a probe might 'see' (Lee & Richards, 1971). However, there is no uniform agreement on what the proper water radius should be. Usually it is chosen to be about 1.4 Å.

22.1.1.5. *Application of geometry calculations: the measurement of packing*

22.1.1.5.1. *Using volume to measure packing efficiency*

Volume calculations are principally applied in measuring packing. This is because the packing efficiency of a given atom is simply the ratio of the space it could minimally occupy to the space that it actually does occupy. As shown in Fig. 22.1.1.8, this ratio can be expressed as the VDW volume of an atom divided by its Voronoi volume (Richards, 1974, 1985; Richards & Lim, 1994). (Packing efficiency also sometimes goes by the equivalent terms 'packing density' or 'packing coefficient'.) This simple definition masks considerable complexities – in particular, how does one determine the volume of the VDW envelope (Petitjean, 1994)? This requires knowledge of what the VDW radii of atoms are, a subject on which there is not universal agreement (see above), especially for water molecules and polar atoms (Gerstein *et al.*, 1995; Madan & Lee, 1994).

Knowing that the absolute packing efficiency of an atom is a certain value is most useful in a comparative sense, *i.e.* when comparing equivalent atoms in different parts of a protein structure. In taking a ratio of two packing efficiencies, the VDW envelope volume remains the same and cancels. One is left with just the ratio

