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).

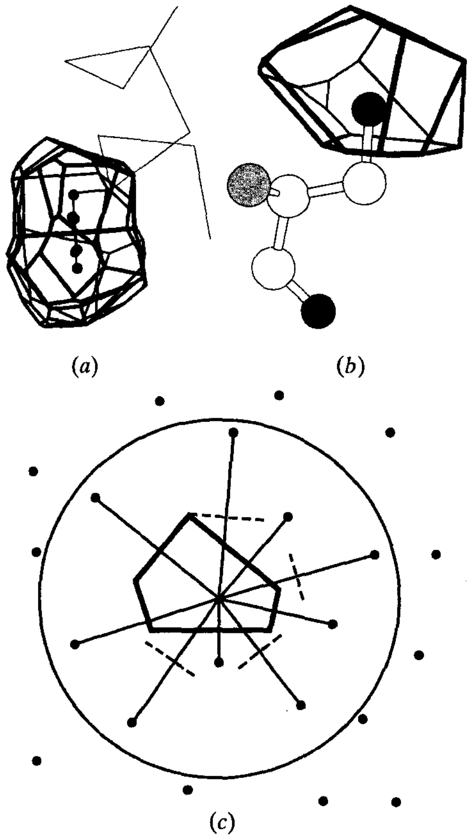

Fig. 22.1.1.1. The Voronoi construction in two and three dimensions. Representative Voronoi polyhedra from 1CSE (subtilisin) are shown. (*a*) Six polyhedra around the atoms in a Phe ring. (*b*) A single polyhedron around the side-chain hydroxyl oxygen (OG) of a serine. (*c*) A schematic showing the construction 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).

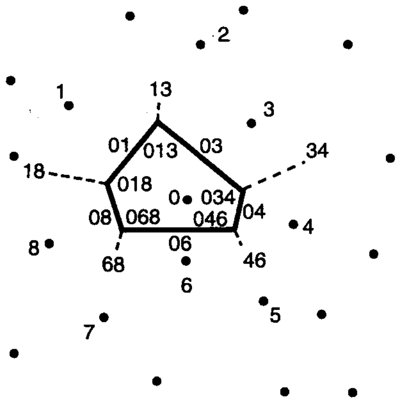

Fig. 22.1.1.2. Labelling parts of Voronoi polyhedra. The central atom is

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 $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$ 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).

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:

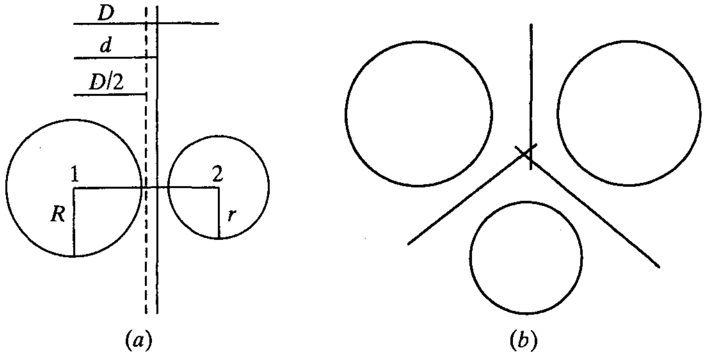

Fig. 22.1.1.3. Positioning of the dividing plane. (*a*) The dividing plane is positioned at a distance d from the larger atom with respect to radii of the larger atom (R) and the smaller atom (r) and the total separation between the atoms (D). (*b*) Vertex error. One problem with using method B is that 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.

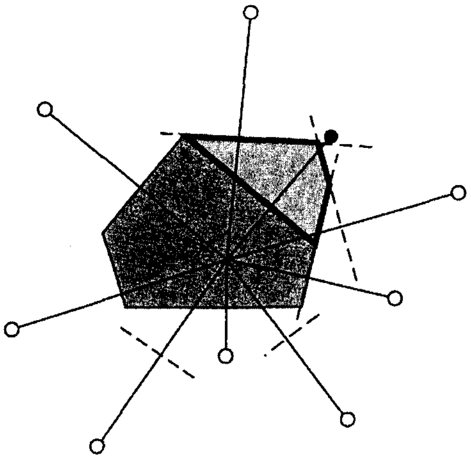

Fig. 22.1.1.4. The 'chopping-down' method of polyhedra construction. This is necessary when using method B for plane positioning, since one can no longer solve for the position of vertices. One starts with a large tetrahedron around the central atom and then 'chops it down' by removing vertices that 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

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

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.

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).

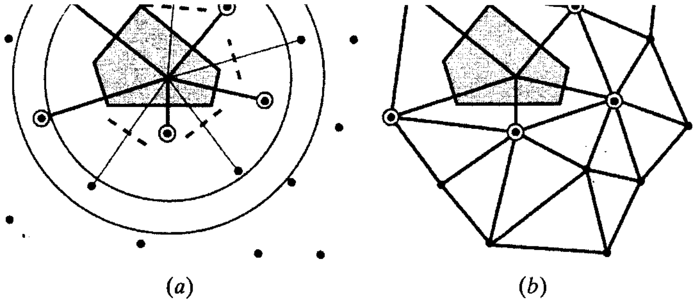

Fig. 22.1.1.5. Delaunay triangulation and its relation to the Voronoi construction. (*a*) A standard schematic of the Voronoi construction. The atoms used to define the Voronoi planes around the central atom are circled. Lines connecting these atoms to the central one are part of the Delaunay triangulation, 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 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).

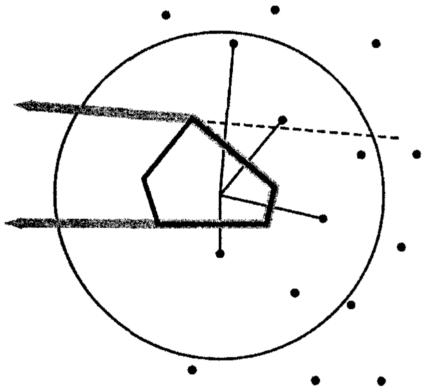

Fig. 22.1.1.6. The problem of the protein surface. This figure shows the difficulty in constructing Voronoi polyhedra for atoms on the protein surface. If all the water molecules near the surface are not resolved in a crystal structure, one often does not have enough neighbours to define a closed polyhedron. 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).

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

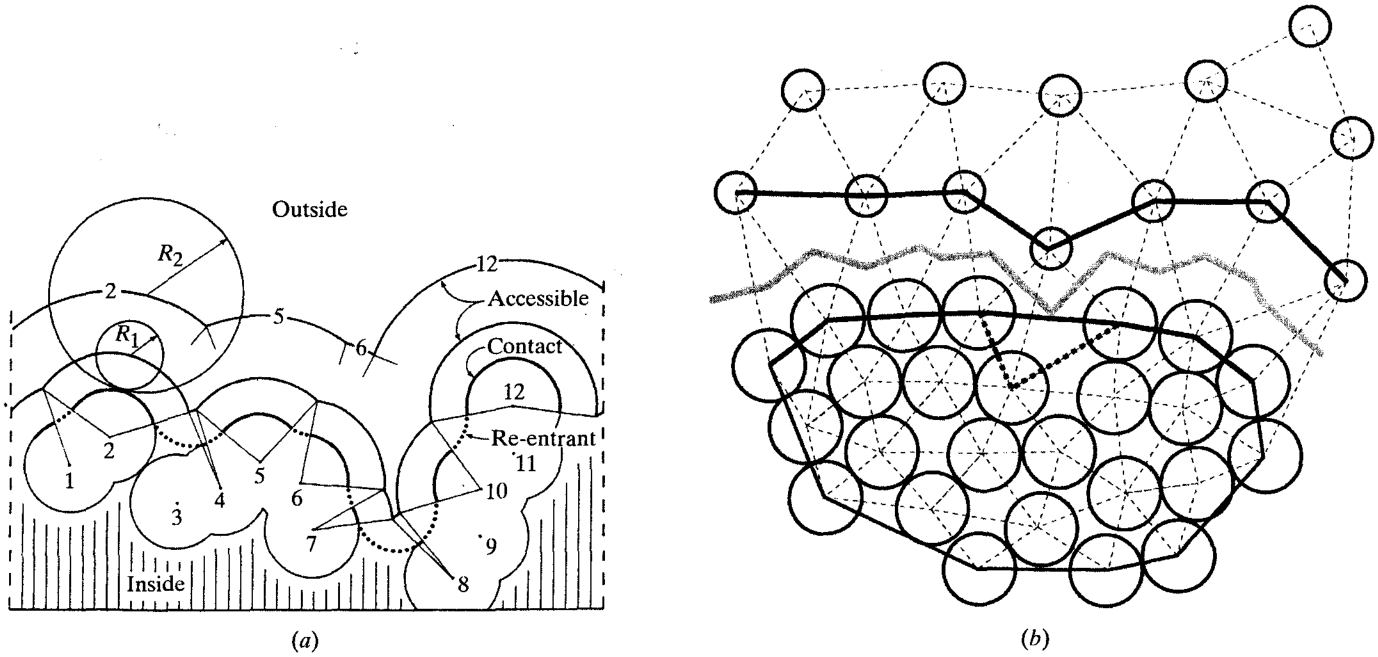

Fig. 22.1.1.7. Definitions of the protein surface. (*a*) The classic definitions of protein surface in terms of the probe sphere, the accessible surface and the molecular surface. (This figure is adapted from Richards, 1977). (*b*) Voronoi polyhedra and Delaunay triangulation can also be used to define 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
⩾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

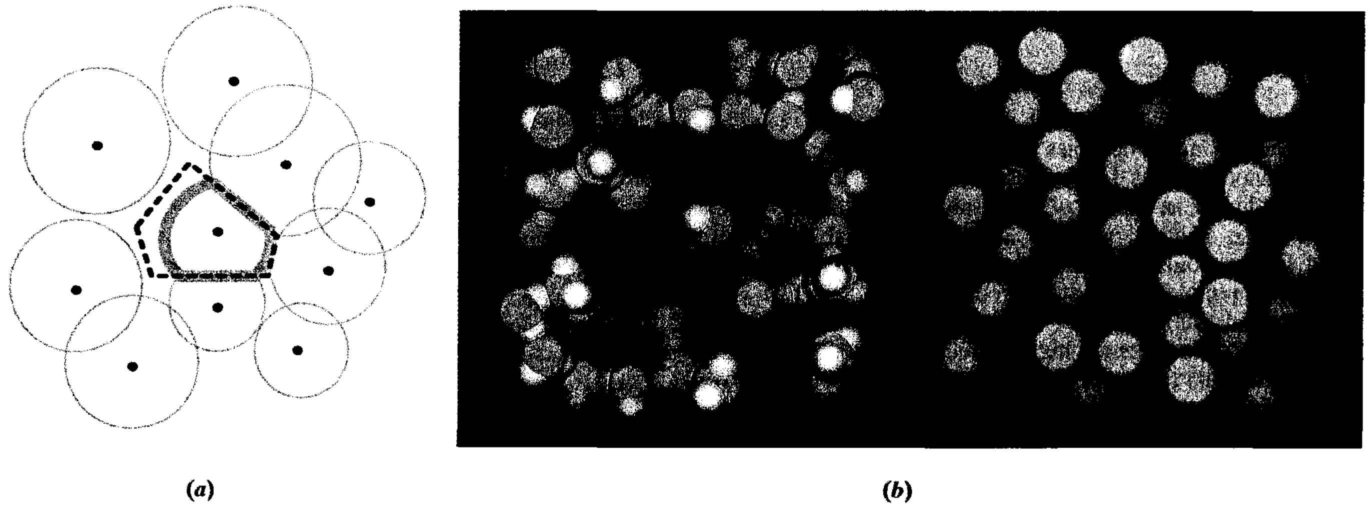

Fig. 22.1.1.8. Packing efficiency. (*a*) The relationship between Voronoi polyhedra and packing efficiency. Packing efficiency is defined as the volume of an object as a fraction of the space that it occupies. (It is also known as the 'packing coefficient' or 'packing density'.) In the context of molecular 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

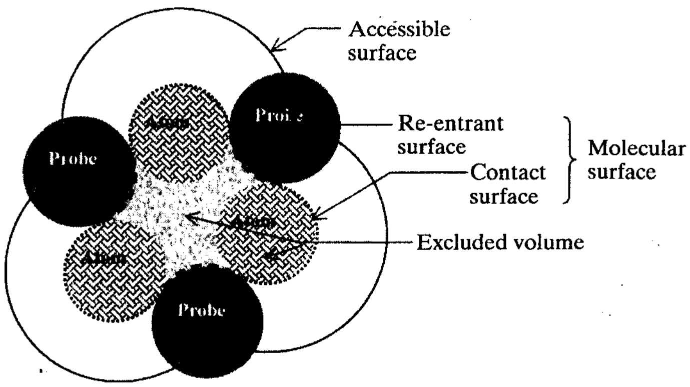

Fig. 22.1.2.1. Surfaces in a plane cut through a hypothetical molecule. The molecular surface consists of the sum of the atomic surfaces that can be contacted by solvent molecules and the surface of the space between atoms from which solvent molecules are excluded. The solvent-accessible surface is the surface formed by the set of the centres of spheres that are in closest contact with the molecular surface.

surface area, also known as the solvent-excluding surface, is the outer surface of the volume from which solvent molecules are excluded. Strictly, this would depend on the orientation of non-spherically symmetric solvents such as water. However, since hydrogen atoms are smaller than oxygen atoms, for current purposes it is sufficient to consider water as a sphere with a radius of 1.4 to 1.7 Å, approximating the 'average' distance from the centre of the oxygen atom to the van der Waals surface of water. The practical definition of the molecular surface is, then, the area of the volume excluded to a spherical probe of 1.4 to 1.7 Å radius.

As an aside, it is important to note that surface-area calculations depend on inexact parameterization. For example, there is no radius of any hard-sphere model that can give a realistic representation of the solvent. Furthermore, the choice of van der Waals radii can depend on whether the distance of zero or minimum potential energy is estimated and the potential-energy function or experimental data used. (Tables of common values are given by Gerstein & Richards in Section 22.1.1.) Thus, calculations of molecular and accessible surfaces are approximate. However, when the errors are averaged over large areas of a macromolecule, the numbers can be precise enough to give important insights into function.

Fig. 22.1.2.1 shows that the molecular surface consists of two components. The contact surface is part of the van der Waals surface. The re-entrant surface encloses the interstitial volume and has components that are the exterior surfaces of atoms (contact surface) and parts of the surfaces of probes placed in positions where they are in contact with van der Waals surfaces of two or more atoms (re-entrant surface).

The occluded molecular surface is an approximate complement to the solvent-accessible surface. It is the part of the surface that would be inaccessible to solvent because of steric conflict with neighbouring macromolecular atoms. It is an approximation in that current calculations use van der Waals surfaces, ignoring the differences between atomic and re-entrant surfaces (see below), and the volume of the probe is not fully accounted for (Pattabiraman *et al.*, 1995). Occluded area is defined as the atomic area whose normals cannot be extended 2.8 Å (the presumptive diameter of a water molecule) without intersecting the van der Waals volume of another atom. This crude approximation to the surface that is inaccessible to water not only increases the speed of calculation, but enables surface areas to be partitioned between the atoms. It is used primarily to evaluate model protein structures by comparing the fraction of each amino acid's surface area that is occluded with that calculated for the same residue types in a database of accurate structures.

22.1.2.1.3. *Hydration surface*

Whether graphically displaying a molecule or examining potential docking interactions, it is usually the molecular surface or solvent-accessible surface that is used. However, macromolecules also interact through the small (solvent) molecules that are more or less tightly bound (Gerstein & Lynden-Bell, 1993*c*). There is a gradation of how tightly solvent molecules are bound and how many are bound around different side chains. With dynamics simulations, Gerstein & Lynden-Bell (1993*c*) showed that the second hydration shell was a reasonable, practical 'average' limit to which water atoms should be considered significantly perturbed by the protein. They defined a hydration surface as the surface of this second shell and presented evidence that it approximates the boundary between bound and bulk solvent. They presented calculations that showed that molecules interact significantly when their hydration surfaces interact, and not just when they are close enough for their molecular surfaces to form contacts. It may be computationally impractical to perform the simulations required to calculate the hydration surfaces of many proteins, but this work reminds us that energetically significant interactions occur over a wider area than the commonly computed contact molecular-surface area.

22.1.2.1.4. *Hydrophobicity*

The hydrophobic effect (Kauzmann, 1959; Tanford, 1997) has its origins in unfavourable entropic terms for water molecules immediately surrounding a hydrophobic group. In the bulk solvent, each water molecule can be oriented in a variety of ways with favourable hydrogen bonding. At the interface with a hydrophobic group, hydrogen bonds are possible only in some directions, with some configurations of the water molecules. When a hydrophobic group is embedded in water, the surrounding solvent molecules have a more restricted set of hydrogen-bonding configurations, resulting in an unfavourable entropic term. The magnitude of the entropic term should be proportional to the number of solvent molecules immediately surrounding the hydrophobic group. This integer number can be considered very approximately proportional to the area of the surface made by the centres of the set of possible solvent probes contacting the solute, *i.e.* the *solvent-accessible surface area* (Fig. 22.1.2.1). When large areas are considered, summed over many hydrophobic atoms, the errors of this non-integer approximation are insignificant. It is now common practice to estimate the hydrophobic effect free-energy contribution by multiplying the change in macromolecular surface area by an energy per unit area [$\sim$80 J mol^{-1} Å^{-2} (Richards, 1985), but see also below].

22.1.2.2. *Calculation of surface area and energies of interaction*

22.1.2.2.1. *Introduction*

The first method to be discussed allows the calculation of an accessible surface. The first method for calculating molecular surface involved raining water down on a model of a macromolecule and constructing a surface by making a net under the spheres in their landing positions (Greer & Bush, 1978). This ignored overhangs and was replaced by the dot surface method. More recently, methods were developed to make polyhedral surfaces of triangles by contouring between lattice points or by delimiting with arcs the spherical and toroidal surfaces and then subdividing the piece-wise quartic molecular surface. The surface is then composed of patches whose areas can be precisely integrated. van der Waals surfaces consist of convex spherical triangles whose areas can be estimated by the Gauss–Bonnet theorem. Re-entrant surfaces are comprised of concave spherical triangles whose areas can be similarly estimated and toroidal saddle-shaped patches whose areas can be calculated by analytical geometry and calculus.

22.1.2.2.2. *Lee & Richards planar slices*

The first method for calculating the accessible surface area overlaid the molecule on a regular stack of finely spaced parallel planes (Lee & Richards, 1971). The advantage of this method was the ease with which the area could be calculated. The intersection of the atomic surfaces with the planes were circular arcs whose lengths were readily calculated and multiplied by the planar spacing to give an approximation to the surface area. Programs that are currently distributed use more sophisticated methods.

22.1.2.2.3. *Connolly dot surface algorithm*

A molecular dot surface is a smooth envelope of points on the molecular surface. A probe sphere is placed at a set of approximately evenly spaced points so that the probe and van der Waals surfaces of a given atom are tangential. If the probe sphere does not overlap any other atom, the point is designated as surface. To define the re-entrant surface, sphere centres are also sampled that are tangential to both van der Waals spheres of a pair of neighbouring atoms and are equidistant from the interatomic axis. Arcs are then drawn between surface points and the arcs are subdivided into a set of finely spaced points to define the re-entrant surface. Similarly, spheres contacting triplets of neighbouring atoms are tested, and approximately evenly spaced points within the concave triangle defined by the three contact points are added to the re-entrant surface.

22.1.2.2.4. *Marching-cube algorithm*

This is conceptually the simplest method and is used in the program *GRASP* (Nicholls *et al.*, 1991). First, grid points of a cubic lattice overlaid on the molecule are segregated into 'interior' and 'exterior' as follows. All points farther from an atom than the sum of the van der Waals radius and a probe radius are flagged as external. External points with an internal neighbour are flagged as an approximate 'accessible surface'. All grid points falling within probe spheres centred at each surface point now join the set of exterior points. Points that remain 'interior' define the volume enclosed by the molecular surface.

All that remains is to contour the molecular surface that lies between interior and exterior grid points. It is a little complicated in three dimensions and is achieved by the marching-cube algorithm. Cubes containing adjacent grid points that are both interior and exterior are used to define potential polyhedral vertices. Triangles are defined by joining the midpoints of unit-cell edges that have one interior and one exterior point. The triangles are joined at their edges in a consistent manner to create a polyhedral surface.

22.1.2.2.5. *Complete and connected rolling algorithms*

Several algorithms start by dividing the surface into regions within which the surface is smooth and continuous. The surface can be efficiently described in terms of a set of arcs and their start and end points. In complete rolling, the probe is placed in all possible positions at which it contacts the van der Waals spheres of three neighbouring atoms. Those surrounding the same atom are paired as the start and end points of an arc. The complete rolling algorithm does not distinguish outer and inner (cavity) surfaces. In the connected rolling algorithm, the process starts at a triple contact point that is far from the centre of mass and therefore likely to be external. The probe is then rolled only along crevices between two atoms, pursuing all alternatives, stopping each pathway only when the probe returns to a place that has already been probed. This algorithm therefore produces only the outer surface.

22.1.2.2.6. *Analytic surface calculations and the Gauss–Bonnet theorem*

An analytical method was also proposed for calculating approximate accessible areas (Wodak & Janin, 1980). It assumed random distributions of neighbouring atoms, but this can be a sufficient approximation when calculating the area of an entire molecule. The areas of spherical and toroidal pieces of surface can be calculated exactly by analytic and differential geometry (Richmond, 1984; Connolly, 1983). An advantage of analytical expressions over the prior numerical approximations is that analytical derivatives of the areas can be calculated, albeit with significant difficulty. This then provides the opportunity to optimize atomic positions with respect to surface area. Pseudo-energy functions that approximate the hydrophobic contribution to free energy with a term proportional to the accessible surface area (Richards, 1977) can therefore be incorporated in energy-minimization programs. Although rigorous, these methods are computationally cumbersome and are not used in all energy-minimization routines. Incorporation of solvent effects may become more universal with the Gaussian atom approximations discussed below.

22.1.2.2.7. *Approximations to the surface*

The methods discussed above are computationally quite cumbersome, especially if they need to be repeated many times. Thus, they are not well suited to comparisons of many structures. They are also not well suited to the calculation of surface-area-dependent energy terms during dynamics simulation or energy minimization, which require the calculation of the derivatives of the surface area with respect to atomic position. It has been argued by several (including A. Nicholls and K. Sharp, personal communications) that simplifying approximations to the surface-area calculations are in order, because the common uses of surface area already embody crude *ad hoc* approximations, such as non-integer numbers of spherical solvent molecules.

In the treatments discussed earlier, the volume of the protein is (implicitly) described by a set of overlapping step functions that have a constant value if close enough to an atom, or zero if not. Several authors have replaced these step functions with continuous spherical Gaussian functions centred on each atom (Gerstein, 1992; Grant & Pickup, 1995) in treatments reminiscent of Ten Eyck's electron-density calculations (Ten Eyck, 1977). This speeds up the calculation and also facilitates the calculation of analytical derivatives of the surface area. A surface can be calculated for graphical display by contouring the continuous function at an appropriate threshold. The final envelope can be modified by using iterative procedures that fill cavities and crevices that are (nearly) surrounded by protein atoms (Gerstein, 1992).

22.1.2.2.8. *Extended atoms account for missing hydrogen atoms*

Structures of macromolecules determined by X-ray crystallography rarely reveal the positions of the hydrogen atoms. It is, of course, possible to add explicit hydrogen atoms at the stereochemically most likely positions, but this is rarely done for surface-area calculations. Instead, their average effect is approximately and implicitly accounted for by increasing the heteroatom van der Waals radius by 0.1 to 0.3 Å. (It is not usual to smear atoms to account for thermal motion.)

22.1.2.3. *Estimation of binding energies*

22.1.2.3.1. *Hydrophobicity*

As previously introduced, hydrophobic energies result primarily from the increased entropy of water molecules at the macromolecule–solvent interface and can be estimated from the accessible surface area. A number of different constants relating area to free energy of transfer from a hydrophobic to aqueous environment have been proposed in the range of 67 to 130 J mol^{-1} Å^{-2} (Reynolds *et al.*, 1974; Chothia, 1976; Hermann, 1977; Eisenberg & McLachlan, 1986), but if a single value is to be used for all of the protein surface, the consensus among crystallographers has been about 80 J mol^{-1} Å^{-2} (Richards, 1985).

There are two widely used enhancements of the basic method. Atomic solvation parameters (ASPs, $\Delta\sigma$) remove the assumption that all protein atoms have equal potential influence on the hydrophobic free energy. Eisenberg & McLachlan (1986) determined separate ASPs for atom types C, N/O, O^{-}, N^{+} and S (treating hydrogen atoms implicitly) by fitting these constants to the experimentally determined octanol/water relative transfer free energies of the 20 amino-acid side chains of Fauchere & Pliska (1983), assuming standard conformations of the side chains. A much improved free energy change of solvation can then be estimated from $\Delta G = \sum_{\text{atoms } i} \Delta\sigma_i A_i$, where the summation is over all atoms with accessible area A and $\Delta\sigma_i$ is specific for the atom type. Their estimates of ASPs are given in Table 22.1.2.1. Use of ASPs rather than a single value for all atoms makes substantial differences to the estimated free energies of association of macromolecular assemblies (Xie & Chapman, 1996). Through calculation of the overall energy of solvation, calculations with ASPs also allow discrimination between proposed structures that are correctly folded (with hydrophobic side chains that are predominantly internal) and those that are not (Eisenberg & McLachlan, 1986).

The work of Sharp *et al.* (1991) indicates that hydrophobicity depends not only on surface area, but curvature. Sharp *et al.* were trying to reconcile long-apparent differences between microscopic and macroscopic measurements of hydrophobicity (Tanford, 1979). Microscopic measurements, the basis of all of our preceding discussions, are derived from the partitioning of dilute solutes between solvents. Macroscopic values can come from the measurements of the surface tension between a liquid bulk of the molecule of interest and water. Macroscopic values for aliphatic carbons are much higher, ~302 J mol^{-1} Å^{-2}. Postulating that the entropic effects at the heart of hydrophobicity depended on the number of water molecules in contact with each other at the molecular surface (Nicholls *et al.*, 1991), Sharp *et al.* pointed out that not all surfaces were equivalent. Relative to a plane, concave solute surfaces would accommodate fewer solvent molecules neighbouring the molecular surface, whereas convex surfaces would accommodate more. Their treatment could be considered to be a second-order approximation to the number of interfacial solvent molecules, compared to the prior first-order consideration of only area.

To calculate the curvature of point *a* on the accessible surface (relative to that of a plane), a sphere of twice the solvent radius is drawn (Nicholls *et al.*, 1991). This represents the locus of the centres of solvent molecules that could be in contact with a solvent at *a*. A curvature correction, *c*, is the proportion of points on the spherical surface that are inside the inaccessible volume, relative to that for a planar accessible surface ($\frac{1}{2}$). In calculating the free energy of transfer, each element of the accessible area is multiplied by its curvature correction. When this is done, the increasingly convex surfaces of small aliphatic molecules account for most of the discrepancy between microscopic and macroscopic hydrophobicities (Nicholls *et al.*, 1991). Furthermore, it emphasizes that, just by their shape, concave surfaces can become relatively hydrophobic. This has been clearly illustrated with *GRASP* surface representations (see below) in which the accessible surface is coloured according to the local curvature (Nicholls *et al.*, 1991). Consideration of curvature also indicates that the energy of macromolecular association is slightly less than it would otherwise be due to the generation of a concave collar at the interface between two binding macromolecules (Nicholls *et al.*, 1991).

Table 22.1.2.1. *The atomic solvation parameters of Eisenberg & McLachlan (1986)*

Atom	$\Delta\sigma$(atom) (J mol^{-1} Å^{-2})
C	67 (8)
N/O	−25 (17)
O^{-}	−101 (42)
N^{+}	−210 (38)
S	88 (42)

22.1.2.3.2. *Estimates of binding energies*

In a molecular association in which (as is often the case) hydrophobic interactions dominate, the binding energy can be estimated from the surfaces of the individual molecules that become buried upon association (Richards, 1985). The buried area is simply the sum of the surfaces of the two molecules (calculated independently) minus the surface of the complex, calculated as if one molecule. Usually, all heteroatoms are regarded as equivalent, and the buried area is multiplied by a uniform constant, say 80 J mol^{-1} Å^{-2} (Richards, 1985). It is only slightly more complicated to use the different ASPs (Eisenberg & McLachlan, 1986) for different atom types and/or to account for curvature (Nicholls *et al.*, 1991). It should be noted that in many crystal structures, the distinction between atom types in some side chains remains indeterminate, *e.g.* N and C in histidines, O and O^{-} in carboxylates, and N and N^{+} in arginines. In such cases, average values of the two ASPs can be used (Xie & Chapman, 1996). Such energy calculations have been put to several uses, including attempts to predict assembly and disassembly pathways for viral capsid assemblies (Arnold & Rossmann, 1990; Xie & Chapman, 1996, and citations therein).

22.1.2.3.3. *Other non-graphical interpretive methods using surface area*

Which are the amino acids most likely to interact with other molecules? It is reasonable to expect them to be surface-accessible. In determining which residues are most surface-exposed, it is necessary to partition molecular or accessible surfaces between atoms. Contact surfaces (Fig. 22.1.2.1) are atom specific. Re-entrant or accessible surfaces can be divided among surface atoms by proximity. Surface areas can then be summed over the atoms in a residue. Accessible surface areas are sometimes reported as *accessibilities* (Lee & Richards, 1971) – fractions of a maximum where the standard is evaluated from a tripeptide in which the residue of interest is surrounded by glycines. A different approach to assessing surface exposure is to ask what is the largest molecular fragment that could contact a given atom. This is commonly assayed by determining the largest sphere that can be placed tangentially to the van der Waals surface without intersecting any other atom. An alternative approach to locating functionally important surface regions was proposed in the mid-1980s, but is currently not used very often. The local irregularity of surface texture was characterized through measurement of the fractal dimension (Lewis & Rees, 1985).

Substrates, drugs and ligands often bind in clefts or pockets that are concave in shape. Conversely, it is the most exposed convex regions that are likely to be antigenic. The surface shape can be determined by placing a large (say 6 Å radius) sphere at each vertex of the polyhedral molecular surface. If more than half of the sphere's volume overlaps the molecular volume, then the surface is concave, while if less than half, the surface is convex.

Are there similarities in the shapes of surfaces at the interfaces of macromolecular complexes? For example, are there similarities between the shapes of evolutionary-related antigens or the hypervariable regions of antibodies that bind to them? Quantitative comparison of surface topologies is far from trivial, with questions of 3D alignment, the metrics to be used in quantifying topology *etc.* In addition to real differences between molecules, their surfaces may appear to differ due to the resolutions at which their structures were determined. Gerstein (1992) has proposed that comparisons be made in reciprocal space so that correlations can be judged as a function of resolution. Coordinates are aligned. Spherical Gaussian functions are placed at each atom, and an envelope is calculated at some threshold value and modified to remove cavities. Gerstein found that comparison of the envelope structure-factor vectors, obtained by Fourier transformation, led to a plausible classification of the hypervariable regions of known antibody structures.

22.1.2.4. *Graphical representations of shape and properties*

22.1.2.4.1. *Realistic*

22.1.2.4.1.1. *Shaded backbone*

With very large complexes, such as viruses, the surface features to be viewed are obvious at low resolution. In a very simple yet effective representation popularized by the laboratories of David Stuart and Jim Hogle, a Cα trace is 'depth cued' (shaded) according to the distance from the centre of mass (Acharya *et al.*, 1990; Fig. 1 for example). The impression of three dimensions probably results from the similarity of the shading to highlighting. The method is most effective for large complexes in which there are sufficient Cα atoms to give a dense impression of a surface.

22.1.2.4.1.2. *'Connolly' and solid polyhedral surfaces*

In one of the earliest surface graphical representations, dots were drawn for each Connolly surface dot, using vector-graphics terminals. With the improved graphics capability of modern computers, dot representations have been replaced by ones in which solid polyhedra are drawn with a large enough number of small triangular faces such that the surface appears smooth. These representations are clearer, because atoms in the foreground obscure those in the background.

22.1.2.4.1.3. *Photorealistic rendering*

Depth and three-dimensional relationships are most easily represented by stereovision or rotation of objects in real time on a computer screen. Graphics engines for interactive computers compromise quality for the speed necessary for interactive response, but simple depth cueing (combined with motion or stereo) is sufficient for good 3D representation. For still and/or non-stereo images more common in publications, more sophisticated rendering is helpful and possible now that speed is not a constraint. In *Raster3D* (Merritt & Bacon, 1997), multiple-light-source shading and highlighting is added, with individual calculations for each fine pixel. These are dependent on the directions of the normals to the surface, which are calculated analytically for spherical surfaces. More complicated surfaces, input as connected triangles, have surfaces rendered raster, pixel by pixel, by interpolating between the surface-normal vectors at the vertices of the surrounding triangle. Together, this leads to a high-quality smooth image that conveys much of the three-dimensionality of molecular surfaces.

22.1.2.4.1.4. *GRASP surfaces*

GRASP is currently perhaps the most popular program for the display of molecular surfaces. Readers are referred to the program documentation (Nicholls, 1992) or a paper that tangentially describes an early implementation (Nicholls *et al.*, 1991). The molecular or accessible surface is determined by the marching-cube algorithm. The surface is filled using methods that make modest compromises on photorealistic light reflection *etc.*, but take advantage of machine-dependent Silicon Graphics surface rendering to perform the display fast enough for interactive adjustment of the view.

The most powerful part of the program is the ability to colour according to properties mapped to the surface (see Fig. 22.1.2.2). These may be values of (say) electrostatic potential interpolated from a three-dimensional lattice. Much has been learned about many proteins from the potentials determined by solution of the Poisson–Boltzmann equation (Nicholls & Honig, 1991). The electrostatic complementarity of binding surfaces has often been readily apparent in ways that were not obvious from Coulombic calculations that ignore screening or from calculations and graphics representations that treat the charges of individual atoms as independent entities.

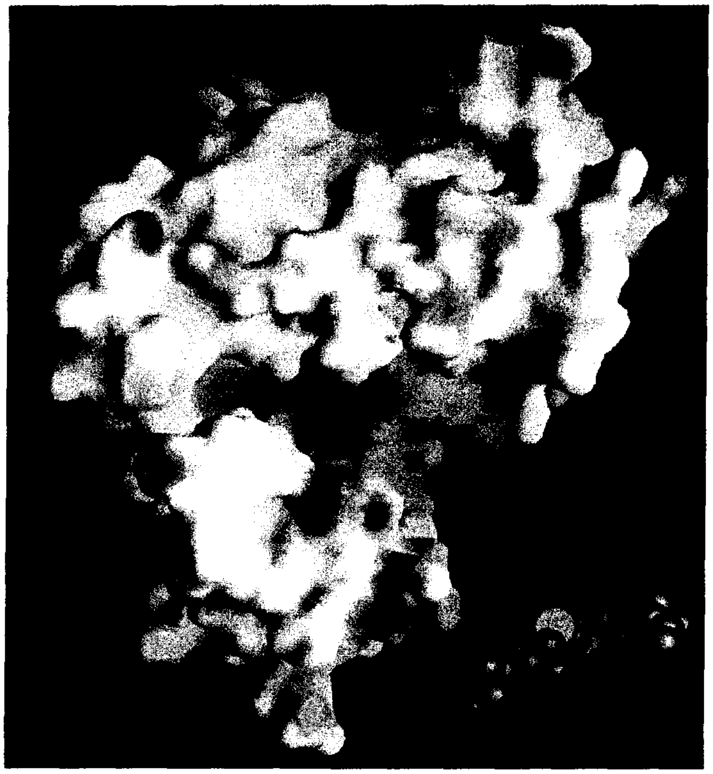

Fig. 22.1.2.2. *GRASP* example. The larger picture shows the molecular surface of arginine kinase (Zhou *et al.*, 1998) with the domains and a loop moved to the open configuration seen in a homologous creatine kinase structure (Fritz-Wolf *et al.*, 1996). The surface, coloured with positive charge blue and negative charge red, demonstrates that the active-site pocket (centre) is the most positively charged part of the structure. It complements the negatively charged phosphates of the transition-state analogue components that are shown, moved as a rigid body to the bottom right. They are shown in van der Waals representation, in which oxygens are red, carbons black and nitrogens blue.

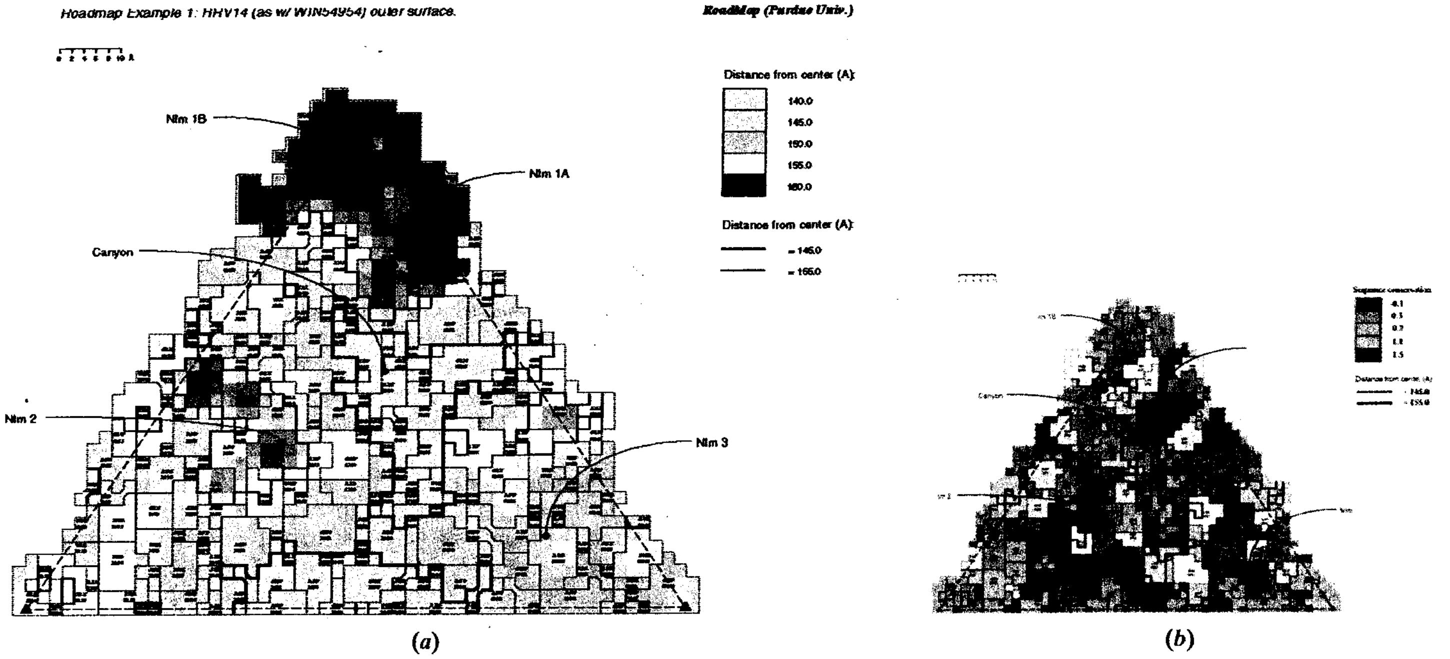
Roadmap Example 1: HRV14 (as w/ WIN54954) outer surface.
RoadMap (Purdue Univ.)
Distance from center (A):
140.0
145.0
150.0
155.0
160.0
Distance from center (A):
= 145.0
= 165.0
NIm 1B
NIm 1A
Canyon
NIm 2
NIm 3
Canyon
(a)
(b)

Fig. 22.1.2.3. (*a*) Solvent-accessible surface topology of a rhinovirus 14–drug complex (Kim *et al.*, 1993). The triangle shows one of the 60 symmetry-equivalent faces of an icosadeltahedron that constitute the entire virus surface. The surface is coloured and contoured according to distance from the centre of the virus, red being the most elevated. Residues are marked with dotted lines and labelled with residue type and number. A letter starting the residue label indicates a symmetry equivalent. The first numeral indicates the protein number (1 to 4), which is followed by the three-digit residue number. A depression, the 'canyon', is where the cellular receptor is bound (Olson *et al.*, 1993). The locations of the dominant neutralizing immunogenic (NIm) sites were determined through the sequencing of escape mutants (Sherry & Rueckert, 1985; Sherry *et al.*, 1986) and are labelled 'NIm'. (*b*) The same view is coloured according to sequence similarity (Palmenberg, 1989; Chapman, 1994), with blue being the most conserved rhinoviral amino acids and red being the most variable. Comparison of diagrams like these suggested the 'canyon hypothesis' (Rossmann, 1989). The prediction has proved largely true in that the sites of receptor attachment in several picornaviruses would be depressed areas whose sequences could be more highly conserved because they were partially inaccessible to antibodies and therefore not under the same selective pressure to mutate. In this and other applications, the schematic nature of these diagrams has helped in the collation of structure with data arising from the known phenotypes of site-directed or natural mutants. Part (*b*) is reproduced from Chapman (1993). Copyright (1993) The Protein Society. Reprinted with the permission of Cambridge University Press.

Many other properties can be mapped to the surface. These include properties of the atoms associated with that part of the surface (such as thermal factors), curvature of the surface calculated from adjacent atoms (Nicholls & Honig, 1991), or distance to the nearest part of the surface of an adjacent molecule. *GRASP* is now used to illustrate complicated molecular structures, in part because it also supports the superimposition of other objects over the molecular surface. These include the representation of molecules with CPK spheres and/or bonds, and the representation of electrostatic potentials with field lines, dipole vectors *etc.*

22.1.2.4.1.5. *Implementations in popular packages*

Commercial packages use variants of the methods discussed above. For example, surfaces are drawn in the *Insight* II molecular modelling system using the Connolly dot algorithm (Molecular Structure Corporation, 1995).

22.1.2.4.2. *Schematic and two-dimensional representations such as 'roadmap'*

For their work on viruses, Rossmann & Palmenberg (1988) introduced a highly schematic representation in which individual amino acids were labelled. The methods were extended by Chapman (1993) to other proteins and to the automatic display of features such as topology, sequence similarity and hydrophobicity. Roadmaps sacrifice a realistic impression of shape for the ability to show the locations and properties of constituent surface atoms or residues. This has been important in combining the power of structure and molecular biology in understanding function. Potential sites of mutation are readily identified without substantial molecular-graphics resources, and phenotypes of mutants are readily mapped to the surface and compared with the physiochemical properties to reveal structure–function correlations.

For a set of projection vectors, the intersection points with the first van der Waals (or solvent-accessible) surface of an atom are calculated by basic vector algebra. The atom is identified so that when the projection is mapped to a plane for display, the boundaries of each atom or amino acid can be determined. The atoms or amino acids can then be coloured, shaded, outlined, contoured, or labelled according to parameters that are either calculated from the coordinates (such as distance from the centre of mass), read from a file (such as sequence similarity), or follow properties that are dependent on the residue type (*e.g.* hydrophobicity) or atom type [*e.g.* atomic solvation parameters (Eisenberg & McLachlan, 1986)].

Several types of projections can be used. The simplest is similar to that used by most other surface-imaging programs. A set of parallel projection vectors is mapped to a 2D grid. An example is shown in Fig. 22.1.2.3. This view avoids distortions, but only one side of the molecule is visualized. Roadmaps are flat, two-dimensional projections that cannot be rotated in real time to reveal other views. Three-dimensionality is limited to an extension by Jean-Yves Sgro that maps the parallel projection of one view to a three-dimensional surface shell that can be rotated with interactive graphics and/or viewed with stereo imaging (Harber *et al.*, 1995; Sgro, 1996). However, the schematic nature of roadmaps leads to the ability to view all parts of the molecule simultaneously.