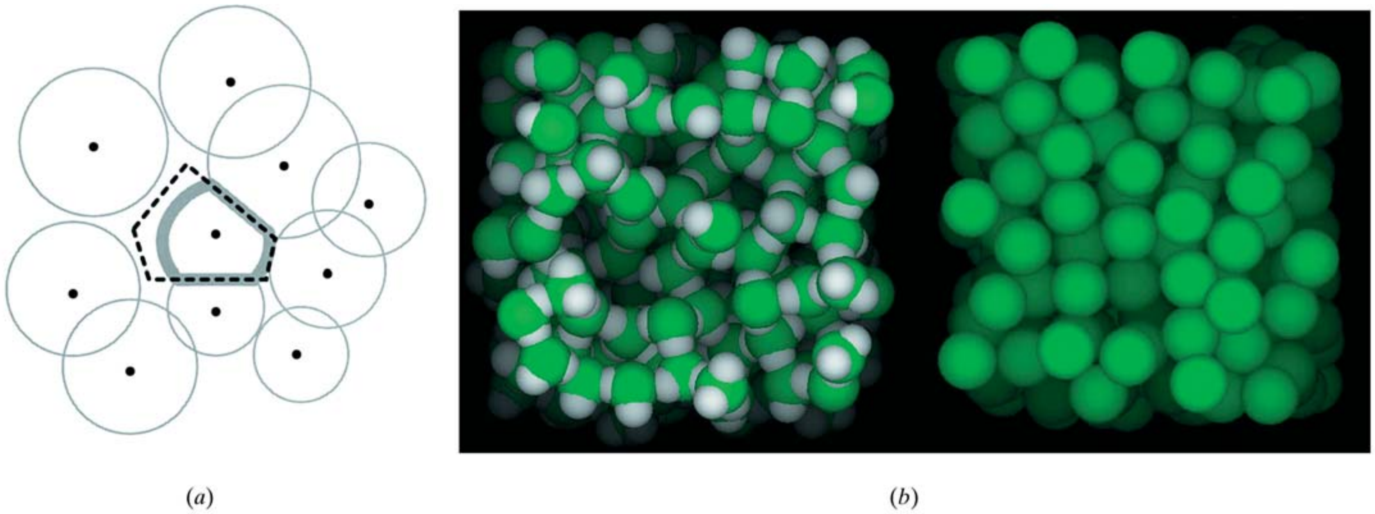

Fig. 22.1.1.8. Packing efficiency. (*a*) The relationship between Voronoi polyhedra and packing efficiency. Packing efficiency is defined as the volume of an object as a fraction of the space that it occupies. (It is also known as the 'packing coefficient' or 'packing density'.) In the context of molecular structure, it is measured by the ratio of the VDW volume ($V_{\rm VDW}$, shown by a light grey line) and Voronoi volume ($V_{\rm Vor}$, shown by a dotted line). This calculation gives absolute packing efficiencies. In practice, one usually measures a relative efficiency, relative to the atom in a reference state: $(V_{\rm VDW}/V_{\rm Vor})/[V_{\rm VDW}/V_{\rm Vor}({\rm ref})]$. Note that in this ratio the unchanging VDW volume of an atom cancels out, leaving one with just a ratio of two Voronoi volumes. Perhaps more usefully, when one is trying to evaluate the packing efficiency P at an interface, one computes $P = p\sum V_i/\sum v_i$, where p is packing efficiency of the reference data set (usually 0.74), V_i is the actual measured volume of each atom i at the interface and v_i is the reference volume corresponding to the type of atom i. (*b*) A graphical illustration of the difference between tight packing and loose packing. Frames from a simulation are shown for liquid water (left) and for liquid argon, a simple liquid (right). Owing to its hydrogen bonds, water is much less tightly packed than argon (packing efficiency of 0.35 *versus* ~0.7). Each water molecule has only four to five nearest neighbours while each argon atom has about ten.

Table 22.1.1.3. *Standard residue volumes*

The mean standard volume, the standard deviation about the mean and the frequency of occurrence of each residue in the protein core are given. Considering cysteine (Cyh, reduced) to be chemically different from cystine (Cys, involved in a disulfide and hence oxidized) gives 21 different residues. These residue volumes are adapted from the *ProtOr* parameter set (also known as the BL+ set) in Tsai *et al.* (1999) and Tsai *et al.* (2001). For this set, the averaging is done over 87 representative high-resolution crystal structures, only buried atoms not in contact with ligands are selected, the radii set shown in the last column of Table 22.1.1.1 is used and the volumes are computed in the presence of the crystal water. The frequencies for buried residues are from Harpaz *et al.* (1994).

Residue	Volume (Å^3)	Standard deviation (Å^3)	Frequency (%)
Ala	89.3	3.5	13
Val	138.2	4.8	13
Leu	163.1	5.8	12
Gly	63.8	2.7	11
Ile	163.0	5.3	9
Phe	190.8	4.8	6
Ser	93.5	3.9	6
Thr	119.6	4.2	5
Tyr	194.6	4.9	3
Asp	114.4	3.9	3
Cys	102.5	3.5	3
Pro	121.3	3.7	3
Met	165.8	5.4	2
Trp	226.4	5.3	2
Gln	146.9	4.3	2
His	157.5	4.3	2
Asn	122.4	4.6	1
Glu	138.8	4.3	1
Cyh	112.8	5.5	1
Arg	190.3	4.7	1
Lys	165.1	6.9	1

of space that an atom occupies in one environment to what it occupies in another. Thus, for the measurement of packing, standard reference volumes are particularly useful. Recently calculated values of these standard volumes are shown in Tables 22.1.1.3 and 22.1.1.4 for atoms and residues (Tsai *et al.*, 1999).

In analysing molecular systems, one usually finds that close packing is the default (Chandler *et al.*, 1983), *i.e.* atoms pack like billiard balls. Unless there are highly directional interactions (such as hydrogen bonds) that have to be satisfied, one usually achieves close packing to optimize the attractive tail of the VDW interaction. Close-packed spheres of the same size have a packing efficiency of ~0.74. Close-packed spheres of different size are expected to have a somewhat higher packing efficiency. In contrast, water is not close-packed because it has to satisfy the additional constraints of hydrogen bonding. It has an open, tetrahedral structure with a packing efficiency of ~0.35. (This difference in packing efficiency is illustrated in Fig. 22.1.1.8*b*)

22.1.1.5.2. *The tight packing of the protein core*

The protein core is usually considered to be the atoms inaccessible to solvent *i.e.* with an accessible surface area of zero or a very small number, such as 0.1 Å^2. Packing calculations on the protein core are usually done by calculating the average volumes of the buried atoms and residues in a database of crystal structures. These calculations were first done more than two decades ago (Chothia & Janin, 1975; Finney, 1975; Richards, 1974). The initial calculations revealed some important facts about protein structure. Atoms and residues of a given type inside proteins have a roughly constant (or invariant) volume. This is because the atoms inside proteins are packed together fairly tightly, with the protein interior better resembling a close-packed solid than a liquid or gas. In fact, the packing efficiency of atoms inside proteins is roughly as expected for the close packing of hard spheres (0.74).