To view all parts of the molecule, cylindrical projections are used that are similar to those used in atlases. This is possible because the

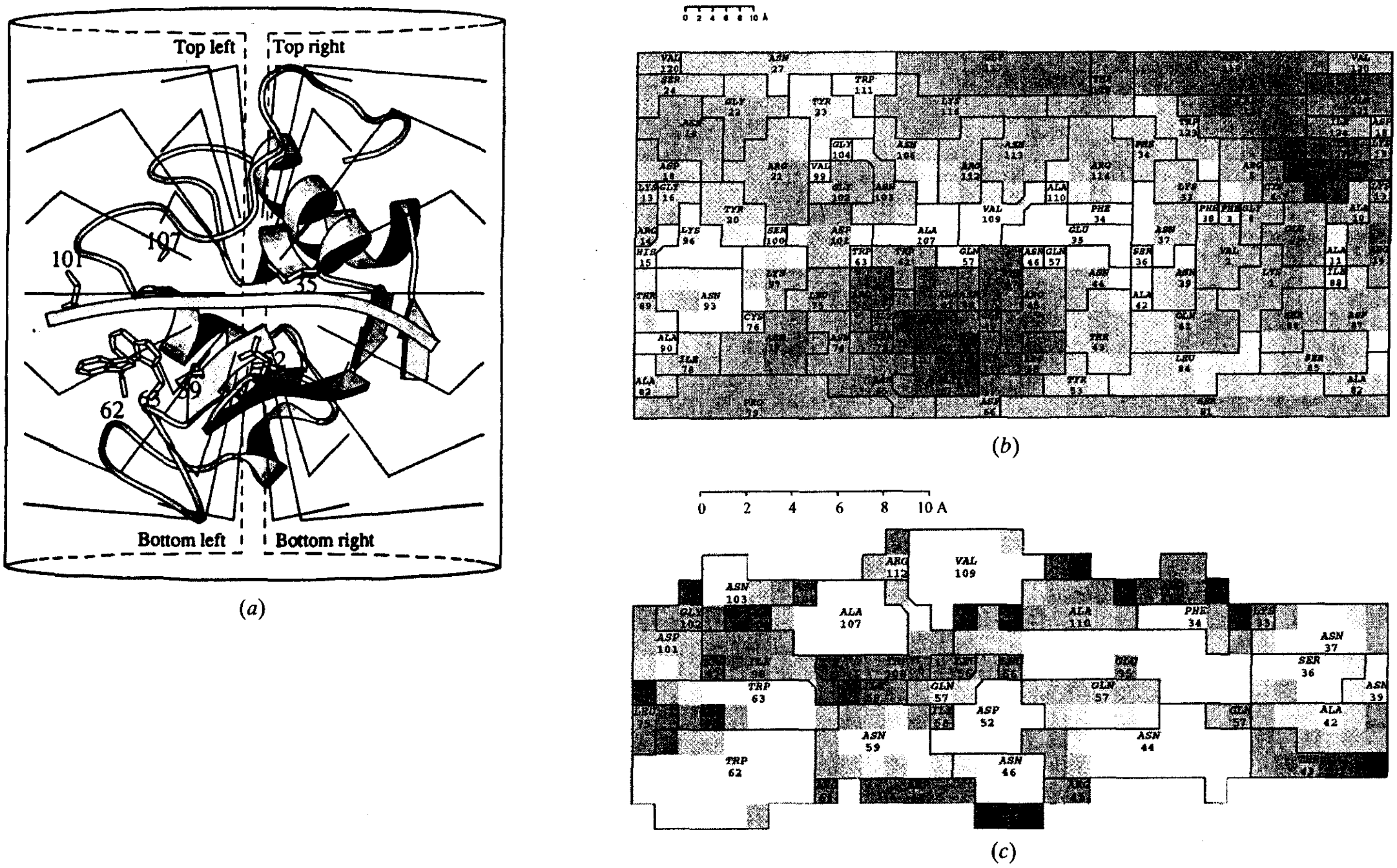

Fig. 22.1.2.4. Different projections illustrated with lysozyme. (*a*) Lysozyme (Blake *et al.*, 1965; Diamond, 1974) is sketched with *MOLSCRIPT* (Kraulis, 1991) and shown with a ribbon indicting the active-site cleft (Kelly *et al.*, 1979). A cylindrical surface shown unrolled in (*b*) is shown in (*a*) wrapped around lysozyme. Vectors orthogonal to the now cylindrical illustrative surface are extended inwards until they intersect with the sphere. Vectors then run towards the centre of the molecule, and their intersections with the solvent-accessible surface are projected back upon the cylinder [unrolled in (*b*)]. (*b*) The surface is shaded according to distance from the centre, revealing the substrate-binding cleft as lighter shading. Details of active-site residues are revealed in (*c*) with a different type of projection. A segmented bent cylinder was traced along the substrate-binding cleft. The surface shows the projection outwards from points on the cylindrical axis. This reveals the amino acids likely to be in most intimate contact with the substrate. Similar plots, coloured according to charge, atomic solvation parameters, or hydrophobicity, can reveal the nature of predominant chemical interactions. This figure is reproduced from Chapman (1993). Copyright (1993) The Protein Society. Reprinted with the permission of Cambridge University Press.

representation is schematic (not realistic), and longitudinal distortion, similar to that near the poles in world maps, is acceptable. The surface is projected outwards radially onto a cylinder that wraps around the macromolecule (Fig. 22.1.2.4). Active-site clefts, drug or inhibitor binding sites and pores can be similarly illustrated by projecting their surfaces outward (from the axis) onto a cylinder that encloses the pore, pocket, or cleft. Such clefts are rarely straight, but with some distortion a satisfactory representation is possible by segmenting the cylinder, so that its axis follows the (curved) centre of the binding site or pore (Fig. 22.1.2.4).

22.1.2.5. *Conclusion*

Both quantitative and qualitative analyses of the surfaces of biomolecules are among the most powerful methods of elucidating functional mechanism from three-dimensional structures. A wide array of methods have been developed to help understand binding interactions and macromolecular assembly and to visualize the shape and physiochemical surface properties of macromolecules. Visualization methods range from those that depict a realistic impression of the topology to those that are more schematic and facilitate collation of structural and genetic information.

Acknowledgements

The authors thank Genfa Zhou for providing Fig. 22.1.2.2. MSC gratefully acknowledges the support of the National Science Foundation (BIR 94-18741 and DBI 98-08098), the National Institutes of Health (GM 55837) and the Markey Charitable Trust. MG acknowledges support from the NSF Database Activities Program DBI-9723182.

22.2.1. Introduction

The hydrogen bond (Huggins, 1971) plays a critical role in the structure and function of biological macromolecules. This is because, uniquely among the non-covalent interactions that stabilize such structures, it combines a strong directional character with its energetic contributions. Thus, hydrogen-bonding patterns define the secondary structures that form the framework of proteins, are responsible for the specificity of base pairing in nucleic acids, shape the loops and irregular features that often determine molecular recognition, and provide for appropriately oriented functional groups in catalytic and/or binding sites.

Much of our present knowledge of hydrogen bonding in biological structures is foreshadowed in Linus Pauling's influential book (Pauling, 1960), and Jeffrey & Saenger (1991) have provided a comprehensive recent review. Other important reviews have covered hydrogen-bonding patterns in globular proteins (Baker & Hubbard, 1984; Stickle *et al.*, 1992), the satisfaction of hydrogen-bonding potential in proteins (McDonald & Thornton, 1994*a*), hydrogen-bonding patterns for side chains (Ippolito *et al.*, 1990) and side-chain hydrogen bonding in relation to secondary structures (Bordo & Argos, 1994).

22.2.2. Nature of the hydrogen bond

Hydrogen bonds are attractive electrostatic interactions of the type D—H$\cdots A$, where the H atom is formally attached to a donor atom, D (assumed to be more negative than H), and is directed towards an acceptor, A. The acceptor A is normally an electronegative atom, usually O or N, but occasionally S or Cl, with a full or partial negative charge and a lone pair of electrons directed towards the H atom. Although most of the hydrogen bonds in proteins and nucleic acids are N—H$\cdots$O or O—H$\cdots$O (less often, N—H$\cdots$N), it is important to be aware that other possibilities exist, including N—H$\cdots$S, O—H$\cdots$S and C—H$\cdots$O, and that these can be very important in specific cases (Adman *et al.*, 1975; Derewenda *et al.*, 1995). Likewise, the π-electron clouds of aromatic rings can also act as acceptors for appropriately oriented D—H groups (Legon & Millen, 1987; Mitchell *et al.*, 1994).

In an ideal hydrogen bond, the donor heavy atom, the H atom, the acceptor lone pair and the acceptor heavy atom should all lie in a straight line (Legon & Millen, 1987), as illustrated in Fig. 22.2.2.1(*a*). The strength of the interaction is also expected to depend on the electronegativities of the atoms involved. Hydrogen bonds are said to be bifurcated when a single D—H group interacts with two acceptors in a three-centred hydrogen bond (Fig. 22.2.2.1*b*); these hydrogen bonds are necessarily nonlinear and weaker. However, the term bifurcated is also sometimes applied to the quite different situation where a donor atom with two H atoms or an acceptor atom with two lone pairs makes two hydrogen bonds, as in Figs. 22.2.2.1(*c*) and (*d*). These interactions can be strong and linear. Some hydrogen-bonding arrangements are said to be cooperative; for example, hydrogen bonding by a peptide C=O group should enhance the polarity of the whole peptide unit and hence the acidity of the amide proton and the strength of its hydrogen bonding (Jeffrey & Saenger, 1991).

22.2.3. Hydrogen-bonding groups

22.2.3.1. *Proteins*

The hydrogen-bonding capacities of the various hydrogen-bonding groups in proteins are shown in Fig. 22.2.3.1. All, with the exception of the peptide NH and Trp side-chain NH groups, can participate in more than one hydrogen-bond interaction. Peptide and side-chain C=O groups, for example, can act as acceptors for two hydrogen bonds by using both lone pairs of electrons on the sp^2-hybridized oxygen. Likewise, the —OH groups of Ser or Thr can act as donors through their single H atom, and acceptors through their two lone pairs. In Tyr side chains, the C—O bond has some double-bond character, and the phenolic —OH is thus likely to prefer only two hydrogen bonds, both in the ring plane. The carboxylate groups of Asp and Glu are normally ionized above pH 4 and their C—O bonds also have partial double-bond character; each carboxylate oxygen should then be able to accept two hydrogen bonds, although the restriction to two may be less severe than for C=O.

Several uncertainties exist. Crystallographically, it is not usually possible to distinguish the amide oxygen and nitrogen atoms of Asn and Gln, and the decision as to which is which has to be made on environmental grounds by considering what hydrogen bonds would be made in each of the two possible arrangements. Likewise, two possibilities exist for His side chains by rotating 180° about C^β—C^γ. This problem has been analysed by McDonald & Thornton (1994*b*), and corrections can be made with *HBPLUS*.

For some side chains, the ionization state is uncertain. Arg and Lys are assumed to be fully protonated, as in Fig. 22.2.3.1, and Asp and Glu are assumed to be fully ionized. Nevertheless, a survey by Flocco & Mowbray (1995) has shown that a small but significant number of short O$\cdots$O distances between Asp and Glu side chains must represent O—H$\cdots$O hydrogen bonds, with one carboxyl group protonated. His side chains, in addition to the orientational uncertainty, have a pK_a (~6.5) that implies that they may be in either their neutral or their protonated form, depending on pH and environment. In the neutral form, only one N atom is protonated (more often $N^{\varepsilon2}$, but sometimes $N^{\delta1}$), but in the protonated form both N atoms carry protons; again, the actual state has to be deduced from their environment.

22.2.3.2. *Nucleic acids*

The three components of nucleic acids, *i.e.* phosphate groups, sugars and bases, all participate in hydrogen bonding to greater or lesser extent. The phosphate oxygen atoms can potentially act as acceptors of two or more hydrogen bonds and are frequently the

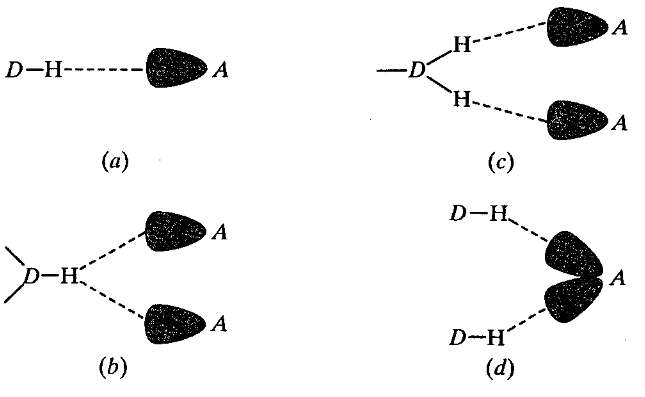

Fig. 22.2.2.1. Hydrogen-bonding configurations. (*a*) The standard two-centre hydrogen bond in which an H atom attached to a donor atom, D, is directed towards a lone pair of an acceptor, A. (*b*) A classic three-centre, or bifurcated, hydrogen bond, with a single H atom shared between the lone pairs of two acceptors. The situations shown in (*c*) and (*d*) are not true three-centre hydrogen bonds since they are essentially equivalent to that in (*a*).

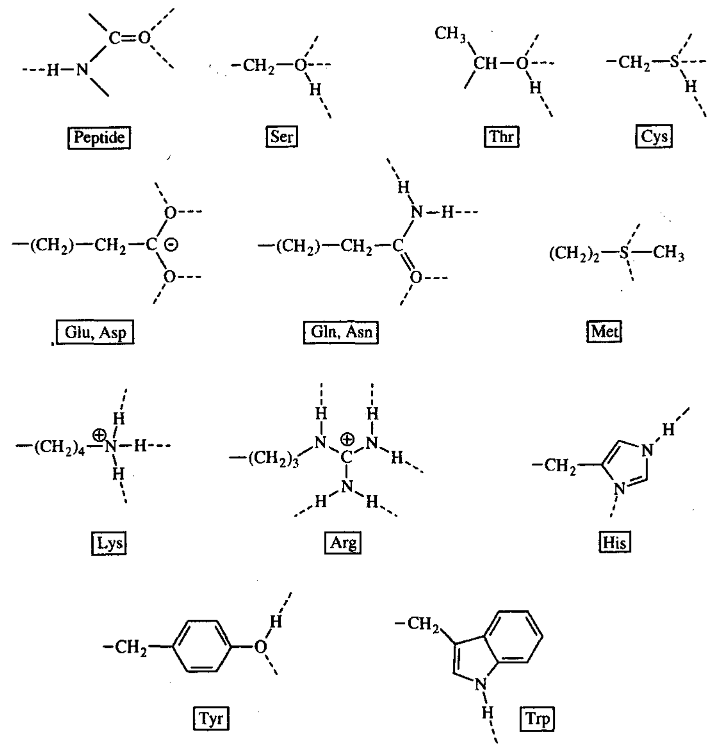

Fig. 22.2.3.1. Hydrogen-bonding potential of protein functional groups. Potential hydrogen bonds are shown with broken lines. Arg, Lys, Asp and Glu side chains are shown in their ionized forms.

recipients of hydrogen bonds from protein side chains in protein–DNA complexes. The sugar residues of RNA have a 2′-OH which can act as both hydrogen-bond donor and acceptor, and the 4′-O of both ribose and deoxyribose can potentially accept two hydrogen bonds.

It is the bases of DNA and RNA that have the greatest hydrogen-bonding potential, however, with a variety of hydrogen-bond donor or acceptor sites. Although each of the bases could theoretically occur in several tautomeric forms, only the canonical forms shown in Fig. 22.2.3.2 are actually observed in nucleic acids. This leads to clearly defined hydrogen-bonding patterns which are critical to both base pairing and protein–nucleic acid recognition. The —NH_2 and >NH groups act only as hydrogen-bond donors, and C=O only as acceptors, whereas the >N— centres are normally acceptors but at low pH can be protonated and act as hydrogen-bond donors.

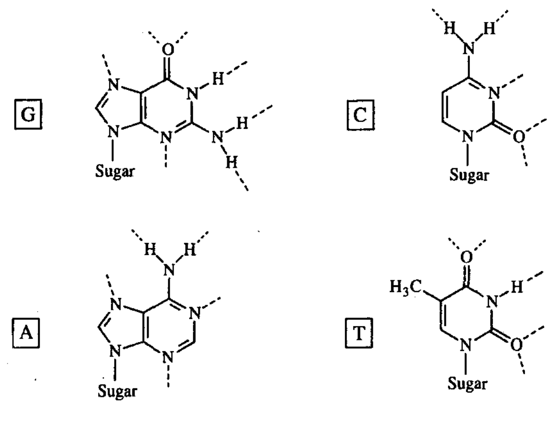

Fig. 22.2.3.2. Hydrogen-bonding potential of nucleic acid bases guanine (G), adenine (A), cytosine (C) and thymine (T) in their normal canonical forms.

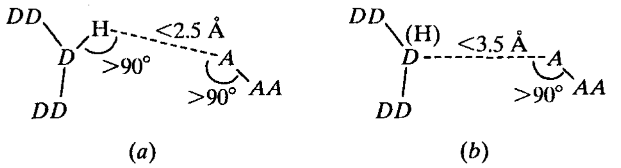

Fig. 22.2.4.1. Suggested criteria for identifying likely hydrogen bonds. *DD* and *AA* represent atoms covalently bonded to the donor atom, *D*, and acceptor atom, *A*, respectively. Here, (*a*) represents the criteria when the donor H atom can be placed, and (*b*) when it cannot be placed. Additional criteria based on the angle *DD—D*···*A* could be incorporated with (*b*). Adapted from Baker & Hubbard (1984) and McDonald & Thornton (1994*a*).

22.2.4. Identification of hydrogen bonds: geometrical considerations

Because hydrogen bonds are electrostatic interactions for which the attractive energy falls off rather slowly (Hagler *et al.*, 1974), it is not possible to choose an exact cutoff for hydrogen-bonding distances. Rather, both distances and angles must be considered together; the latter are particularly important because of the directionality of hydrogen bonding. Inferences drawn from distances alone can be highly misleading. An approach with an N—H···O angle of 90° and an H···O distance of 2.5 Å would be very unfavourable for hydrogen bonding, yet it translates to a N···O distance of 2.7 Å. This could (wrongly) be taken as evidence of a strong hydrogen bond.

For macromolecular structures determined by X-ray crystallography, problems also arise from the imprecision of atomic positions and the fact that H atoms cannot usually be seen. Thus, the geometric criteria must be relatively liberal. H atoms should also be added in calculated positions where this is possible; this can be done reliably for most NH groups (peptide NH, side chains of Trp, Asn, Gln, Arg, His, and all >NH and NH_2 groups in nucleic acid bases).

The hydrogen-bond criteria used by Baker & Hubbard (1984) are shown in Fig. 22.2.4.1. Very similar criteria are used in the program *HBPLUS* (McDonald & Thornton, 1994*a*), which also adds H atoms in their calculated positions if they are not already present in the coordinate file. In general, hydrogen bonds may be inferred if an interatomic contact obeys *all* of the following criteria:

(1) The distance H···*A* is less than 2.5 Å (or *D*···*A* less than 3.5 Å if the donor is an —OH or —NH_3^+ group or a water molecule).

(2) The angle at the H atom, *D*—H···*A*, is greater than 90°.

(3) The angle at the acceptor, *AA*—*A*···H (or *AA*—*A*···*D* if the H-atom position is unreliable), is greater than 90°.

Other criteria can be applied, for example taking into account the hybridization state of the atoms involved and the degree to which any approach lies in the plane of the lone pair(s). In all analyses of hydrogen bonding, however, it is clear that a combination of distance and angle criteria is effective in excluding unlikely hydrogen bonds.

22.2.5. Hydrogen bonding in proteins

22.2.5.1. *Contribution to protein folding and stability*

The net contribution of hydrogen bonding to protein folding and stability has been the subject of much debate over the years. The current view is that although the hydrophobic effect provides the driving force for protein folding (Kauzmann, 1959), many polar groups, notably peptide NH and C=O groups, inevitably become buried during this process, and failure of these groups to find hydrogen-bonding partners in the folded protein would be strongly destabilizing. This, therefore, favours the formation of secondary structures and other structures that permit effective hydrogen bonding in the folded molecule. Not surprisingly, the contribution of specific hydrogen bonds to stability depends on their location in the structure (Fersht & Serrano, 1993). Mutagenesis studies have shown that even the loss of a single hydrogen bond can be significantly destabilizing (Alber *et al.*, 1987) and that the energetic contribution can vary depending on whether or not the groups involved are charged (Fersht *et al.*, 1985).

22.2.5.2. *Saturation of hydrogen-bond potential*

A consistent conclusion from analyses of protein structures is that virtually all polar atoms either form explicit hydrogen bonds or are at least in contact with external water. The extent to which their *full* hydrogen-bond potential is fulfilled in a folded protein (for example, the potential of an Arg side chain to make five hydrogen bonds) has been examined in several studies. Baker & Hubbard (1984) considered the explicit hydrogen bonds made by main-chain and side-chain atoms in a number of refined protein structures and established general patterns for both, but did not differentiate buried and solvent-exposed atoms or allow for unmodelled solvent. Savage *et al.* (1993) used the solvent accessibilities of polar groups to estimate their assumed numbers of hydrogen bonds to external water. This supplemented the explicit hydrogen bonds that could be derived from the atomic coordinates and allowed an estimate of the extent to which potential hydrogen bonds are lost during protein folding. McDonald & Thornton (1994*a*) focused specifically on buried hydrogen-bond donors and acceptors in order to determine the extent to which the hydrogen-bond potential of these is utilized.

The results of these analyses can be summarized as follows. Almost all polar groups do in fact make at least one hydrogen bond. Hydrogen-bond donors are almost always hydrogen bonded; only 4% of NH groups 'lose' hydrogen bonds as a result of protein folding (Savage *et al.*, 1993). On the other hand, hydrogen-bond acceptors often do not exert their full hydrogen-bonding potential. For example, for main-chain C=O groups, which are expected to accept two hydrogen bonds, 24% of possible hydrogen bonds are estimated to be lost during folding (Savage *et al.*, 1993). Among buried C=O groups, although very few make no hydrogen bonds (as little as 2% if hydrogen-bonding criteria are relaxed), the majority fail to form a second hydrogen bond (McDonald & Thornton, 1994*a*). Steric factors, particularly in β-sheets or where Pro residues are adjacent, restrict hydrogen-bonding possibilities, although some of the 'lost' interactions may be recovered through C—H$\cdots$O interactions (see Section 22.2.7.1). McDonald & Thornton also point out that failure to form a second hydrogen bond is less energetically expensive than failure to form the first. Among polar side chains, the ionizable side chains (Asp, Glu, Arg, Lys, His) show a very strong tendency to be fully hydrogen bonded or solvent exposed. Buried Arg side chains, for example, frequently form all five possible hydrogen bonds. The side chains that most often fail to fulfil their full hydrogen-bond potential are Ser, Thr and Tyr; these almost always donate one hydrogen bond but frequently fail to accept one.

22.2.5.3. *Secondary structures*

Secondary structures provide the means whereby the polar C=O and NH groups of the polypeptide chain can remain effectively hydrogen bonded when they are buried within a folded globular protein. In doing so, they provide the framework of folding patterns and account for the majority of hydrogen bonds within protein structures. The three secondary-structure classes (helices, β-sheets and turns) are each characterized by specific hydrogen-bonding patterns, which can be used for objective identification of these structures (Stickle *et al.*, 1992).

22.2.5.3.1. *Helices*

Helices have traditionally been defined in terms of their N—H$\cdots$O=C hydrogen-bonding patterns as α-helices ($i \rightarrow i-4$), 3_{10}-helices ($i \rightarrow i-3$), or π-helices ($i \rightarrow i-5$); in an α-helix, for example, the peptide NH of residue 5 hydrogen bonds to the C=O of residue 1. In fact, the vast majority of helices in proteins are α-helices; 3_{10}-helices are rarely more than two turns (six residues) in length, and discrete π-helices have not been seen so far.

The residues within helices have characteristic main-chain torsion angles, (φ, ψ), of around (−63°, −40°) that cause the C=O groups to tilt outwards by about 14° from the helix axis (Baker & Hubbard, 1984). This results in somewhat less linear hydrogen bonding than in the original Pauling model (Pauling *et al.*, 1951), with a degree of distortion towards 3_{10}-helix geometry. Thus, weak $i \rightarrow i-3$ interactions are often made in addition to the more favourable $i \rightarrow i-4$ hydrogen bonds, giving hydrogen-bond networks that may enhance helix elasticity (Stickle *et al.*, 1992). Tilting outwards also makes the C=O groups more accessible for additional hydrogen bonds from side chains or water molecules. For the α-type, $i \rightarrow i-4$ interactions, the hydrogen-bond angles at both donor and acceptor atoms are quite tightly clustered (N—H$\cdots$O ~157° and C=O$\cdots$H ~ 147°). The hydrogen-bond lengths in helices average 2.06 (16) Å (O$\cdots$H) or 2.99 (14) Å (O$\cdots$N) (Baker & Hubbard, 1984).

Few helices are regular throughout their length. Many are curved or kinked such that one side (often the outer, solvent-exposed side) of the helix is opened up a bit and has longer hydrogen bonds (Blundell *et al.*, 1983; Baker & Hubbard, 1984). The bends are often associated with additional hydrogen bonds from water molecules or side chains to C=O groups that are tilted out more than usual. Curved helices are normal in coiled-coil structures and can enable long helices to pack more effectively in globular structures. Sometimes a kink can be functionally important, as in manganese superoxide dismutase, where a kink in a long helix, incorporating a π-type ($i \rightarrow i-5$) hydrogen bond, enables the optimal positioning of active-site residues (Edwards *et al.*, 1998).

The beginnings and ends of helices are sites of hydrogen-bonding variations which can be seen as characteristic 'termination motifs'. At helix N-termini, 3_{10}-type $i \rightarrow i-3$ (or bifurcated $i \rightarrow i-3$ and $i \rightarrow i-4$) hydrogen bonds are often found. At C-termini, two common patterns occur. In one, labelled $\alpha_{\rm C1}$ by Baker & Hubbard (1984), there is a transition from α-type, $i \rightarrow i-4$ to 3_{10}-type, $i \rightarrow i-3$ hydrogen bonding, often with genuine bifurcated hydrogen bonds, as in Fig. 22.2.2.1(*b*), at the transition point. The other, labelled $\alpha_{\rm C2}$ (Baker & Hubbard, 1984) or referred to as the 'Schellman motif' (Schellman, 1980), has a π-type, $i \rightarrow i-5$ hydrogen bond coupled with a 3_{10}-type, $i-1 \rightarrow i-4$ hydrogen bond; residue $i-1$ has a left-handed α configuration and is often Gly. The beginnings and ends of helices are also the sites of specific side-chain hydrogen-bonding patterns, referred to as N-caps and C-caps (Presta & Rose, 1988; Richardson & Richardson, 1988); these are described below.

22.2.5.3.2. *β-sheets*

β-sheets consist of short strands of polypeptide (typically 5–7 residues) running parallel or antiparallel and cross-linked by N—H$\cdots$O=C hydrogen bonds. Although the (φ, ψ) angles of residues within β-sheets can be quite variable, the hydrogen-bonding patterns within these segments tend to be quite regular, as in the original Pauling models (Pauling & Corey, 1951). Occasional β-bulges in the middle of β-strands can interrupt the hydrogen-bonding pattern (Richardson *et al.*, 1978), but otherwise disruptions occur only at the ends of strands. The hydrogen bonds in β-sheets appear to be slightly shorter than those in helices, by ~0.1 Å, and also more linear (N—H···O ~ 160°, compared with ~157° in helices) (Baker & Hubbard, 1984). There also appears to be no difference between parallel and antiparallel β-sheets in the hydrogen-bond lengths and angles.

22.2.5.3.3. *Turns*

By far the most common type of turn is the β-turn, a sequence of four residues that brings about a reversal in the polypeptide chain direction. Hydrogen bonding does not seem to be essential for turn formation, but a common feature is a hydrogen bond between the C=O group of residue 1 and the NH group of residue 4, a 3_{10}-type, $i \rightarrow i-3$ interaction. Turns are also often associated with characteristic side-chain–main-chain hydrogen-bond configurations (see below). The hydrogen bonds in turns tend to be longer and less linear than those in helices and β-sheets; in particular, the angle at the acceptor oxygen atom C—O···H is around 120° (Baker & Hubbard, 1984).

In addition to β-turns, a small but significant number of γ-turns are found. In these three-residue turns, a hydrogen bond is formed between the C=O of residue 1 and the NH of residue 3, an $i \rightarrow i-2$ interaction. Although the approach to the acceptor oxygen atom is highly nonlinear (C—O···H ~ 100°), the nonlinearity at the H atom is less pronounced (N—H···O ~ 130–150°) (Baker & Hubbard, 1984). γ-turns are again of several types, depending on the configuration of the central residue. The classic γ-turn, first recognised by Matthews (1972) and Nemethy & Printz (1972), has a central residue with (φ, ψ) angles around (70°, −60°), which puts it in the normally disallowed region of the Ramachandran plot. More common, however, are structures in which an $i \rightarrow i-2$ hydrogen bond is associated with a central residue with a configuration around (90°, −70°) (Baker & Hubbard, 1984); these structures are not necessarily true turns in the sense of bringing about a sharp chain reversal, however.

22.2.5.3.4. *Aspects of in-plane geometry*

For hydrogen bonds involving sp^2 donors and/or acceptors, optimal interaction is expected to occur when the donor D—H group and the acceptor lone-pair orbital are coplanar (Taylor *et al.*, 1983). Analysis of 'in-plane' and 'out-of-plane' components of N—H···O hydrogen bonds in proteins shows that these have characteristic values for different secondary structures (Artymiuk & Blake, 1981; Baker & Hubbard, 1984). The out-of-plane component is tightly clustered at ~25° for helices and ~60° for the most common β-turns (type I and type III), but is widely scattered around a mean of 0° for β-sheets. The latter reflects different twists or curvature of β-sheets. The large out-of-plane component for turns is consistent with a relatively weak interaction.

22.2.5.4. *Side-chain hydrogen bonding*

An important concept in understanding the patterns of side-chain hydrogen bonding in proteins is that of local *versus* non-local interactions; local means that a side chain hydrogen bonds to another residue that is relatively close to it in the linear amino-acid sequence. Baker & Hubbard (1984) were first to introduce this distinction, with local defined as ±4 residues. Bordo & Argos (1994) define local as ±6 residues and Stickle *et al.* (1992) as ±10 residues. The distinction is not important, but the distributions in all three analyses show that ±5 would encompass all the significant populations of local hydrogen bonds. Local hydrogen bonds, in which side chains interact with nearby main-chain atoms or other side chains, are evidently critical for protein folding. Non-local hydrogen bonds, although fewer in number (see below), in turn can be very important for stabilization of the folded protein.

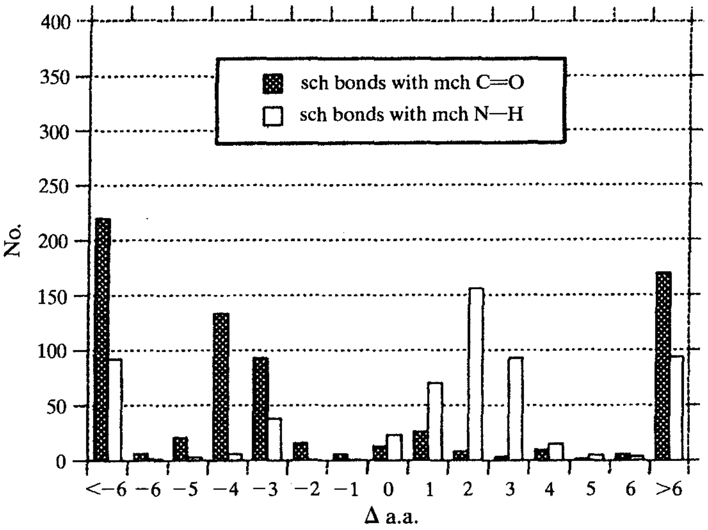

Fig. 22.2.5.1. Distribution of side-chain–main-chain hydrogen bonds as a function of the separation (Δ a.a.) along the polypeptide between the side-chain (sch) and main-chain (mch) groups involved, *i.e.* Δ a.a. = −*n* means that a side chain interacts with a main-chain group *n* residues earlier in the polypeptide (towards the N-terminus). Reproduced with permission from Bordo & Argos (1994). Copyright (1994) Academic Press.

If hydrogen bonds with water are excluded, a rule of thirds applies. Approximately one-third of the hydrogen bonds made by side chains (sch's) are with main-chain (mch) C=O groups, one-third are with main-chain NH groups, and one-third with other side chains. Within these populations, however, there are significant differences. For sch–mch(C=O) hydrogen bonds, approximately 45% are local; for sch–mch(NH) hydrogen bonds, a much higher proportion is local (69%), and for sch–sch hydrogen bonds, the proportion is much less (35%) (Bordo & Argos, 1994).

The distribution of local sch–mch(NH) hydrogen bonds shows a marked positional preference (Fig. 22.2.5.1) that highlights consistent hydrogen-bonding motifs found in all proteins (Fig. 22.2.5.2). The major peak involves side chains that interact with an NH group two residues further on in the polypeptide, an n-NH($n+2$) hydrogen bond. This motif primarily involves Asp, Asn, Ser and Thr side chains and is most often found (i) in turns, where a side chain from position 1 hydrogen bonds to the NH of residue 3, (ii) in loop regions where it stabilizes the local structure

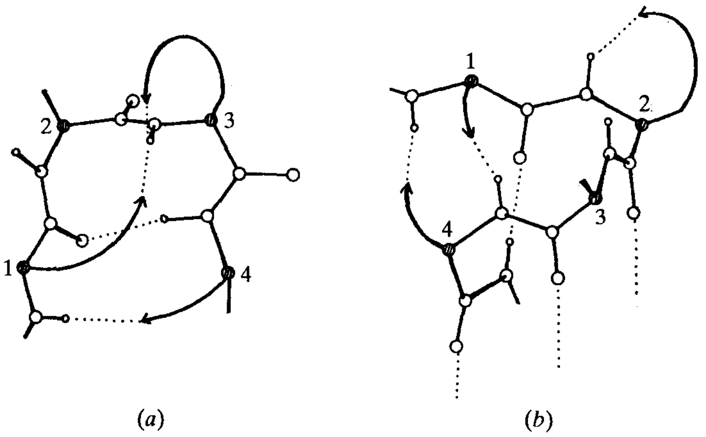

Fig. 22.2.5.2. Schematic representations of common classes of side-chain–main-chain hydrogen bonds (*a*) in turns and (*b*) at helix N-termini. Arrows represent side chains that hydrogen bond to main-chain CO or NH groups (NH identified by the small circle for H).

but is not necessarily associated with chain reversal, and (iii) at helix N-termini.

Helix N-termini are also the site of other characteristic local side-chain–NH hydrogen-bonding motifs (Baker & Hubbard, 1984; Presta & Rose, 1988; Richardson & Richardson, 1988; Harper & Rose, 1993; Bordo & Argos, 1994). Prominent among these are sch–NH($n+3$) hydrogen bonds involving Ser, Thr, Asp and Asn side chains, but sch–NH($n-3$) interactions, in which Glu or Gln side chains hydrogen bond back to a main-chain NH, form an important lesser category. Other motifs, such as that in which a Glu or Gln side chain bends round to hydrogen bond to its own NH group, are also found. Collectively, these contribute to helix capping motifs (Fig. 22.2.5.2*b*) that help satisfy the hydrogen bonding of the 'free' NH groups of the helix N-terminus and in effect extend the helix; the sch–mch(NH) hydrogen bond mimics the mch–mch $C{=}O\cdots HN$ hydrogen bonds of the helix. Helix N-capping by side chains is probably a very important influence in protein folding, acting as a stereochemical code for helix initiation (Presta & Rose, 1988; Harper & Rose, 1993).

The distribution of sch–mch(CO) hydrogen bonds also shows a striking preference, this time for positions -3 and -4. These sch–CO($n-3$) or sch–CO($n-4$) hydrogen bonds account for the vast majority of local hydrogen bonds between side chains and main-chain C=O groups. Almost all (~85%) are in helices, with most of the remainder in turns. They involve predominantly (~80%) Ser and Thr side chains but other side chains (Asn, His, Arg) can also participate. These local hydrogen bonds can occur at any point along a helix, where they are often associated with helix bending or kinking (Baker & Hubbard, 1984). However, they are most frequently found at helix C-termini (Bordo & Argos, 1994) and may constitute a termination motif.

Local side-chain–side-chain hydrogen bonds, although common, do not seem to fit into any obvious patterns; the only recurring interaction identified so far is between side chains on succeeding turns of helices, *i.e.* separated by approximately four residues. These frequently involve charged side chains, which can form hydrogen-bonded ion pairs. In sections of extended chain, side chains that are two residues apart may similarly interact.

Non-local hydrogen bonding by side chains is less easy to categorize but is no less significant; more than 50% of side-chain–main-chain(C=O) hydrogen bonds are non-local, as are ~65% of side-chain–side-chain hydrogen bonds. In most proteins, a small number of polar side chains with multiple hydrogen-bonding capability act as the centre for networks of hydrogen bonds; these appear to be particularly important for stabilizing non-repetitive polypeptide chain structures (coil, loops). Examples are given in Baker & Hubbard (1984). Most often these involve larger side chains with more than one hydrogen-bonding centre (Asn, Asp, Gln, Glu, Arg, His) which cross-link different sections of the polypeptide. Arg side chains interacting with main-chain C=O groups seem to be particularly effective; Ser and Thr, on the other hand, are seldom used, even though both have the potential to form three hydrogen bonds.

The geometry of side-chain hydrogen bonding has been analysed by Baker & Hubbard (1984) and, more extensively, by Ippolito *et al.* (1990). The former concentrate on hydrogen-bond lengths and angles and show that the preferred angles fit well with stereochemical expectations. Ippolito *et al.* examine the preferences for the various hydrogen-bonding sites around each side-chain type by means of scatter plots (Fig. 22.2.5.3) from which probability densities are computed. These show that well defined preferences exist, determined by both steric and electronic effects.

22.2.5.5. *Hydrogen bonds with water molecules*

Water molecules, with their small size and double-donor, double-acceptor hydrogen-bonding capability, are ideal for completing intramolecular hydrogen-bonding networks, *e.g.* by linking two proton acceptor atoms, or two protein donor atoms, that cannot otherwise interact. Thus, buried water molecules, making multiple hydrogen bonds, help satisfy the hydrogen-bond potential of internal polar atoms and contribute to protein stability; internal waters average about three hydrogen bonds each (Baker & Hubbard, 1984; Williams *et al.*, 1994). From the survey of Williams *et al.* (1994), most (58%) occupy discrete cavities, while 22% are in clusters housing two waters and 20% are in larger clusters; some examples of larger clusters are given in Baker & Hubbard (1984). Buried waters are often conserved between homologous proteins (Baker, 1995), and each buried water–protein hydrogen bond is estimated to stabilize a folded protein by, on average, 0.6 kcal mol^{-1} (1 kcal mol^{-1} = 4.184 kJ mol^{-1}) (Williams *et al.*, 1994). More loosely bound external waters exchange much more rapidly and presumably contribute less energetically.

Several patterns of hydrogen bonding are consistently observed. Water molecules are most often seen interacting with oxygen atoms rather than nitrogen atoms and acting as hydrogen-bond donors rather than acceptors. Possible reasons include the greater number of acceptor sites in proteins and the fewer geometrical restrictions imposed by acceptors (Baker & Hubbard, 1984; Baker, 1995). There is also a predominance of interactions with main-chain atoms rather than side-chain atoms: on average ~40% with main-chain C=O groups, 15% with main-chain NH and 45% with side-chain groups (Baker & Hubbard, 1984; Thanki *et al.*, 1988). Favoured main-chain binding sites include the N- and C-termini of helices, C=O groups on the solvent-exposed sides of helices, the edge strands of β-sheets, and the ends of strands where they add extra inter-strand hydrogen bonds at the position where the strands diverge (Thanki *et al.*,

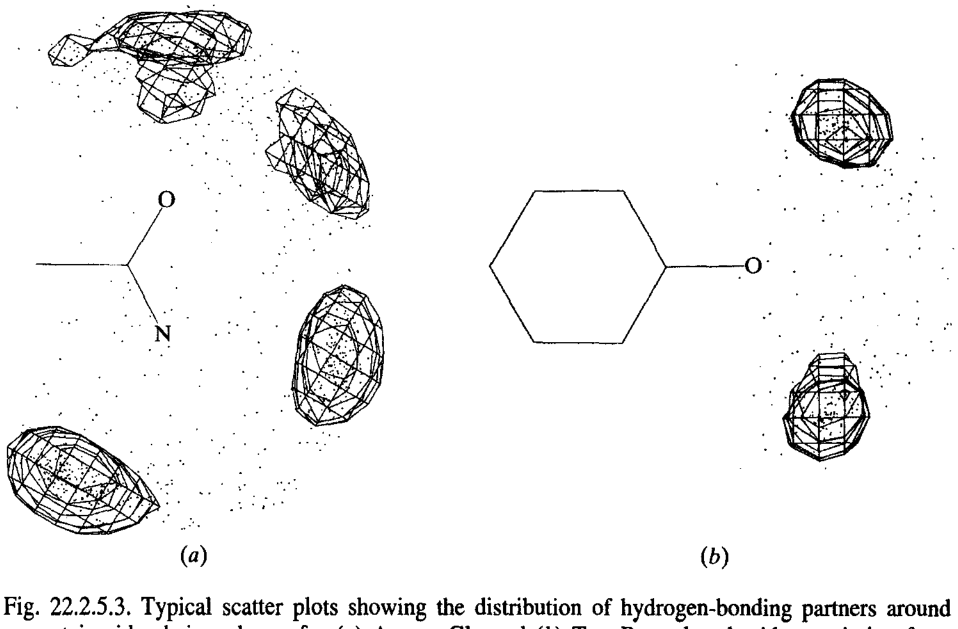

Fig. 22.2.5.3. Typical scatter plots showing the distribution of hydrogen-bonding partners around protein side chains, shown for (*a*) Asn or Gln and (*b*) Tyr. Reproduced with permission from Ippolito *et al.* (1990). Copyright (1990) Academic Press.

1991). Among side chains, the most highly hydrated appear to be Asp and Glu, whose COO^- groups bind, on average, two water molecules each (Baker & Hubbard, 1984; Thanki *et al.*, 1988). On the other hand, the best-ordered water sites are created by residues whose side chains simultaneously make hydrogen bonds to other protein atoms (His, Asp, Asn, Arg) or may be sterically restricted (Tyr, Trp).

The distributions of water molecules around protein groups follow the geometrical patterns expected from simple bonding ideas (Baker & Hubbard, 1984; Thanki *et al.*, 1988). Interactions with NH groups are linear, and those with C=O groups show a preferred angle of ~130° at the oxygen-atom acceptor, consistent with interaction with an oxygen-atom lone pair; restriction to the peptide plane is not very strong, however. Although the distributions around polar side chains generally follow the expected patterns (Thanki *et al.*, 1988), there is little evidence of ordered water clusters around non-polar groups. This may be because water clusters need to be 'anchored' by hydrogen bonding to polar groups to be seen crystallographically.

22.2.6. Hydrogen bonding in nucleic acids

Hydrogen bonding by purine and pyrimidine bases is, together with base stacking, a major determinant of nucleic acid structure. With so many hydrogen-bonding groups, there are many potential modes of interaction between bases (Jeffrey & Saenger, 1991). Those that are actually found in DNA and RNA structures are, however, much more restricted in number, at least based on presently available experimental data.

22.2.6.1. *DNA*

DNA structure is dominated by the prevalence of duplex structures and hence by the classic Watson–Crick hydrogen-bonding pattern of A–T and G–C base pairs. This hydrogen-bonding pattern is not affected by whether the double helix has A-form, B-form, or Z-form geometry. Other hydrogen-bonding modes in DNA are probably very rare, arising only as a result of mutations (which produce mismatches), chemical modifications, such as methylation, or other disturbances, such as the binding of drugs or proteins so as to alter DNA conformation. Mismatches can give stable hydrogen bonding but at the expense of local perturbations of the DNA structure.

22.2.6.2. *RNA*

In contrast to DNA, RNA molecules generally form single-stranded structures, which are correspondingly much more complex and less regular. This means that catalytic and other activities can be generated in addition to their information-carrying roles. Current knowledge of detailed RNA three-dimensional structure is limited to transfer RNAs and several ribozymes, including a large ribosomal RNA domain (Cate *et al.*, 1996). Even from this small sample, however, it is clear that a great diversity of hydrogen-bonding interactions exists; RNA molecules contain regions of double-helical structure, often with classical Watson–Crick A–U and G–C base pairing, but these regions are interspersed with loops and bulges and tertiary interactions between the various secondary-structural (double-helical) elements. These interactions include many unconventional base pairings (*e.g.* see Fig. 22.2.6.1).