More recent calculations measuring the packing in proteins (Harpaz *et al.*, 1994; Tsai *et al.*, 1999) have shown that the packing inside of proteins is somewhat tighter (by ~4%) than that observed initially and that the overall packing efficiency of atoms in the protein core is greater than that in crystals of organic molecules. When molecules are packed this tightly, small changes in packing efficiency are quite significant. In this regime, the limitation on close packing is hard-core repulsion, which is expected to have a twelfth power or exponential dependence, so even a small change is energetically quite substantial. Furthermore, the number of allowable configurations that a collection of atoms can assume without core overlap drops off very quickly as these atoms approach the close-packed limit (Richards & Lim, 1994).

The exceptionally tight packing in the protein core seems to require a precise jigsaw puzzle-like fit of the residues. This appears to be the case for the majority of atoms inside of proteins (Connolly, 1986). The tight packing in proteins has, in fact, been proposed as a quality measure in protein crystal structures (Pontius *et al.*, 1996). It is also believed to be a strong constraint on protein flexibility and motions (Gerstein *et al.*, 1993; Gerstein, Lesk & Chothia, 1994). However, there are exceptions, and some studies have focused on these, showing how the packing inside proteins is punctuated by defects, or cavities (Hubbard & Argos, 1994, 1995; Kleywegt & Jones, 1994; Kocher *et al.*, 1996; Rashin *et al.*, 1986; Richards, 1979; Williams *et al.*, 1994). If these defects are large enough, they can contain buried water molecules (Baker & Hubbard, 1984; Matthews *et al.*, 1995; Sreenivasan & Axelsen, 1992).

Surprisingly, despite the intricacies of the observed jigsaw puzzle-like packing in the protein core, it has been shown that one can simply achieve the 'first-order' aspect of this, getting the overall volume of the core right rather easily (Gerstein, Sonnhammer & Chothia, 1994; Kapp *et al.*, 1995; Lim & Ptitsyn, 1970). This has to do with simple statistics for summing random numbers and the fact that the distribution of sizes for amino acids usually found inside proteins is rather narrow (Table 22.1.1.3). In fact, the similarly sized residues Val, Ile, Leu and Ala (with volumes 138, 163, 163 and 89 Å^3) make up about half of the residues buried in the protein core. Furthermore, aliphatic residues, in particular, have a relatively large number of adjustable degrees of freedom per Å^3, allowing them to accommodate a wide range of packing geometries. All of this suggests that many of the features of protein sequences may only require random-like qualities for them to fold (Finkelstein, 1994).

22.1.1.5.3. *Looser packing on the surface*

Measuring the packing efficiency inside the protein core provides a good reference point for comparison, and a number of other studies have looked at this in comparison with other parts of the protein. The most obvious thing to compare with the protein inside is the protein outside, or surface. This is particularly interesting from a packing perspective, since the protein surface is covered by water, and water is packed much less tightly than protein and in a distinctly different fashion. (The tetrahedral packing geometry of water molecules gives a packing efficiency of less than half that of hexagonal close-packed solids.)

Table 22.1.1.4. *Standard atomic volumes*

Tsai *et al.* (1999) and Tsai *et al.* (2001) clustered all the atoms in proteins into the 18 basic types shown below. Most of these have a simple chemical definition, *e.g.* '=O' are carbonyl carbons. However, some of the basic chemical types, such as the aromatic CH group ('≥CH'), need to be split into two subclusters (bigger and smaller), as is indicated by the column labelled 'Cluster'. Volume statistics were accumulated for each of the 18 types based on averaging over 87 high-resolution crystal structures (in the same fashion as described for the residue volumes in Table 22.1.1.3). No. is the number of atoms averaged over. The final column ('Symbol') gives the standardized symbol used to describe the atom in Tsai *et al.* (1999). The atom volumes shown here are part of the *ProtOr* parameter set (also known as the BL+ set) in Tsai *et al.* (1999).