Some RNA structural motifs may prove to be of widespread general importance in RNA molecules. One example is a sharp turn with sequence CUGA in the hammerhead ribozyme that exactly matches turns in tRNAs (Pley *et al.*, 1994). Another is the GNRA tetraloop structure (N = any base, R = purine). This loop has a well defined structure, stabilized by hydrogen bonding and stacking involving its own bases, and it also presents further hydrogen-bonding groups that can dock into 'receptor' structures in other parts of the RNA molecule. This results in triple or quadruple base interactions (Fig. 22.2.6.1) that tie different parts of the RNA structure together; the parallel with hydrogen-bonding side chains in proteins is very strong. The 2′-hydroxyls of ribose groups are also used in some of these interactions (Fig. 22.2.6.1). Further ribose interactions involve interdigitated ribose groups that line the interfaces between adjacent helices such that pairs of riboses interact by hydrogen bonding through their 2′-hydroxyl groups, forming 'ribose zippers' As many more RNA structures are determined experimentally, it is likely that more hydrogen-bonding motifs will be recognized, and their full role in RNA structure can be better assessed than at our present, imperfect state of knowledge.

22.2.7. Non-conventional hydrogen bonds

The vast majority of hydrogen bonds in biological macromolecules involve nitrogen and oxygen donors exclusively. Nevertheless, several other interactions have all the characteristics of hydrogen bonds and clearly contribute to structure and stability where they occur.

22.2.7.1. *C—H···O hydrogen bonds*

Sutor (1962) first summarized evidence for C—H···O hydrogen bonds following earlier suggestions by Pauling (1960), and current evidence has been nicely summarized in several recent articles (Derewenda *et al.*, 1995; Wahl & Sundaralingam, 1997). The energy of C—H···O hydrogen bonds has been generally estimated as ~0.5 kcal mol^{-1} (about 10% of an N—H···O interaction) but may be higher, especially in hydrophobic environments. It also depends on the acidity of the C—H proton, with methylene (CH_2) and methyne (CH) groups being most favourable.

A number of examples of C—H···O hydrogen bonds can be found in nucleic acid structures (Wahl & Sundaralingam, 1997). The best known is that between the backbone O5′ oxygen and a purine C(8)—H or pyrimidine C(6)—H, when the bases are in the *anti* conformation. Another example is given by a U–U base pair, in which the two bases form a conventional N(3)—H···O(4) hydrogen bond and a C(5)—H···O hydrogen bond.

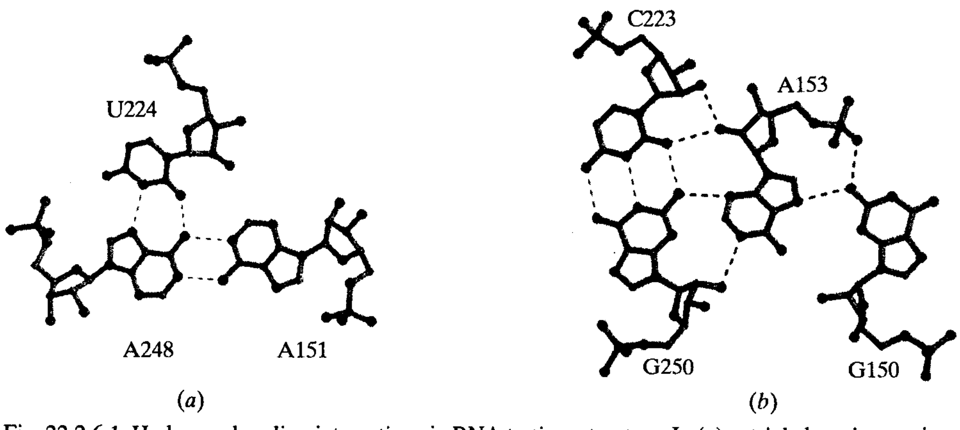

Fig. 22.2.6.1. Hydrogen-bonding interactions in RNA tertiary structure. In (*a*), a triple base interaction is shown. In (*b*), G150 and A153 of a GAAA tetraloop participate in multiple hydrogen-bond interactions involving bases, riboses and phosphate. Reprinted with permission from Cate *et al.* (1996). Copyright (1996) American Association for the Advancement of Science.

In proteins, two groups are regarded as being particularly significant (Derewenda *et al.*, 1995). These are the $C^{\epsilon}H$ of His side chains and the methylene H atoms of the main-chain α-carbon atoms. C—H···O hydrogen bonds involving His side chains have been found for the active-site His residues of proteins of the lipase/esterase family and in other proteins (Derewenda *et al.*, 1994). The $C^{\alpha}H$ atoms appear to provide much more widespread C—H···O hydrogen bonding, however, especially in β-sheets, where they are directed towards the 'free' lone pairs of the main-chain C=O groups. C—H···O hydrogen bonds may thus play a previously unrecognised role in satisfying the hydrogen-bond potential of C=O groups. In general, Derewenda *et al.* (1995) find a significant number of C···O contacts that meet the criteria for C—H···O hydrogen bonds; the H···O distance peaks at 2.45 Å (C···O 3.5 Å), which is less than the van der Waals distance of 2.7 Å, and the angles indicate that the H atoms are directed at the acceptor lone-pair orbitals.

22.2.7.2. *Hydrogen bonds involving sulfur atoms*

Sulfur atoms are larger and have a more diffuse electron cloud than oxygen or nitrogen, but are nevertheless capable of participating in hydrogen bonds. Given that the radius of sulfur is ~0.4 Å greater than that of oxygen, hydrogen bonds can be assumed if the distance H···S is less than ~2.9 Å, or S···O(N) is less than ~3.9 Å, providing the angular geometry is right. In proteins, the SH group of cysteine can be a hydrogen-bond acceptor or donor, whereas the sulfur atoms in disulfide bonds and in Met side chains can act only as acceptors.

The clearest example of hydrogen bonding involving Cys residues is given by the NH···S hydrogen bonds in Fe-S proteins (Adman *et al.*, 1975); here, peptide NH groups are oriented to point directly at the S atoms of metal-bound Cys residues, with H···S distances of 2.4–2.9 Å. Similar NH···S hydrogen bonds are found in blue copper proteins, involving the Cys ligands. In these cases, the cysteine sulfur is deprotonated and therefore more negative, making it a stronger hydrogen-bond acceptor, and it is likely that hydrogen bonding to cysteine S^- atoms is common. A large survey of Cys and Met side chains in proteins has given evidence of both N—H···S and S—H···O hydrogen bonds involving the SH groups of Cys side chains (Gregoret *et al.*, 1991). In particular, Cys residues in helices frequently hydrogen bond to the main-chain C=O group four residues back in the helix in SH(n)···O($n-4$) interactions analogous to those seen for Ser and Thr residues in helices. On the other hand, O—H···S or N—H···S hydrogen bonds to the S atoms of Met or half-cystine side chains, although they do exist, are rare (Gregoret *et al.*, 1991; Ippolito *et al.*, 1990).

22.2.7.3. *Amino-aromatic hydrogen bonding*

Surveys of protein structures have shown that aromatic rings (of Trp, Tyr, or Phe) are frequently in close association with side-chain NH groups of Lys, Arg, Asn, Gln, or His (Burley & Petsko, 1986). Energy calculations further suggest that where an N—H group, as donor, is directed towards the centre of an aromatic ring, as acceptor, a hydrogen-bonded interaction with an energy of ~3 kcal mol^{-1} (about half that of a normal N—H···O or O—H···O hydrogen bond) can result (Levitt & Perutz, 1988). Whether the close associations observed by Burley & Petsko can truly be regarded as hydrogen bonds has been controversial, however. Mitchell *et al.* (1994) have analysed amino–aromatic interactions and shown that by far the most common form of association between sp^2 nitrogen atoms and aromatic rings involves approximately plane-to-plane stacking, which cannot represent hydrogen bonding. There is still, however, a significant number of cases where the H atoms of N—H groups are directed towards aromatic rings, and these represent genuine hydrogen bonds (Mitchell *et al.*, 1994). It is clearly essential to consider the donor–acceptor geometry, both distances and angles, before assuming an amino–aromatic hydrogen bond; the N···ring distance should be less than ~3.8 Å, and N—H···*C* angle greater than 120°, where *C* is the ring centre (Mitchell *et al.*, 1994).

Acknowledgements

The author gratefully acknowledges Dr Clyde Smith for help with figures, and the Health Research Council of New Zealand and the Howard Hughes Medical Institute for research support.

22.3. Electrostatic interactions in proteins

By K. A. Sharp

22.3.1. Introduction

Electrostatic interactions play a key role in determining the structure, stability, binding affinity, chemical properties, and hence the biological reactivity, of proteins and nucleic acids. Interactions where electrostatics play an important role include:

(1) Ligand/substrate association. Long-range electrostatic forces can considerably enhance association rates by facilitating translational and rotational diffusion or by reduction in the dimensionality of the diffusion space.

(2) Binding affinity. Tight specific binding is often a prerequisite for biological activity, and electrostatics make important contributions to desolvation and formation of chemically complementary interactions during binding.

(3) Modification of chemical and physical properties of functional groups such as cofactors (haems, metal ions *etc.*), alteration of the ionization energy (pK_a) of side chains and shifting of redox midpoints.

(4) The creation of potentials or fields in the active sites to stabilize functionally important charged or dipolar intermediates in processes such as catalysis.

In this chapter I will discuss, within the framework of classical electrostatics, how such effects can be modelled starting from the structural information provided by X-ray crystallography. Nevertheless, many of the concepts of classical electrostatics can be used in combination with molecular dynamics (MD), quantum mechanics (QM) and other computational methods to study a wider range of macromolecular properties, for example specific protein motions, the breaking or forming of bonds, determination of intrinsic pK_a's, determination of electronic energy levels *etc.*

The central aim in studying the electrostatic properties of macromolecules is to take the structural information provided by crystallography (typically the atomic coordinates, although B-factor information may also be of use) and obtain a realistic description of the electrostatic potential distribution $\varphi(\mathbf{r})$. The electrostatic potential distribution can then be used in a variety of ways: (i) graphical analysis may reveal deeper aspects of the structure and help identify functionally important regions or active sites; (ii) the potentials may be used to calculate energies and forces, which can then be used to calculate equilibrium or kinetic properties; and (iii) the electrostatic potentials may be used in conjunction with other computational methods such as QM and MD.

Three problems must be solved to obtain the electrostatic potential distribution. The first is to model the macromolecular charge distribution, usually by specifying the location and charge of all its atoms. Although the coordinates of the molecule are determined by crystallographic methods, the charge distribution is not. A number of atomic charge distributions have been developed for proteins and nucleic acids using quantum mechanical methods and/or parameterization to different experimental data. The second problem is that the positions of the water molecules and solvent ions are generally not known. (Water molecules and ions seen in even the best crystal structures usually constitute a small fraction of the total important in solvating the molecule. Moreover, the orientation of the crystallographic water molecules, crucial in determining the electrostatic potential, is rarely known.) Within the framework of classical electrostatics, inclusion of the *effect* of the solvating water molecules and ions is handled not by treating them explicitly, but implicitly in terms of an 'electrostatic response' to the field created by the molecular charge distribution. The third problem is that incorporation of the available structural information at atomic resolution results in a complicated spatial distribution of charge, dielectric response *etc.* Numerical methods for rapidly and accurately solving the electrostatic equations that determine the potential are therefore essential.

22.3.2. Theory

22.3.2.1. *The response of the system to electrostatic fields*

The response to the electrostatic field arising from the molecular charge distribution arises from three physical processes: electronic polarization, reorientation of permanent dipolar groups and redistribution of mobile ions in the solvent. Movement of ionized side chains, if significant, is sometimes viewed as part of the dielectric response of the protein, and sometimes explicitly as a conformational change of the molecule.

Electronic polarizability can be represented either by point inducible dipoles (Warshel & Åqvist, 1991) or by a dielectric constant. The latter approach relates the electrostatic polarization, $\mathbf{P}(\mathbf{r})$ (the mean dipole moment induced in some small volume V) to the Maxwell (total) field, $\mathbf{E}(\mathbf{r})$, and the local dielectric constant representing the response of that volume, $\varepsilon(\mathbf{r})$, according to

$$\mathbf{P}(\mathbf{r}) = [\varepsilon(\mathbf{r}) - 1]\mathbf{E}(\mathbf{r})/4\pi. \qquad (22.3.2.1)$$

The contribution of electronic polarizability to the dielectric constant of most organic material and water is fairly similar. It can be evaluated by high-frequency dielectric measurements or the refractive index, and it is in the range 2–2.5.

The reorientation of groups such as the peptide bond or surrounding water molecules which have large permanent dipoles is an important part of the overall response. This response too may be treated using a dielectric constant, *i.e.* using equation (22.3.2.1) with a larger value of the dielectric constant that incorporates the additional polarization from dipole reorientation. An alternative approach to equation (22.3.2.1) for treating the dipole reorientation contribution of water surrounding the macromolecules is the Langevin dipole model (Lee *et al.*, 1993; Warshel & Åqvist, 1991; Warshel & Russell, 1984). Four factors determine the degree of response from permanent dipoles: (i) the dipole-moment magnitude; (ii) the density of such groups in the protein or solvent; (iii) the freedom of such groups to reorient; and (iv) the degree of cooperativity between dipole motions. Thus, water has a high dielectric constant ($\varepsilon = 78.6$ at 25 °C). For electrostatic models based on dielectric theory, the experimental solvent dielectric constant, reflecting the contribution of electronic polarizability and dipole reorientation, is usually used. From consideration of the four factors that determine the dielectric response, macromolecules would be expected to have a much lower dielectric constant than the solvent. Indeed, theoretical studies of the dielectric behaviour of amorphous protein solids (Gilson & Honig, 1986; Nakamura *et al.*, 1988) and the interior of proteins in solution (Simonson & Brooks, 1996; Simonson & Perahia, 1995; Smith *et al.*, 1993), and experimental measurements (Takashima & Schwan, 1965) provide an estimate of $\varepsilon = 2.5$–4 for the contribution of dipolar groups to the protein dielectric.

The Langevin model can account for the saturation of the response at high fields that occurs if the dipoles become highly aligned with the field. The dielectric model can also be extended to incorporate saturation effects (Warwicker, 1994), although there is a compensating effect of electrostriction, which increases the local dipole density (Jayaram, Fine *et al.*, 1989). While the importance of saturation effects would vary from case to case, linear solvent dielectric models have proven sufficiently accurate for most protein applications to date.

Charge groups on molecules will attract solvent counter-ions and repel co-ions. The most common way of treating this charge rearrangement is *via* the Boltzmann model, where the net charge density of mobile ions is given by

$$\rho^m(\mathbf{r}) = \sum_i z_i e c_i^o \exp[-z_i e \varphi(\mathbf{r})/kT], \qquad (22.3.2.2)$$

where c_i^o is the bulk concentration of an ion of type i, valence z_i, and $\varphi(\mathbf{r})$ is the average potential (an approximation to the potential of mean force) at position **r**. The Boltzmann approach neglects the effect of ion size and correlations between ion positions. Other models for the mobile-ion behaviour that account for these effects are integral equation models and MC models (Bacquet & Rossky, 1984; Murthy *et al.*, 1985; Olmsted *et al.*, 1989, 1991; Record *et al.*, 1990). These studies show that ion size and correlation effects do not compromise the Boltzmann model significantly for monovalent (1–1) salts at mid-range concentrations 0.001–0.5 *M*, and consequently it is widely used for modelling salt effects in proteins and nucleic acids.

22.3.2.2. *Dependence of the potential on the charge distribution*

The potential at a point in space, **r**, arising from some charge density distribution $\rho(\mathbf{s})$ and some dipole density distribution $\mathbf{P}(\mathbf{s})$ (which includes polarization) is given by

$$\varphi(\mathbf{r}) = \textstyle\int \rho(\mathbf{s})/(|\mathbf{s}-\mathbf{r}|) + \mathbf{P}(\mathbf{s})(\mathbf{s}-\mathbf{r})/(|\mathbf{s}-\mathbf{r}|^3)\ \mathrm{d}\mathbf{s}. \qquad (22.3.2.3)$$

The total charge distribution is the sum of the explicit charge distribution on the molecule and that from the mobile solvent ion distribution, $\rho = \rho^e + \rho^m$. Substituting for the dielectric polarization using equation (22.3.2.1) and for the mobile ion charge distribution using equation (22.3.2.2), the potential may be expressed in terms of a partial differential equation, the Poisson–Boltzmann (PB) equation:

$$\nabla\varepsilon(\mathbf{r})\nabla\varphi(\mathbf{r}) + 4\pi\sum_i z_i e c_i^o \exp[-z_i e\varphi(\mathbf{r})/kT] + 4\pi\rho^e(\mathbf{r}) = 0, \qquad (22.3.2.4)$$

which relates the potential, molecular charge and dielectric distributions, $\varphi(\mathbf{r})$, $\rho^e(\mathbf{r})$ and $\varepsilon(\mathbf{r})$, respectively. Contributions to the polarizability from electrons, a molecule's permanent dipoles and solvent dipoles are incorporated into this model by using an appropriate value for the dielectric for each region of protein and solvent. Values for protein atomic charges, radii and dielectric constants suitable for use with the Poisson–Boltzmann equation are available in the literature (Jean-Charles *et al.*, 1990; Mohan *et al.*, 1992; Simonson & Brünger, 1994; Sitkoff *et al.*, 1994). For protein applications, the Boltzmann term in equation (22.3.2.4) is usually linearized to become $-8\pi\varphi(\mathbf{r})I/kT$ where I is the ionic strength, whereas for nucleic acids and molecules of similarly high charge density the full nonlinear equation is used.

22.3.2.3. *The concepts of screening, reaction potentials, solvation, dielectric, polarity and polarizability*

Application of a classical electrostatic view to macromolecular electrostatics involves a number of useful concepts that describe the physical behaviour. It should first be recognized that the potential at a particular charged atom i includes three physically distinct contributions. The first is the direct or Coulombic potential of j at i. The second is the potential at i generated by the polarization (of a molecule, water and ion atmosphere) induced by j. This is often referred to as the screening potential, since it opposes the direct Coulombic potential. The third arises from the polarization induced by i itself. This is often referred to as the reaction or self-potential, or if solvent is involved, as the solvation potential.

When using models that apply the concept of a dielectric constant (a measure of polarizability) to a macromolecule, it is important to distinguish between polarity and polarizability. Briefly, polarity may be thought of as describing the density of charged and dipolar groups in a particular region. Polarizability, by contrast, refers to the *potential* for reorganizing charges, orienting dipoles and inducing dipoles. Thus polarizability depends both on the polarity and the freedom of dipoles to reorganize in response to an applied electric field. When a protein is folding or undergoing a large conformational rearrangement, the peptide groups may be quite free to reorient. In the folded protein, these may become spatially organized so as to stabilize another charge or dipole, creating a region with high polarity, but with low polarizability, since there is much less ability to reorient the dipolar groups in response to a new charge or dipole without significant disruption of the structure. Thus, while there is still some discussion about the value and applicability of a protein dielectric constant, it is generally agreed that the interior of a macromolecule is a less polarizable environment compared to solvent. This difference in polarizability has a significant effect on the potential distribution.

Formally charged groups on proteins, particularly the longer side chains on the surface of proteins, Arg, Lys, and to a lesser extent Glu and Asp, have the ability to alter their conformation in response to electrostatic fields. In addition, information about fluctuations about their mean position may need to be included in calculating average properties. Three approaches to modelling protein formal charge movements can be taken. The first is to treat the motions within the dielectric response. In this approach, the protein may be viewed as having a dielectric higher than 2.5–4 in the regions of these charged groups, particularly at the surface, where the concentration and mobility of these groups may give an effective dielectric of 20 or more (Antosiewicz *et al.*, 1994; Simonson & Perahia, 1995; Smith *et al.*, 1993). A second approach is to model the effect of charge motions on the electrostatic quantity of interest explicitly, *e.g.* with MD simulations (Langsetmo *et al.*, 1991; Wendoloski & Matthew, 1989). This involves generating an ensemble of structures with different atomic charge distributions. The third approach is based on the fact that one is often interested in a specific biological process $A \rightarrow B$ in which one can evaluate the structure of the protein in states A and B (experimentally or by modelling), and any change in average charge positions is incorporated at the level of different average explicit charge distribution inputs for the calculation, modelling only the electronic, dipolar and salt contributions as the response.

The term 'effective' dielectric constant is sometimes used in the literature to describe the strength of interaction between two charges, q_1 and q_2. This is defined as the ratio of the observed or calculated interaction strength, U, to that expected between the same two charges in a vacuum:

$$\varepsilon^{\rm eff} = [(q_1q_2)/r_{12}]/U, \qquad (22.3.2.5)$$

where r_{12} is the distance between the charges. If the system were completely homogeneous in terms of its electrostatic response and involved no charge rearrangement then $\varepsilon^{\rm eff}$ would describe the dielectric constant of the medium containing the charges. This is generally never the case: the strength of interaction in a protein system is determined by the net contribution from protein, solvent and ions, so $\varepsilon^{\rm eff}$ does not give information about the dielectric property of any particular region of space. In fact, in the same system different charge–charge interactions will generally yield different values of $\varepsilon^{\rm eff}$. Thus $\varepsilon^{\rm eff}$ is really no more than its definition – a measure of the strength of interaction – and it cannot be used directly to answer questions about the protein dielectric constant, for example. Rather, it is one of the quantities that one aims to extract from theoretical models to compare with an experiment.

22.3.2.4. *Calculation of energies and forces*

Once the electrostatic potential distribution has been obtained, calculation of experimental properties usually requires evaluation of the electrostatic energy or force. For a linear system (where the dielectric and ionic responses are linear) the electrostatic free energy is given by

$$\Delta G^{\rm el} = 1/2\sum_i \varphi_i q_i, \qquad (22.3.2.6)$$

where $\varphi_{\rm i}$ is the potential at an atom with charge q_i. The most common source of nonlinearity is the Boltzmann term in the PB equation (22.3.2.4) for highly charged molecules such as nucleic acids. The total electrostatic energy in this case is (Reiner & Radke, 1990; Sharp & Honig, 1990; Zhou, 1994)

$$\Delta G^{\rm el} = \int_V \{\rho^e\varphi - (\varepsilon E^2/8\pi) - kT\sum_i c_i^0[\exp(-z_i e\varphi/kT) - 1]\}\,{\rm d}{\bf r}, \qquad (22.3.2.7)$$

where the integration is now over all space.

The general expression for the electrostatic force on a charge q is given by the gradient of the total free energy with respect to that charge's position,

$$\mathbf{f}_q = -\nabla_{\mathbf{r}q}(G^{\rm el}). \qquad (22.3.2.8)$$

If the movement of that charge does not affect the potential distribution due to the other charges and dipoles, then equation (22.3.2.8) can be evaluated using the 'test charge' approach, in which case the force depends only on the gradient of the potential or the field at the charge:

$$\mathbf{f} = q\mathbf{E}. \qquad (22.3.2.9)$$

However, in a system like a macromolecule in water, which has a non-homogeneous dielectric, forces arise between a charge and any dielectric boundary due to image charge (reaction potential) effects. A similar effect to the 'dielectric pressure' force arises from solvent-ion pressure at the solute–solvent boundary. This results in a force acting to increase the solvent exposure of charged and polar atoms. An expression for the force that includes these effects has been derived within the PB model (Gilson *et al.*, 1993):

$$\mathbf{f} = \rho^e\mathbf{E} - (1/2)E^2\nabla\varepsilon - kT\sum_i c_i^0[\exp(-z_i e\varphi/kT) - 1]\nabla A, \qquad (22.3.2.10)$$

where A is a function describing the accessibility to solvent ions, which is 0 inside the protein, and 1 in the solvent, and whose gradient is nonzero only at the solute–solvent surface. Similarly, in a two-dielectric model (solvent plus molecule) the gradient of ε is nonzero only at the molecular surface. The first term accounts for the force acting on a charge due to a field, as in equation (22.3.2.9), while the second and third terms account for the dielectric surface pressure and ionic atmosphere pressure terms respectively. Equation (22.3.2.10) has been used to combine the PB equation and molecular mechanics (Gilson *et al.*, 1995).

22.3.2.5. *Numerical methods*

A variety of numerical methods exist for calculating electrostatic potentials of macromolecules. These include numerical solution of self-consistent field electrostatic equations, which has been used in conjunction with the protein dipole–Langevin dipole method (Lee *et al.*, 1993). Numerical solution of the Poisson–Boltzmann equation requires the solution of a three-dimensional partial differential equation, which can be nonlinear. Many numerical techniques, some developed in engineering fields to solve differential equations, have been applied to the PB equation. These include finite-difference methods (Bruccoleri *et al.*, 1996; Gilson *et al.*, 1988; Nicholls & Honig, 1991; Warwicker & Watson, 1982), finite-element methods (Rashin, 1990; Yoon & Lenhoff, 1992; Zauhar & Morgan, 1985), multigridding (Holst & Saied, 1993; Oberoi & Allewell, 1993), conjugate-gradient methods (Davis & McCammon, 1989) and fast multipole methods (Bharadwaj *et al.*, 1994; Davis, 1994). Methods for treating the nonlinear PB equation include under-relaxation (Jayaram, Sharp & Honig, 1989) and powerful inexact Newton methods (Holst *et al.*, 1994). The nonlinear PB equation can also be solved *via* a self-consistent field approach, in which one calculates the potential using equation (22.3.2.5), then the mobile charge density is calculated using equation (22.3.2.3), and the procedure is repeated until convergence is reached (Pack & Klein, 1984; Pack *et al.*, 1986). The method allows one to include more elaborate models for the ion distribution, for example incorporating the finite size of the ions (Pack *et al.*, 1993). Approximate methods based on spherical approximations (Born-type models) have also been used (Schaeffer & Frommel, 1990; Still *et al.*, 1990). Considerable numerical progress has been made in finite methods, and accurate rapid algorithms are available. The reader is referred to the original references for numerical details.

22.3.3. Applications

An exhaustive list of applications of classical electrostatic modelling to macromolecules is beyond the scope of this chapter. Three general areas of application are discussed.

22.3.3.1. *Electrostatic potential distributions*

Graphical analysis of electrostatic potential distributions often reveals features about the structure that complement analysis of the atomic coordinates. For example, Fig. 22.3.3.1(*a*) shows the distribution of charged residues in the binding site of the proteolytic enzyme thrombin. Fig. 22.3.3.1(*b*) shows the resulting electrostatic potential distribution on the protein surface. The basic (positive) region in the fibrinogen binding site, which could be inferred from close inspection of the distribution of charged residues in Fig. 22.3.3.1(*a*), is clearly more apparent in the potential distribution. Fig. 22.3.3.1(*c*) shows the effect of increasing ionic strength on the potential distribution, shrinking the regions of strong potential. Fig. 22.3.3.1(*d*) is calculated assuming the same dielectric for the solvent and protein. The more uniform potential distribution compared to Fig. 22.3.3.1(*b*) shows the focusing effect that the low dielectric interior has on the field emanating from charges in active sites and other cleft regions.

22.3.3.2. *Charge-transfer equilibria*

Charge-transfer processes are important in protein catalysis, binding, conformational changes and many other functions. The primary examples are acid–base equilibria, electron transfer and ion binding, in which the transferred species is a proton, an electron or a salt ion, respectively. The theory of the dependence of these three equilibria within the classical electrostatic framework can be treated in an identical manner, and will be illustrated with acid–base equilibria. A titratable group will have an intrinsic ionization equilibrium, expressed in terms of a known intrinsic pK_a^0, where p$K_a^0 = -\log_{10}(K_a^0)$, K_a^0 is the dissociation constant for the reaction ${\rm H}^+A = {\rm H}^+ + A$ and A can be an acid or a base. The pK_a^0 is determined by all the quantum-chemical, electrostatic and environmental effects operating on that group in some reference state. For example, a reference state for the aspartic acid side-chain ionization might be the isolated amino acid in water, for which $pK_a^0 = 3.85$. In the environment of the protein, the pK_a will be altered by three electrostatic effects. The first occurs because the group is positioned in a protein environment with a different polarizability, the second is due to interaction with permanent dipoles in the protein, the third is due to charged, perhaps titratable, groups. The effective pK_a is given by

$$pK_a = pK_a^0 + (\Delta\Delta G^{\rm rf} + \Delta\Delta G^{\rm perm} + \Delta\Delta G^{\rm tit})/2.303kT, \qquad (22.3.3.1)$$

where the factor of $1/2.303kT$ converts units of energy to units of pK_a. The first contribution, $\Delta\Delta G^{\rm rf}$, arises because the completely solvated group induces a strong favourable reaction field (see Section 22.3.2.3) in the high dielectric water, which stabilizes the charged form of the group. (The neutral form is also stabilized by the solvent reaction field induced by any dipolar groups, but to a lesser extent.) Desolvating the group to any degree by moving it into a less polarizable environment will preferentially destabilize the charged form of that group, shifting the pK_a by an amount

$$\Delta\Delta G^{\rm rf} = (1/2)\sum_i \left(q_i^d \Delta\varphi_i^{{\rm rf},d} - q_i^p \Delta\varphi_i^{{\rm rf},p}\right), \qquad (22.3.3.2)$$

where q_i^p and q_i^d are the charge distributions on the group, $\Delta\varphi_i^{{\rm rf},p}$ and $\Delta\varphi_i^{{\rm rf},d}$ are the changes in the group's reaction potential upon moving it from its reference state into the protein, in the protonated (superscript p) and deprotonated (superscript d) forms, respectively, and the sum is over the group's charges. The contribution of the permanent dipoles is given by

$$\Delta\Delta G^{\rm tit} = \sum_i (q_i^d - q_i^p)\varphi_i^{\rm perm}, \qquad (22.3.3.3)$$

where $\varphi_i^{\rm perm}$ is the interaction potential at the ith charge due to all the permanent dipoles in the protein, including the effect of screening. It is observed that intrinsic pK_a's of groups in proteins are rarely shifted by more than 1 pK_a unit, indicating that the effects of desolvation are often compensated to a large degree by the $\Delta\Delta G^{\rm perm}$ term (Antosiewicz *et al.*, 1994). The final term accounts for the contribution of all the other charged groups:

$$\Delta\Delta G^{\rm tit} = \sum_i \left(q_i^d \langle\varphi_i\rangle^d_{{\rm pH},c,\Delta V} - q_i^p \langle\varphi_i\rangle^p_{{\rm pH},c,\Delta V}\right), \qquad (22.3.3.4)$$

where $\langle\varphi_i\rangle$ is the mean potential at group charge i from all the other titratable groups. The charge states of the other groups in the protein depend in turn on their intrinsic 'pK_a's', on the external pH if they are acid–base groups, the external redox potential, ΔV, if they are redox groups and the concentration of ions, c, if they are ion-binding sites, as indicated by the subscript to $\langle\varphi_i\rangle$. Moreover, the charge state of the group itself will affect the equilibrium at the other sites. Because of this linkage, exact determination of the complete charged state of a protein is a complex procedure. If there are N such groups, the rigorous approach is to compute the titration-state partition function by evaluating the relative electrostatic free energies of all 2^N ionization states for a given set of pH, c, ΔV. From this one may calculate the mean ionization state of any group as a function of pH, ΔV *etc*. For large N this becomes impractical, but various approximate schemes work well, including a Monte Carlo procedure (Beroza *et al.*, 1991; Yang *et al.*, 1993) or partial evaluation of the titration partition function by clustering the groups into strongly interacting sub-domains (Bashford & Karplus, 1990; Gilson, 1993; Yang *et al.*, 1993).

Calculation of ion-binding and electron-transfer equilibria in proteins proceeds exactly as for calculation of acid–base equilibria, the results usually being expressed in terms of an association constant, K_a, or a redox midpoint potential E_m (defined as the external reducing potential at which the group is half oxidized and half reduced, usually at pH 7), respectively.

22.3.3.3. *Electrostatic contributions to binding energy*

The electrostatic contribution to the binding energy of two molecules is obtained by taking the difference in total electrostatic energies in the bound (AB) and unbound $A + B$ states. For the linear case,

$$\Delta\Delta G^{\rm elec}_{\rm bind} = (1/2)\sum_i^{N_A} q_i^A(\varphi_i^{AB} - \varphi_i^A) + (1/2)\sum_j^{N_B} q_i^B(\varphi_j^{AB} - \varphi_j^B), \qquad (22.3.3.5)$$

where the first and second sums are over all charges in molecule A and B, respectively, and φ^x is the total potential produced by $x = A$, B, or AB. From equation (22.3.3.5), it should be noted that the electrostatic free energy change of each molecule has contributions from intermolecular charge–charge interactions, and from changes in the solvent reaction potential of the molecule itself when solvent is displaced by the other molecule. Equation (22.3.3.5) allows for the possibility that the conformation may change upon binding, since different charge distributions may be used for the complexed and uncomplexed forms of A, and similarly for B. However, other energetic terms, including those involved in any conformational change, have to be added to equation (22.3.3.5) to obtain net binding free energy changes. Nevertheless, changes in binding free energy due to charge modifications or changes in external factors such as pH and salt concentration may be estimated using equation (22.3.3.5) alone. For the latter, salt effects are usually only significant in highly charged molecules, for which the nonlinear form for the total electrostatic energy, equation (22.3.2.4), must be used. The salt dependence of binding of drugs and proteins to DNA has been studied using this approach (Misra, Hecht *et al.*, 1994; Misra, Sharp *et al.*, 1994; Sharp *et al.*, 1995), including the pH dependence of drug binding (Misra & Honig, 1995). Other applications include the binding of sulfate to the sulfate binding protein (Åqvist *et al.*, 1991) and antibody and antigen interactions (Lee *et al.*, 1992; Slagle *et al.*, 1994).

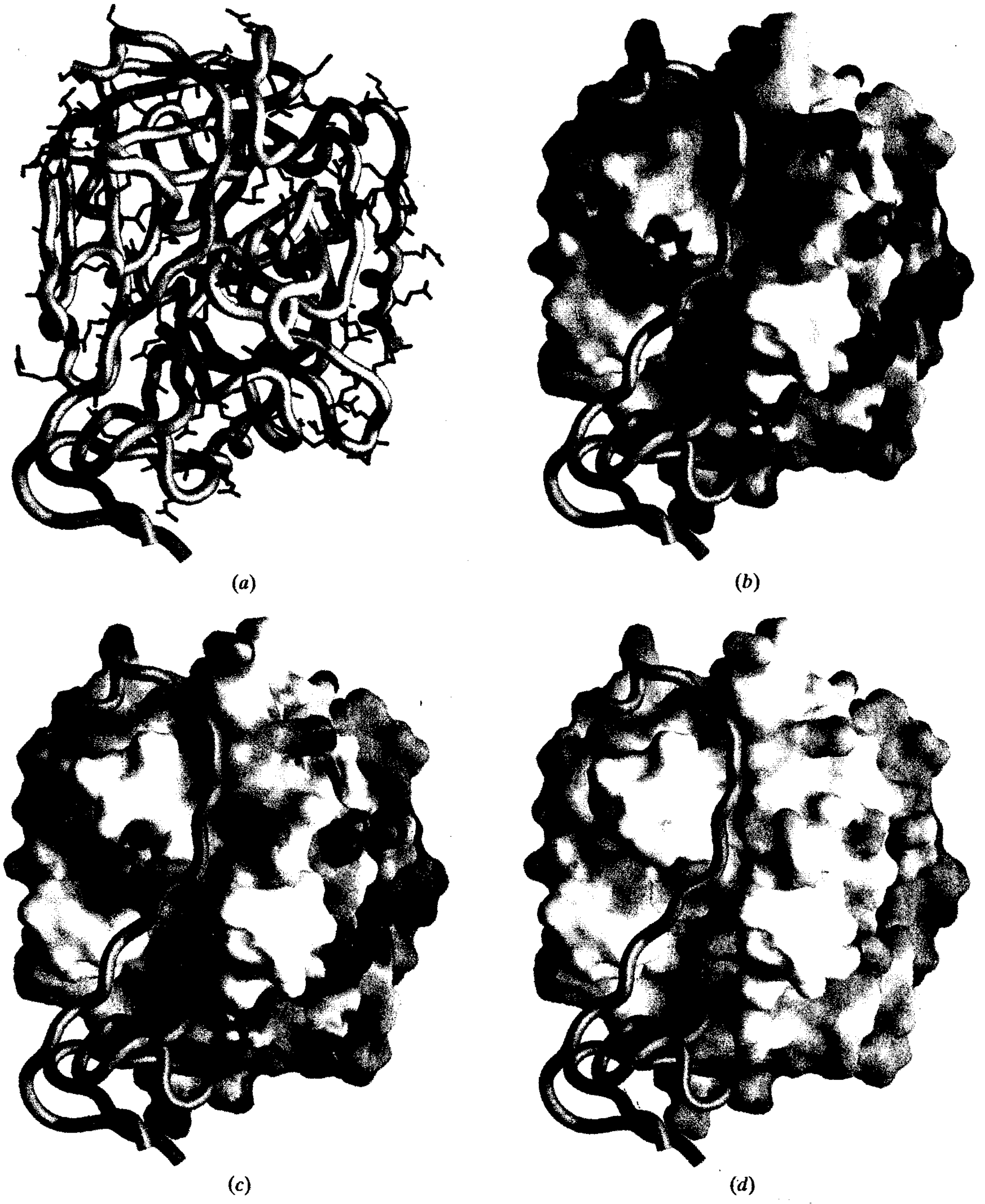

Fig. 22.3.3.1. (*a*) The proteolytic enzyme thrombin (yellow backbone worm) complexed with an inhibitor, hirudin (blue backbone worm). The negatively charged (red) and positively charged (blue) side chains of thrombin are shown in bond representation. (*b*) Solvent-accessible surface of thrombin coded by electrostatic potential (blue: positive, red: negative). Hirudin is shown as a blue backbone worm. Potential is calculated at zero ionic strength. (*c*) Solvent-accessible surface of thrombin coded by electrostatic potential (blue: positive, red: negative). Hirudin is shown as a blue backbone worm. Potential is calculated at physiological ionic strength (0.145 *M*). (*d*) Solvent-accessible surface of thrombin coded by electrostatic potential (blue: positive, red: negative). Hirudin is shown as a blue backbone worm. Potential is calculated using the same polarizability for protein and solvent.

22.4. The relevance of the Cambridge Structural Database in protein crystallography

By F. H. Allen, J. C. Cole and M. L. Verdonk

22.4.1. Introduction

At its inception in the late 1960s, the Cambridge Structural Database (CSD: Allen, Davies *et al.*, 1991; Kennard & Allen, 1993) was one of the first scientific databases for which numerical data were the primary objective of the compilation. Thus, the CSD provides not only a fully retrospective bibliography of the structure determination of organic and metallo-organic compounds, but also gives immediate access to the primary results of each diffraction experiment: the space group, cell dimensions and fractional coordinates that define each structure at atomic resolution. In the late 1960s, the world output of small-molecule structures was just a few hundred per year and it was possible to use existing printed compilations to ensure that the developing CSD was fully retrospective. Despite this comprehensive nature, it has taken time for the CSD to have significant scientific impact as a research tool in its own right, and to be recognized as a source of structural knowledge that is applicable across a broad spectrum of structural chemistry.

There are two reasons for this rather gradual uptake. First, it took time to devise and implement software for the validation and organization of the data. Secondly, and most importantly, it was necessary to develop software for database searching, particularly for locating chemical substructures, and for data analysis and visualization. It was not until the late 1970s that the first comprehensive software systems became available and began to be widely distributed to scientists in academia and industry. Nevertheless, a number of highly influential database analyses were performed prior to 1980, and the proper numerical analysis and statistical treatment of bulk geometrical data began to receive attention (see *e.g.* Murray-Rust & Bland, 1978; Murray-Rust & Motherwell, 1978; Taylor, 1986). This software and its successors at last allowed the types of geometrical surveys, analyses and tabulations carried out manually by early practitioners such as Pauling (1939), Sutton (1956, 1959) and Pimental & McClellan (1960) to be executed automatically in a few minutes of increasingly powerful CPU time.

The early development of applications software simultaneously with methods for the acquisition and validation of new structural data was crucial for the CSD. Developments in structure-determination theory, allied to technological improvements in data collection and the ever increasing speed and capacity of modern computers, led to such a rapid expansion that the archive of May 1999 now contains more than 200000 crystal structures, a total that doubles approximately every seven years. The literature is now so vast, so chemically diverse and so widely spread that it is virtually impossible for individual scientists to maintain current awareness without recourse to database facilities. It is now impossible to carry out viable systematic analyses without recourse to database technology. This chapter focuses primarily on the structural knowledge that is provided by such analyses, and that is relevant to the determination, refinement, validation and systematic study of macromolecular structures. However, the validity of these results depends crucially on two factors: the *completeness* of the archive and the *accuracy* with which the data are recorded. Hence, it is appropriate to preface the chapter with some comparative comment on these fundamental issues as they apply to the small-molecule and macromolecular structure archives.

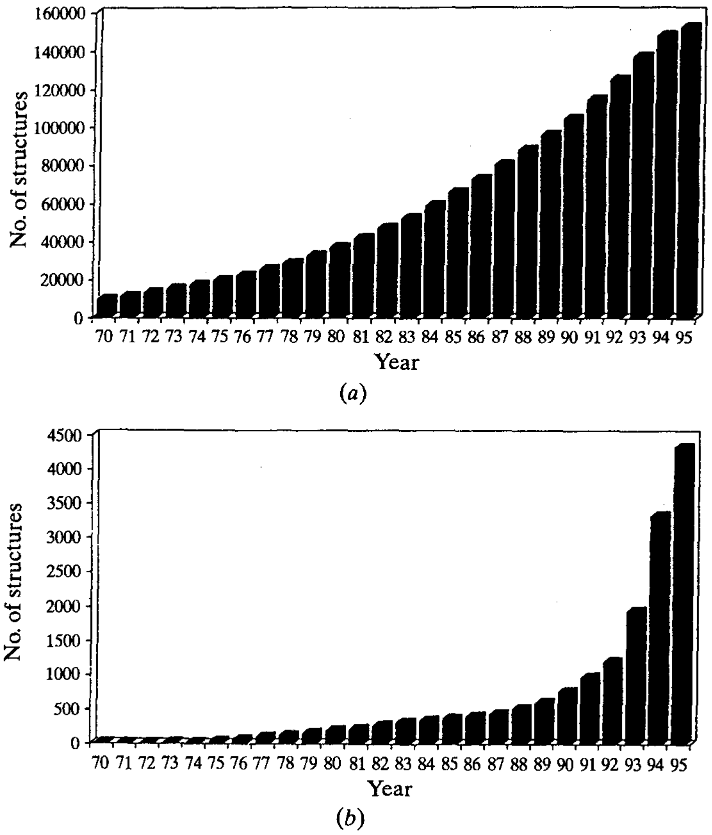

Fig. 22.4.2.1. (*a*) Growth rate of the CSD and (*b*) growth rate of the PDB, in terms of the numbers of structures published per annum for the period 1970–1995.

22.4.2. The CSD and the PDB: data acquisition and data quality

22.4.2.1. *Statistical inferences*

With a current total of 200000 structures and a doubling period of seven years (Fig. 22.4.2.1*a*), we may expect at least half a million small-molecule crystal structures to be in the CSD by the year 2010. The Protein Data Bank (PDB) (Abola *et al.*, 1997; Berman *et al.*, 2000), which began operations in the mid-1970s, and the Nucleic Acid Database (NDB) (Berman *et al.*, 1992) are the international repositories for macromolecular structure information. Input to the PDB was initially slow but is now showing a rapid growth rate reminiscent of the CSD of the 1970s (Fig. 22.4.2.1*b*). The PDB archive has a current total of *ca* 8500 structures (mid-1999) and a doubling period of close to two years. As with the CSD, this early *high rate* of growth will almost certainly decrease, thus increasing the doubling period. Nevertheless, by the year 2010, we might expect the PDB to contain more than 100000 structures.

22.4.2.2. *Data acquisition and completeness*

Given the size and diversity of the CSD, it is amazing that searches for some common chemical substructures often yield far fewer hits than might have been expected. Sometimes, the absence of just a few key CSD entries would have negated a successful systematic analysis: some points in a graph would have been missing and a correlation would not have been detected. Similarly, completeness of the PDB is vital for the future of 'data mining' or 'knowledge engineering' in the macromolecular arena.

Data acquisition by the PDB has always had one valuable advantage in comparison with the CSD. The volume of numerical data generated by a protein structure determination is far too large for primary publication or hard-copy deposition. Thus, the PDB has always acquired data through direct deposition in electronic form, and authors have usually been involved in the validation of their entries. Further, it is a mandatory requirement of the vast majority of journals, and a clear recommendation of appropriate professional organizations, that prior deposition with the PDB is an essential precursor to primary publication. This key involvement of the PDB in the publication process acts as a vital guarantee of the completeness of the archive. The prior-deposition rule must be rigidly adhered to for the long-term benefit of science.