Atom type	Cluster	Description	Average volume (Å^3)	Standard deviation (Å^3)	No.	Symbol
>C=	Bigger	Trigonal (unbranched), aromatics	9.7	0.7	4184	C3H0b
>C=	Smaller	Trigonal (branched)	8.7	0.6	11876	C3H0s
≥CH	Bigger	Aromatic, CH (facing away from main chain)	21.3	1.9	2063	C3H1b
≥CH	Smaller	Aromatic, CH (facing towards main chain)	20.4	1.7	1742	C3H1s
>CH–	Bigger	Aliphatic, CH (unbranched)	14.4	1.3	3642	C4H1b
>CH–	Smaller	Aliphatic, CH (branched)	13.2	1.0	7028	C4H1s
–CH$_2$–	Bigger	Aliphatic, methyl	24.3	2.1	1065	C4H2b
–CH$_2$–	Smaller	Aliphatic, methyl	23.2	2.3	4228	C4H2s
–CH$_3$		Aliphatic, methyl	36.7	3.2	3497	C4H3u
>N–		Pro N	8.7	0.6	581	N3H0u
>NH	Bigger	Side chain NH	15.7	1.5	446	N3H1b
>NH	Smaller	Peptide	13.6	1.0	10016	N3H1s
–NH$_2$		Amino or amide	22.7	2.1	250	N3H2u
–NH$_3^+$		Amino, protonated	21.4	1.2	8	N4H3u
=O		Carbonyl oxygen	15.9	1.3	7872	O1H0u
–OH		Alcoholic hydroxyl	18.0	1.7	559	O2H1u
–S–		Thioether or –S–S–	29.2	2.6	263	S2H0u
–SH		Sulfhydryl	36.7	4.2	48	S2H1u

Calculations based on crystal structures and simulations have shown that the protein surface has intermediate packing, being packed less tightly than the core but not as loosely as liquid water (Gerstein & Chothia, 1996; Gerstein *et al.*, 1995). One can understand the looser packing at the surface than in the core in terms of a simple trade-off between hydrogen bonding and close packing, and this can be explicitly visualized in simulations of the packing in simple toy systems (Gerstein & Lynden-Bell, 1993*a*,*b*).

22.1.2. Molecular surfaces: calculations, uses and representations

(M. S. Chapman and M. L. Connolly)

22.1.2.1. *Introduction*

22.1.2.1.1. *Uses of surface-area calculations*

Interactions between molecules are most likely to be mediated by the properties of residues at their surfaces. Surfaces have figured prominently in functional interpretations of macromolecular structure. Which residues are most likely to interact with other molecules? What are their properties: charged, polar, or hydrophobic? What would be the estimated energy of interaction? How do the shapes and properties complement one another? Which surfaces are most conserved among a homologous family? At the centre of these questions that are often asked at the start of a structural interpretation lies the calculation of the molecular and/or accessible surfaces.

Surface-area calculations are used in two ways. Graphical surface representations help to obtain a quick intuitive understanding of potential molecular functions and interactions through visualization of the shape, charge distribution, polarity, or sequence conservation on the molecular surface (for example). Quantitative calculations of surface area are used *en route* to approximations of the free energy of interactions in binding complexes.

Part of this subject area was the topic of an excellent review by Richards (1985), to which the reader is referred for greater coverage of many of the methods of calculation. This review will attempt to incorporate more recent developments, particularly in the use of graphics, both realistic and schematic.

22.1.2.1.2. *Molecular, solvent-accessible and occluded surface areas*

The concept of molecular surface derives from the behaviour of non-bonded atoms as they approach each other. As indicated by the Lennard–Jones potential, strong unfavourable interactions of overlapping non-bonding electron orbitals increase sharply according to $1/r^{12}$, and atoms behave almost as if they were hard spheres with *van der Waals* radii that are characteristic for each atom type and nearly independent of chemical context. Of course, when orbitals combine in a covalent bond, atoms approach much more closely. Lower-energy attractions between atoms, such as hydrogen bonds or aromatic ring stacking, lead to modest reductions in the distance of closest approach. The van der Waals surface is the area of a volume formed by placing van der Waals spheres at the centre of each atom in a molecule.

Non-bonded atoms of the same molecule contact each other over (at most) a very small proportion of their van der Waals surface. The surface is complicated with gaps and crevices. Much of this surface is inaccessible to other atoms or molecules, because there is insufficient space to place an atom without resulting in forbidden overlap of non-bonded van der Waals spheres (Fig. 22.1.2.1). These crevices are excluded in the *molecular surface area*. The molecular

references