22.4.2.3. *Standard formats: CIF and mmCIF*

The CSD, on the other hand, reflects the published literature, and much of its data content has been re-keyboarded from hard-copy material. The Cambridge Crystallographic Data Centre (CCDC) is now beginning to receive significant amounts of electronic input, a development that owes much to the rapid international acceptance of an agreed standard electronic interchange format, the crystallographic information file or CIF (Hall *et al.*, 1991), and the rapid incorporation of CIF generators within most major structure solution and refinement packages. The CIF offers many advantages, some of which are only just being addressed within the CSD: (*a*) a clear definition of input data items and their representation; (*b*) a significant reduction in time spent correcting simple typographical errors; and (*c*) the possibility of enhancing the overall database content through the electronic availability of *all* information from the analysis, *i.e.* more than could reasonably be re-typed from hard-copy material. For the PDB, the recent adoption of the macromolecular CIF (mmCIF) as the agreed international standard offers similar advantages. This development, together with advances in communications technology, now make it possible to automate the deposition process more effectively, but the advantages of mmCIF can only be fully realized once it also becomes a standard output format of all of the relevant software packages.

22.4.2.4. *Structure validation*

The value of research results derived from the CSD and the PDB depends crucially on the accuracy of the underlying data [see *e.g.* Hooft *et al.* (1996) with respect to protein data]. As with the early CSD, much current research involves use of data from the developing PDB to establish rules and protocols for the validation of new protein structures (see *e.g.* Laskowski *et al.*, 1993). This activity, in turn, means that earlier entries in the archive may have to be reassessed periodically to bring their representations into line with best current practice. This sequence of events was commonplace in the CSD of the 1970s and, even now, new structure types entering the CSD can still provoke a reassessment of subclasses of earlier entries.

Secondly, it is important that errors and warnings raised by validation software have clear meanings and that validation results are clearly encoded within each entry. The end user can then make informed choices about which entries to include (or not) in any given application. Recent moves to apply a range of agreed and unambiguous primary checks to new data, and to require resolution of any problems prior to the issue of a publication ID code, represent an important development.

22.4.3. Structural knowledge from the CSD

22.4.3.1. *The CSD software system*

Structural knowledge from the CSD is reflected principally in the geometries of individual molecules, extended crystal structures and, most importantly, through systematic studies of the geometrical characteristics of large subsets of related substructural units. Software facilities for search, retrieval, analysis and visualization of CSD information are fully described in Chapter 24.3. The system allows for the calculation of a very wide range of geometrical parameters, both intramolecular and intermolecular. Most importantly, chemical substructural search fragments may be specified using normal covalent bonding definitions (single, double, triple *etc.*), limiting non-covalent contact distances and other geometrical constraints. For each instance of a search fragment located in the CSD, the system will compute a user-defined set of geometrical descriptors. The full matrix, $G(N, p)$, of the p geometrical parameters for each of the N fragments located in the CSD can then be analysed using numerical, statistical and visualization techniques to display individual parameter distributions, to compute medians, means and standard deviations, and to examine the geometrical data for correlations or discrete clusters of observations that may exist in the p-dimensional parameter space.

22.4.3.2. *CSD structures and substructures of relevance to protein studies*

Table 22.4.3.1 presents statistics for the 3137 structures of amino acids and peptides that are available in the CSD of April 1998 (containing 181309 entries). Although this represents less than 2% of CSD information, some may consider that these are the only entries of real interest in molecular biology. In certain cases, *e.g.* for the derivation of very precise molecular dimensions and for some conformational work, this may be true. However, the real issue concerns the *transferability* of CSD-derived information to the protein environment. It is the biological relevance of a chemical

Table 22.4.3.1. *Summary of amino-acid and peptide structures available in the CSD (April 1998, 181309 entries)*

(*a*) Overall statistics

Structures	No. of entries
α-Amino acids (any organic) *	3137
Peptides (standard or modified standard α-amino acids) †	1430

(*b*) Peptide statistics

	No. of CSD entries	
No. of residues	Acyclic	Cyclic
2	543	123
3	249	45
4	76	50
5	62	44
6	20	73
7	14	15
8	19	32
10	16	19
11	4	10
12	2	11
13	—	—
14	1	—
15	3	2
16	3	—

* Any organic structure containing the α-amino acid functionality.

† The standard amino acids (those normally found in proteins) may be modified by substitution in these peptides.

Table 22.4.3.2. *CSD entry statistics for selected metal-containing structures*

CSD entries ($R < 0.10$) containing M and (N or O). No additional transition metals were allowed to occur in the Na, K, Mg and Ca structures cited.

Metal	No. of CSD entries
Na	1189
K	987
Mg	510
Ca	469
Zn	1996

substructure (inter- or intramolecular) that is important, and this consideration immediately brings much larger subsets of CSD entries into play. Information such as van der Waals radii can be derived from the CSD as a whole, while more specific information concerning, for example, biologically important metal coordination geometries can be derived from appreciable subsets of the total database, as shown in the statistics of Table 22.4.3.2.

22.4.3.3. *Geometrical parameters of relevance to protein studies*

Precise geometrical knowledge from atomic resolution studies of small molecules is important in the macromolecular domain since it provides: (*a*) geometrical restraints and standards to be applied during protein structure determination, refinement and validation; (*b*) model geometries for liganded small molecules and information about their preferred modes of interaction with the host protein; (*c*) details of metal coordination spheres and geometries that are likely to be observed in metalloproteins; and (*d*) information from which force field and other parameters may be derived. Thus, the types of study discussed in this chapter are concerned with retrieving systematic knowledge concerning:

(1) molecular dimensions: bond lengths and valence angles;

(2) conformational features: torsion angles that describe acyclic and cyclic systems;

(3) metal coordination-sphere geometries: coordination numbers, metal–ligand distances and inter-ligand valence angles;

(4) general non-bonded contact distances: van der Waals radii;

(5) hydrogen-bond geometries: distances, angles, directional properties;

(6) other non-bonded interactions: identification and geometrical description;

(7) formation of preferred atomic arrangements or motifs involving non-covalent interactions.

In this short overview, which deals with such a broad range of structural information, our literature coverage is, of necessity, highly selective. In each area, we have tried to cite the more recent papers, from which leading references to earlier studies can be located. We also draw attention to a number of recent monographs in which a variety of CSD analyses are comprehensively cited and discussed: *Structure Correlation* (Bürgi & Dunitz, 1994), *Crystal Structure Analysis for Chemists and Biologists* (Glusker *et al.*, 1994), *Hydrogen Bonding in Biological Structures* (Jeffrey & Saenger, 1991) and *Crystal Engineering: the Design of Organic Solids* (Desiraju, 1989). Finally, we note the CCDC's own database of published research applications of the CSD. The DBUSE database currently contains literature references and short descriptive abstracts for nearly 700 papers. It forms part of each biannual CSD release and is fully searchable using the *Quest3D* program.

22.4.4. Intramolecular geometry

22.4.4.1. *Mean molecular dimensions*

The work of Pauling (1939) represented the first systematic attempt to derive mean values for bond lengths and valence angles from the limited structural data available at that time. This work resulted in the definition of covalent bonding radii for the common elements and had a seminal influence on the development of chemistry over the past half century. Further tabulations appeared sporadically until the publication in 1956 and 1959 of the major compilation *Tables of Interatomic Distances and Configuration in Molecules and Ions,* edited by Sutton (1956, 1959), by The Chemical Society of London. Kennard (1962) extended the available data for bonds between carbon and other elements.

In the mid-1980s, the CCDC and its collaborators compiled updated tables of mean bond lengths for both organic (Allen *et al.*, 1987) and organometallic and metal coordination compounds (Orpen *et al.*, 1989). Both compilations were based on the CSD of September 1985 containing 49854 entries. Of these, 10324 organic structures and 9802 organometallics or metal complexes satisfied a variety of secondary selection criteria, and were used in the analysis. For each bond length, both compilations present the mean, its estimated standard deviation and the sample standard deviation, together with the median value of the distribution and its upper and lower quartile values. The organic section describes 682 discrete chemical bond types involving 65 element pairs. Of these, 511 (75%) involve carbon, and 428 (63%) involving only carbon, nitrogen and oxygen are relevant to protein studies. The organometallic and metal complex compilation presents similar statistics for 325 different bond types involving *d*- and *f*-block metals. It is planned to automate and systematize the production of such tabulations, so that they can be dynamically updated in computerized form, as part of CCDC's ongoing development of knowledge-based structural libraries.

More recently, Engh & Huber (1991) have generated sets of mean bond lengths and valence angles from peptidic structures retrieved from the 80000 entries then available in the CSD. Their compilations are based on 31 atom types which are most appropriate to the protein environment and are well represented in CSD structures. These authors note that such knowledge, together with torsional and other information, is vital to the determination, refinement and validation of protein structures. Prior to their detailed CSD analysis, some of the parameters used for these purposes had been determined with a lower accuracy than was required by the diffraction data. For this reason, and particularly for use with higher-resolution protein data, they recommend that the most accurate parameters possible should always be used.

Systematic use of CSD data generates mean values together with standard deviations for both the sample and the mean. The sample standard deviations provide information about the spread of each parameter distribution, *i.e.* information about the variability of each parameter which can be parameterized as force constants. Comparative refinements of selected proteins showed that the new CSD-based parameters yielded significant improvements in R factors and in geometry statistics. Finally, Engh & Huber (1991) remark that their results should be updated regularly as the quantity and quality of data in the CSD increase with time. Apart from producing more precise estimates of mean values, incorporation of more protein-relevant atom types into the schema should then be possible.

22.4.4.2. *Conformational information*

Torsion angles are the natural measures of conformational relationships within molecules. If we specify a chemical substructure involving a central bond of interest, then the CSD system

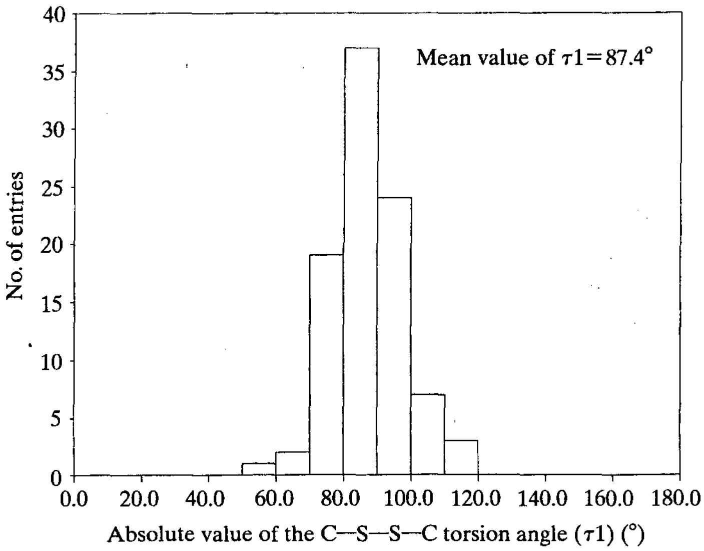

Fig. 22.4.4.1. Distribution of torsion angles in C(sp^3)—S—S—C(sp^3) substructures located in the CSD.

will display the distribution of torsion angles about that bond, computed from the tens, hundreds, or even thousands of instances located in the database. Examination of these *univariate distributions* will reveal any conformational preferences that may exist in small-molecule crystal structures. This approach is illustrated by the histogram of Fig. 22.4.4.1, which shows the torsional distribution about S—S bridge bonds in C(sp^3)—S—S—C(sp^3) substructures located in the CSD. Clearly, there is a preference for a perpendicular conformation in the CS—SC unit. This corresponds well with values observed for cysteine bridges in protein structures, and with theoretical calculations on small model compounds.

The interrelationship between two torsion angles can be visualized by plotting them against each other on a conventional 2D scattergram. In the small-molecule area, the distribution of data points in these scattergrams can reveal conformational interconversion pathways (Rappoport *et al.*, 1990) or show areas of high data density corresponding to conformational preferences (Schweizer & Dunitz, 1982). The best known *bivariate distribution* is the Ramachandran plot of peptidic φ–ψ angles, which is universally used to assess the quality of protein structures and to identify structural features. Ashida *et al.* (1987) performed an extensive analysis of peptide conformations available in the CSD and present torsional histograms, a Ramachandran plot, and a variety of other visual and descriptive statistics that summarize this data set.

It is often necessary to use three or more torsion angles to define the conformation of, *e.g.*, a side chain or flexible ring. Here, *multivariate* statistical techniques (Chatfield & Collins, 1980; Taylor, 1986) have proved valuable for extracting information from the matrix $T(N, k)$ that contains the k torsion angles computed for each of the N examples of the substructure in the CSD. Two methods, both available within the CSD system software described in Chapter 24.3, are commonly used to visualize the k-dimensional data set and to locate natural sub-groupings of data points within it.

Principal component analysis (PCA) (Murray-Rust & Motherwell, 1978; Allen, Doyle & Auf der Heyde, 1991, Allen, Howard & Pitchford, 1996) is a dimension-reduction technique which analyses the variance in $T(N, k)$ in terms of a new set of uncorrelated, orthogonal variables: the principal components, or PCs. The PCs are generated in decreasing order of the percentage of the variance that is explained by each of them. The hope is that the number of PCs, p, that explains most of the variance in the data set is such that $p \ll k$, so that a few pairwise scatter plots with respect to the new PC axes will provide useful visualizations of the complete data set. For cyclic fragments, PCA results are closely related to those obtained using the ring-puckering methodology of Cremer & Pople (1975). Cluster analysis (CA) (Everitt, 1980; Allen, Doyle & Taylor, 1991) is a purely numerical method that attempts to locate discrete groupings of data points within a multivariate data set. CA uses 'distances' or 'dissimilarities' between pairs of points in a k-dimensional space as its working basis, and a very large number of clustering algorithms exist. The mathematical basis of both of these techniques, the modifications that are needed to account for topological symmetry in the search fragment and examples of their application have been reviewed by Taylor & Allen (1994).

Preliminary work using the concepts of machine learning (Carbonell, 1989) for knowledge discovery and classification have also been carried out using the CSD (see *e.g.* Allen *et al.*, 1990; Fortier *et al.*, 1993). In particular, conceptual clustering methods have been applied to a number of substructures (Conklin *et al.*, 1996) and the results compared with those obtained by the statistical and numerical methods described above. Similar techniques are also being used for the classification of protein structures (see *e.g.* Blundell *et al.*, 1987).

22.4.4.3. *Crystallographic conformations and energies*

Crystallographic conformations obviously represent energetically accessible forms. However, for use in molecular-modelling applications, the key question must be asked: Are the condensed-phase crystallographic observations a good guide to conformational preferences in other phases? The indications are that the answer is 'yes' from the types of studies exemplified or cited in the previous section: there appears to be a clear *qualitative* relationship between crystallographic conformer distributions and the low-energy features of the appropriate potential energy hypersurface, although the estimation of absolute energies from the relative populations of these distributions is not appropriate (Bürgi & Dunitz, 1988).

Allen, Harris & Taylor (1996) addressed this question in a systematic manner for a series of 12 one-dimensional (univariate) conformational problems. All of the chosen substructures [simple derivatives of ethane, involving a single torsion angle (τ) about the central C—C bond] were expected to show one symmetric (*anti*, $\tau \simeq 180°$) energy minimum and two symmetry-related asymmetric (*gauche*, $\tau \simeq \pm 60°$) minima. For each substructure, the crystallographic torsional distribution was determined from the CSD and compared with the 1D potential-energy profile, computed using *ab initio* molecular-orbital methods and the 6-31G* basis set. Close agreement was observed between the experimental condensed phase results and the computed *in vacuo* data. Taken over all 12 substructures, the *ab initio* optimized values of the asymmetric (*gauche*) torsion angle vary from $< 55°$ to $> 80°$, and a scatter plot of these optimized values *versus* the mean crystallographic values for *gauche* conformers is linear, with a correlation coefficient of 0.831. Two other results of the study were that: (*a*) torsion angles with higher strain energies ($>$4.5 kJ mol^{-1}) are rarely observed in crystal structures ($<$5%); and (*b*) taken over many structures, conformational distortions due to crystal packing appear to be the exception rather than the rule.

22.4.4.4. *Conformational libraries*

In essence, the CSD can be regarded as a huge library of individual molecular conformations. However, to be of general value, it is necessary to distil, store and present this knowledge in an ordered manner, in the form of torsional distributions for specific atomic tetrads A—B—C—D. Protein-specific libraries of this type derived from high-resolution PDB structures are commonly used as aids to protein structure determination, refinement and validation (Bower *et al.*, 1997; Dunbrack & Karplus, 1993). The information can either be stored in external databases, or hardwired into the program in the form of rules. However, CSD usage has tended to concentrate on analyses of individual substructures, as noted above, both for their intrinsic interest and to develop novel methods of data analysis. Recently, Klebe & Mietzner (1994) have described the generation of a small library containing 216 torsional distributions derived from the CSD, together with 80 determined from protein–ligand complexes in the PDB. The library was used in a knowledge-based approach for predicting multiple conformer models for putative ligands in the computational modelling of protein–ligand docking. Conformer prediction is accomplished by the computer program *MIMUMBA*. As part of its programme for the development of knowledge-based libraries from the CSD, the CCDC has now embarked on the generation of a more comprehensive torsional library. Here, information is being hierarchically ordered according to the level of specificity of the chemical substructures for which torsional distributions are available in the library.

22.4.4.5. *Metal coordination geometry*

Some 54% of the information content of the CSD relates to organometallics and metal complexes. This reflects the crucial role of single-crystal diffraction analyses in the renaissance of inorganic chemistry since the 1950s, and the fundamental importance of the technique in characterizing the many novel molecules synthesized over the past 40 years. Since ligands containing nitrogen, oxygen and sulfur are ubiquitous, the CSD contains much information that is relevant to the binding of metal ions by proteins [*e.g.* zinc (Miller *et al.*, 1985), calcium (Strynadka & James, 1989) *etc.*]. Some statistics for the occurrence of some common metals having N and/or O ligands are presented in Table 22.4.3.2.

One of the earliest studies (Einspahr & Bugg, 1981) concerned the geometry of Ca–carboxylate binding, with special reference to biological systems. Since that time, a variety of other studies of biologically relevant metal coordination modes have appeared from the laboratories of Glusker, Dunitz and others (see *e.g.* Glusker, 1980; Chakrabarti & Dunitz, 1982; Carrell *et al.*, 1988, 1993; Chakrabarti, 1990*a*,*b*). These studies show, *inter alia*, that α-hydroxycarboxylates and imidazoles such as histidine tend to bind metal ions in their planes, but that alkali metal cations tend to bind carboxylate groups indiscriminately both in-plane and out-of-plane. Chapter 17 of Glusker *et al.* (1994) is a significant source of additional information and leading references to work in this area over the past two decades.

22.4.5. Intermolecular data

Non-bonded interaction geometries observed in small-molecule crystal structures are of great value in the determination and validation of protein structures, in furthering our understanding of protein folding, and in investigating the recognition processes involved in protein–ligand interactions. The CSD continues to provide vital information on all of these topics.

22.4.5.1. *van der Waals radii*

The hard-sphere atomic model is central to chemistry and molecular biology and, to an approximation, atomic van der Waals radii can be regarded as transferable from one structure to another. They are heavily used in assessing the general correctness of all crystal-structure models from metals and alloys to proteins. Pauling (1939) was the first to provide a usable tabulation for a wide range of elements, but the values of Bondi (1964) remain the most highly cited compilation in the modern literature. His values, assembled from a variety of sources including crystal-structure information, were selected for the calculation of molecular volumes and, in his original paper, Bondi (1964) issues a caution about their general validity for the calculation of limiting contact distances in crystals. In view of the huge amount of non-bonded contact information available in the CSD, Rowland & Taylor (1996) recently tested Bondi's statement as it might apply to the common nonmetallic elements, *i.e.* H, C, N, O, F, P, S, Cl, Br and I. They found remarkable agreement (within 0.02 Å) between the crystal-structure data and the Bondi values for S and the halogens, and agreement within 0.05 Å for C, N and O (new values all larger). The only significant discrepancy was for H, where averaged neutron-normalized small-molecule data yield a van der Waals radius of 1.1 Å, 0.1 Å shorter than the Bondi (1964) value. In the specific area of amino-acid structure, Gould *et al.* (1985) have studied the crystal environments and geometries of leucines, isoleucines, valines and phenylalanines. Their work provides estimates of minimum non-bonded contact distances and indicates the preferred van der Waals interactions of these primary building blocks.

22.4.5.2. *Hydrogen-bond geometry and directionality*

The hydrogen bond is the strongest and most frequently studied of the non-covalent interactions that are observed in crystal structures. As with intramolecular geometries, the first surveys of non-bonded interaction geometries all concerned hydrogen bonds, and were reported long before the CSD existed (Pauling, 1939; Donohue, 1952; Robertson, 1953; Pimentel & McClellan, 1960). The review by Donohue (1952) already contained a plot of N···O distances *versus* C—N···O angles in crystal structures (the C—N groups are terminal charged amino groups), while the review by Pimentel & McClellan (1960) contained histograms of hydrogen-bond distances. Up to the mid-1970s, numerous other studies appeared, *e.g.* Balasubramanian *et al.* (1970), Kroon & Kanters (1974) and Kroon *et al.* (1975), in which all of the statistical analyses were performed manually.

With the advent of the CSD and its developing software system, these kinds of studies became much more accessible and easier to perform, although the non-bonded search facility was only generalized and fully integrated within *Quest3D* in 1992. Thus, Taylor and colleagues reported studies on N—H···O=C hydrogen bonds (Taylor & Kennard, 1983; Taylor *et al.*, 1983, 1984*a*,*b*), Jeffrey and colleagues reported detailed studies on the O—H···O hydrogen bond (Ceccarelli *et al.*, 1981), hydrogen bonds in amino acids (Jeffrey & Maluszynska, 1982; Jeffrey & Mitra, 1984), and hydrogen bonding in nucleosides and nucleotides, barbiturates, purines and pyrimidines (Jeffrey & Maluszynska, 1986), while Murray-Rust & Glusker (1984) studied the directionalities of O—H···O hydrogen bonds to ethers and carbonyls. These studies indicated that hydrogen bonds are often very directional. For example, the distribution of the O—H···O hydrogen-bond angle, after correction for a geometrical factor, peaks at 180° (*i.e.* there is a clear preference for linear hydrogen bonds) and, in carbonyls and carboxylate groups, hydrogen bonds tend to form along the lone-pair directions of the O-atom acceptors (Fig. 22.4.5.1). For ethers, however, lone-pair directionality is not observed, as is illustrated in Fig. 22.4.5.2.

Software availability has facilitated CSD studies of a wide range of individual hydrogen-bonded systems in the recent literature, including studies of resonance-assisted hydrogen bonds (Bertolasi *et al.*, 1996) and resonance-induced hydrogen bonding to sulfur (Allen, Bird *et al.*, 1997*a*). These statistical studies are often combined with molecular-orbital calculations of interaction energies. Some of these studies are cited in this chapter, but the monograph of Jeffrey & Saenger (1991) and the CCDC's DBUSE database are valuable reference sources.

(*a*)

(*b*)

(*c*)

Fig. 22.4.5.1. The IsoStar knowledge-based library of intermolecular interactions: interaction of O—H donors (contact groups) with one of the $>C=O$ acceptors of a carboxylate group (the central group). (*a*) Direct scatter plot derived from CSD data, (*b*) contoured scatter plot derived from CSD data and (*c*) direct scatter plot derived from PDB data.

Fig. 22.4.5.2. Distribution of O—H donors around ether oxygen acceptors (CSD data from the IsoStar library, see text).

22.4.5.3. *C—H$\cdots$X hydrogen bonds*

An important and often underestimated interaction in biological systems is the C—H$\cdots$*X* hydrogen bond. These bonds have been extensively studied in small-molecule crystal structures, especially in relation to the ongoing discussion as to whether or not they should be called hydrogen bonds. Although Donohue (1968) concluded that the question 'The C—H$\cdots$O hydrogen bond: what is it?' had only one answer: 'It isn't', a survey of 113 neutron-diffraction structures showed clear statistical evidence for an attractive interaction between C—H groups and oxygen and nitrogen acceptors (Taylor & Kennard, 1982). Later, more evidence for this hypothesis was found, and it was even shown that some C—H$\cdots$O interactions are directional (Berkovitch-Yellin & Leiserowitz, 1984; Desiraju, 1991; Steiner & Saenger, 1992; Desiraju *et al.*, 1993; Steiner *et al.*, 1996). A continuing area of interest has been to establish the relative donor abilities of C—H in different chemical environments, since spectroscopic data had indicated that donor ability decreased in the order C(*sp*)—H > C(sp^2)—H > C(sp^3)—H. This general hydrogen-acidity requirement was noted by Taylor & Kennard (1982), and systematically addressed using CSD information by Desiraju & Murty (1987), and by Pedireddi & Desiraju (1992), who derived a novel scale of carbon acidity based on C$\cdots$O separations in a wide variety of systems containing C—H$\cdots$O hydrogen bonds. A recent paper (Derewenda *et al.*, 1995) highlights the importance of C—H$\cdots$O=C bonds in stabilizing protein secondary structure.

22.4.5.4. *O—H$\cdots\pi$ and N—H$\cdots\pi$ hydrogen bonds*

Spectroscopic evidence for the existence of N,O—H$\cdots\pi$ hydrogen bonding to acetylenic, olefinic and aromatic acceptors is well documented (Joris *et al.*, 1968). To our knowledge, the first survey of these interactions in the CSD was carried out by Levitt & Perutz (1988), prompted by observations made in protein structures. A more recent CSD survey of this type of bonding (Viswamitra *et al.*, 1993) has shown that intermolecular examples are clearly observed and that these bonds, although very weak, can be both structurally and energetically significant. Recently, Steiner *et al.* (1995) have presented novel crystal structures, database evidence and quantum-chemical calculations on $C\equiv C—H\cdots\pi(C\equiv C)$ and π(phenyl) bonding. They cite $H\cdots C\equiv C$ (midpoint) distances as short as 2.51 Å and observe hydrogen-bond cooperativity in extended systems with hydrogen-bond energies in the range 4.2–

9.2 kJ mol^{-1}. Finally, we note that electron-rich transition metals can act as proton acceptors in hydrogen-bond interactions with O—H, N—H and C—H donors. Brammer *et al.* (1995) have reviewed progress in this developing area.

22.4.5.5. *Other non-covalent interactions*

The hydrogen bond, $X(\delta-)$—$H(\delta+)\cdots Y(\delta-)$—$Z(\delta+)$, can be viewed as an (almost) linear dipole–dipole interaction, whose ubiquity in nature is due to the presence of many donor–hydrogen dipoles. In a recent review of supramolecular synthons and their application in crystal engineering, Desiraju (1995) illustrates the structural importance of a wide range of attractive non-bonded interactions that do not involve hydrogen mediacy, and notes the long-term value of the CSD in identifying and characterizing these interactions. The area of weak intermolecular interactions is now a burgeoning one in which the combination of CSD analysis and high-level *ab initio* molecular-orbital calculations is proving important in establishing both preferred geometries and estimates of interaction energies. In this context, the intermolecular perturbation theory (IMPT) of Hayes & Stone (1984), a methodology which is free of basis-set superposition errors, is proving particularly useful.

Some of the earliest CSD studies concerned the geometry and directionality of approach of N and O nucleophiles to carbonyl centres, leading to the mapping of (dynamic) reaction pathways through systematic analysis of many examples of related (static) crystal structures (see Bürgi & Dunitz, 1983, 1994). This work was also extended to a study of the directional preferences of non-bonded atomic contacts at sulfur atoms, initially using S in amino acids but later including other examples of divalent sulfur (Rosenfield *et al.*, 1977). It was shown that C—S—C groups tend to bind positively charged electrophiles in directions that are approximately perpendicular to the C—S—C plane, while negatively charged nucleophiles prefer to bind to S along an extension of one of the C—S bonds.

The strong tendency for halogens X = Cl, Br and I to form short contacts to other halogens, and especially to electronegative O and N atoms (Nyburg & Faerman, 1985) is well known (Price *et al.*, 1994). Recent combined CSD/IMPT studies of C—$X\cdots$O=C (Lommerse *et al.*, 1996) and C—$X\cdots$O(nitro) (Allen, Lommerse *et al.*, 1997) systems showed a marked preference for the $X\cdots$O interaction to form along the extension of the C—X bond, with interaction energies in the range -7 to -10 kJ mol^{-1}. These interactions have been used (Desiraju, 1995) to engineer a variety of novel small-molecule crystal structures, and the few $X\cdots$O interactions observed in protein structures generally conform to the geometrical preferences observed in small-molecule studies.

Interactions involving other functional groups are also of importance, and Taylor *et al.* (1990) used CSD information to construct composite crystal-field environments for carbonyl and nitro groups in their search for isosteric replacements in modelling protein–ligand interactions. Their work showed that many of the short intermolecular contacts made by carbonyl groups are to other carbonyl groups in the extended crystal structure. More recently, Maccallum *et al.* (1995*a*,*b*) have demonstrated the importance of Coulombic interactions between the C and O atoms of proximal CONH groups in proteins as an important factor in stabilizing α-helices, β-sheets and the right-hand twist often observed in β-strands. Their calculations indicate an attractive carbonyl–carbonyl interaction energy of about -8 kJ mol^{-1} in specific cases, and they remark that these interactions are *ca* 80% as strong as the CO$\cdots$HN hydrogen bonds within their computational model. Allen, Baalham *et al.* (1998) have used combined CSD/IMPT analysis in a more detailed study of carbonyl–carbonyl interactions and have shown that (*a*) the interaction is commonly observed in small-molecule structures; (*b*) that the preferred interaction geometry is a dimer motif involving two antiparallel C$\cdots$O interactions, although numerous examples of a perpendicular motif (one C$\cdots$O interaction) were also observed; and that (*c*) the total interaction energies for the antiparallel and perpendicular motifs are about -20 and -8 kJ mol^{-1}, respectively, the latter value being comparable to that computed by Maccallum *et al.* (1995*a*,*b*). In studies with protein structures, it has also been noted that carbonyl–carbonyl interactions stabilize the partially allowed Ramachandran conformations of asparagine and aspartic acid (Deane *et al.*, 1999)

22.4.5.6. *Intermolecular motif formation in small-molecule crystal structures*

Desiraju (1995) has stressed that the design process in crystal engineering depends crucially on the high probabilities of formation of certain well known intermolecular motifs, *e.g.* the hydrogen-bonded dimer frequently formed by pairs of carboxylate groups. By analogy with molecular synthesis, he describes these general non-covalent motifs (which often contain strong hydrogen bonds) as supramolecular *synthons*, and points to their importance in supramolecular chemistry as a whole (see *e.g.* Lehn, 1988; Whitesides *et al.*, 1995). Since protein–protein and protein–ligand interactions are also supramolecular phenomena, it follows that information about common interaction motifs is also of importance in structural biology. A computer program is now being written at the CCDC to establish the topologies, chemical constitutions and probabilities of formation of intermolecular motifs directly from the CSD. Initial results (Allen, Raithby *et al.*, 1998; Allen *et al.*, 1999) provide statistics for the most common cyclic hydrogen-bonded motifs, and it is likely that motif information will be included in the developing IsoStar knowledge-based library described in Section 22.4.5.8.

22.4.5.7. *The answer 'no'*

Previous sections have illustrated the location and characterization of some important non-covalent interactions. Equally important is a knowledge of when such interactions *do not* occur although chemical sensibility might indicate that they should. We provide four examples from the CSD: (*a*) only 4.8% of more than 1000 thioether S atoms form hydrogen-atom contacts that are within van der Waals limits, despite the obvious analogy with the potent

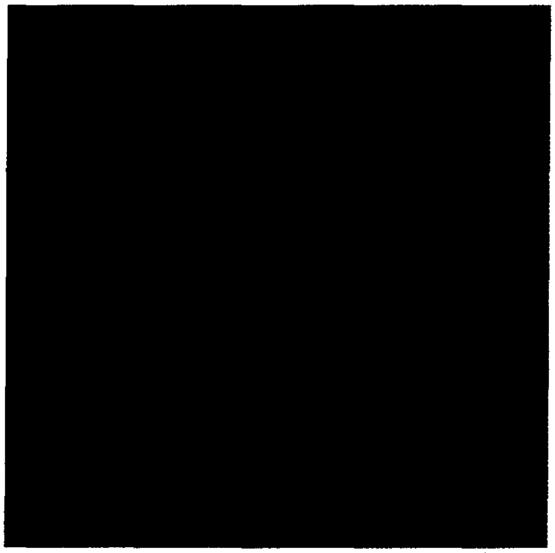

Fig. 22.4.5.3. Distribution of oxygen atoms around C(aromatic)—I (CSD data from the IsoStar library, see text).

acceptor C—O—C (Allen, Bird *et al.*, 1997*b*); (*b*) of 118 instances in which a furan ring coexists with N—H or O—H donors, the O atom forms hydrogen bonds on only three occasions (Nobeli *et al.*, 1997); (*c*) the ester oxygen $(R_1)(\mathrm{O}{=})\mathrm{C}{-}\underline{\mathrm{O}}{-}R_2$ almost never forms strong hydrogen bonds, although the adjunct carbonyl oxygen atom is well known as a highly potent acceptor (Lommerse *et al.*, 1997); and (*d*) covalently bound fluorine atoms rarely form hydrogen bonds (Dunitz & Taylor, 1997).

22.4.5.8. *IsoStar: a library of non-bonded interactions*

The previous sections show that the amount of data in the CSD on intermolecular geometries is vast, and CSD-derived information for a number of specific systems is available in the literature at various levels of detail. If not, the CSD must be searched for contacts between the relevant functional groups. To provide structured and direct access to a more comprehensive set of derived information, a knowledge-based library of non-bonded interactions (IsoStar: Bruno *et al.*, 1997) has been developed at the CCDC since 1995. IsoStar is based on experimental data, not only from the CSD but also from the PDB, and contains some theoretical results calculated using the IMPT method. Version 1.1 of IsoStar, released in October 1998, contains information on non-bonded interactions formed between 310 common functional groups, referred to as *central groups*, and 45 *contact groups*, *e.g.* hydrogen-bond donors, water, halide ions *etc.* Information is displayed in the form of scatter plots for each interaction. Version 1.1 contains about 12000 scatter plots: 9000 from the CSD and 3000 from the PDB. IsoStar also reports results for 867 theoretical potential-energy minima.

For a given contact between between a central group (*A*) and a contact group (*B*), CSD search results were transformed into an easily visualized form by overlaying the *A* moieties. This results in a 3D distribution (scatter plot) showing the experimental distribution of *B* around *A*. Fig. 22.4.5.1(*a*) shows an example of a scatter plot: the distribution of OH groups around carboxylate anions, illustrating hydrogen-bond formation along the lone-pair directions of the carboxylate oxygens. The IsoStar software provides a tool that enables the user to inspect quickly the original crystal structures in which the contacts occur *via* a hyperlink to the original CSD entries. This is very helpful in identifying outliers, motifs and biases. Another tool generates contoured surfaces from scatter plots, which show the density distribution of the contact groups. A similar approach was first used by Rosenfield *et al.* (1984). Contouring aids the interpretation of the scatter plot and the analysis of preferred geometries. Fig. 22.4.5.1(*b*) shows the contoured surface of the scatter plot in Fig. 22.4.5.1(*a*); the lone-pair directionality now becomes even more obvious.

The fact that carboxylate anions form hydrogen bonds along their lone-pair directions may be well known, although force fields do not always use this information. However, the IsoStar library also contains information on many less well understood functional groups. The interaction between aromatic halo groups and oxygen atoms (Lommerse *et al.*, 1996) is referred to above, and Fig. 22.4.5.3 shows the distribution of oxygen acceptor atoms around aromatic iodine groups. It is clear that the contact O atoms are preferentially observed along the elongation of the C—I bond.

The PDB scatter plots in IsoStar only involve interactions between non-covalently bound ligands and proteins, *i.e* side chain–side chain interactions are excluded. Similar work was presented by Tintelnot & Andrews (1989), but at that time the PDB contained only 40 structures of protein–ligand complexes. The IsoStar library contains data derived from almost 800 complexes having a resolution better than 2.5 Å. Fig. 22.4.5.1(*c*) shows an example of a scatter plot from the PDB (the distribution of OH groups around carboxylate groups). Here, although the hydrogen atoms are missing in the PDB plot, the close similarity between Figs. 22.4.5.1(*c*) (PDB) and 22.4.5.1(*a*) (CSD) is obvious.

22.4.5.9. *Protein–ligand binding*

The reluctance to use data from the CSD because they do not relate directly to biological systems has been noted earlier. However, in principle, the same forces that drive the inclusion of a new molecule into a growing crystal should also apply to the binding of a ligand to a protein. In both cases, molecule and target need to be de-solvated first (although in the first case not necessarily from a water environment) and then interact in the most favourable way.

Nicklaus and colleagues suggested that on average, the conformational energy of ligands in the protein-bound state is 66 (48) kJ mol^{-1} above that of the global minimum-energy conformation *in vacuo* (Nicklaus *et al.*, 1995). This result was based on 33 protein–ligand complexes from the PDB for which the ligand also occurs in a small-molecule structure in the CSD. The same investigation also showed that, although ligand conformations in the protein-bound state are generally different from those observed in small-molecule crystal structures, on average the conformational energy of the ligand in the CSD crystal-structure conformation is 66 (47) kJ mol^{-1} above that of the global minimum-energy conformation *in vacuo*, although Boström *et al.* (1998) have shown that these conformational energies are much lower if calculated in a water environment. The computational work indicates that the forces that affect the conformation of a ligand are of comparable magnitude at a protein binding site to those in a small-molecule crystal-structure environment. Thus, if small-molecule crystal-structure statistics tell us that a given structure fragment can only adopt one conformation, generally there is no reason to believe that a ligand that contains this fragment will adopt a different conformation when it binds to a protein.

In principle, the information on non-bonded interactions derived from the CSD and assembled in the IsoStar library should be very important for the understanding and prediction of interaction geometries. However, in light of the comments above, it is important to know whether these data *are* generally relevant to interactions that occur in the protein binding site. Work by Klebe (1994) indicated that, at least for a limited set of test cases, the geometrical distributions derived from ligand–protein complexes are similar to those derived from small-molecule crystal structures. Since the IsoStar library contains information from both the PDB and the CSD, it provides the ultimate basis for establishing similarities (or not) between the interaction geometries observed in small-molecule crystal structures and those observed in protein–ligand complexes. Comparing CSD scatter plots with their corresponding plots from the PDB is an obvious way of establishing the relevance of non-bonded interaction data from small-molecule crystal structures to biological systems.

A full systematic comparison of PDB and CSD scatter plots or, more accurately, of PDB and CSD *density maps* has recently been performed by Verdonk (1998). He calculated residual densities, obtained by subtracting one density map from the other, for each pair of density maps. It appears that, in general, CSD and PDB plots (and thus interaction geometries) are very similar indeed: the average residual density is only 10 (10)%, indicating that 90% of the density in the PDB map is also observed in the CSD map. In Fig. 22.4.5.4(*a*), the average residual densities of each PDB–CSD comparison are plotted *versus* the average concentration of contact groups in the scatter plot. The filled circles represent comparisons for which the protonation state of the central group is unambiguous (*i.e.* carboxylic acid, imidazole *etc.* were excluded). It appears that the residual density decreases with the amount of data in the plots,

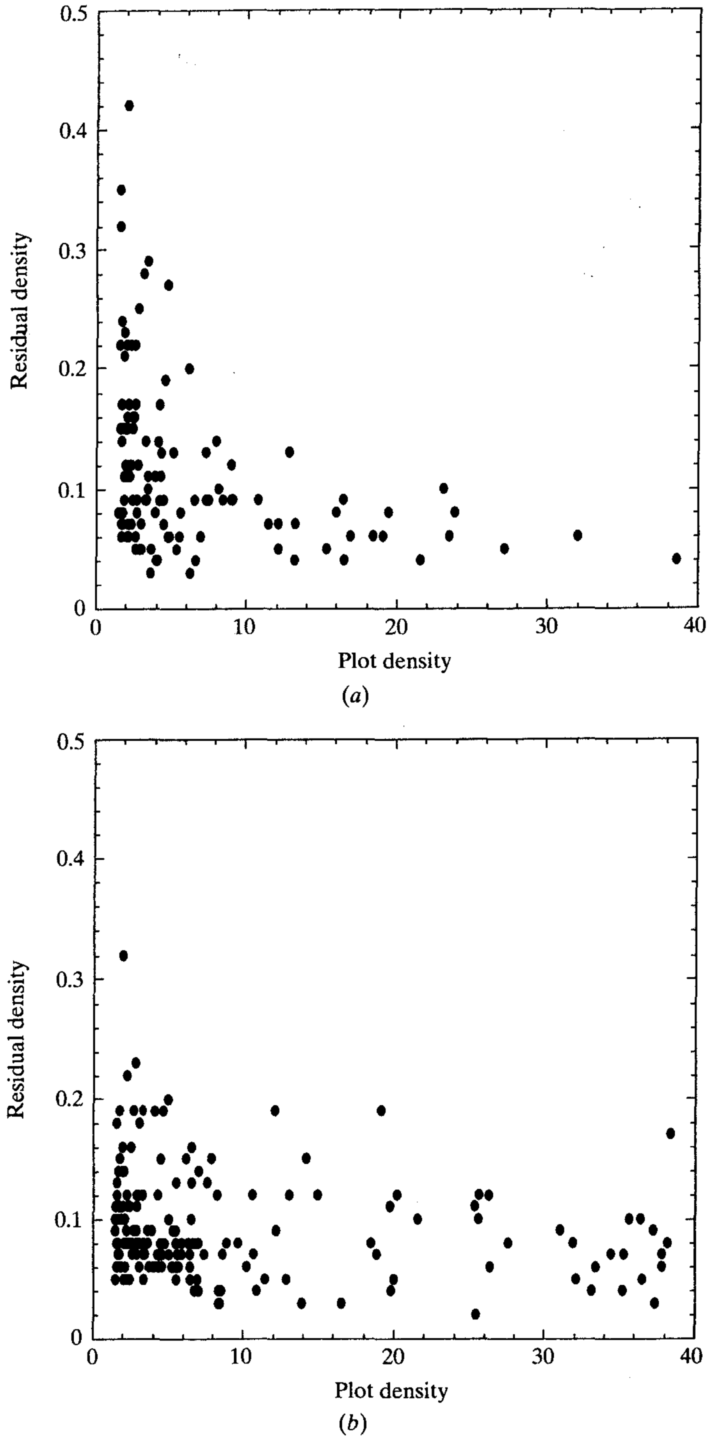

Fig. 22.4.5.4. Pairwise comparison of intermolecular-interaction density maps from the CSD and the PDB. Plots of *residual density* $|\rho(\text{CSD}) - \rho(\text{PDB})|$ *versus plot density*, *i.e.* the average density in the least dense situation (CSD or PDB), for situations where the protonation state of the central group is (*a*) unambiguous, and (*b*) ambiguous.

obviously caused by the more accurate calculation of the residual density. The 'true' residual density seems to be as low as about 6%.

Fig. 22.4.5.4(*b*) shows a similar graph, but now for those density maps in which the protonation state of the central group is ambiguous. As expected, the spread in the calculated residual densities is much higher, even for very dense plots. By comparing the density map from the PDB with the CSD maps for the different protonation states of the central group, the most frequent protonation state of this central group in the protein structures can be predicted. In Table 22.4.5.1, for example, the residual densities for protein carboxylic acid groups are shown, compared with the CSD plots of the neutral carboxylic acid and with those of the charged carboxylate anion. In all cases, the residual density is lower if the PDB map is compared with the CSD map for charged carboxylate anions. This indicates that the majority of glutamate and aspartate side chains are charged, which is consistent with other evidence.

Table 22.4.5.1. *Residual densities for carboxylic acid groups*

The PDB density maps are compared with the CSD maps for uncharged carboxylic acid and for charged carboxylate anions.

	Residual density (CCO_2H)	Residual density ($CCOO^-$)
Any (N,O,S)—H	0.06	0.04
Any N—H nitrogen	0.07	0.05
Any O—H oxygen	0.07	0.05
Non-donating oxygen	0.12	0.04
Carbonyl oxygen	0.13	0.07
Carbonyl carbon	0.12	0.04
Water oxygen	0.07	0.05
Any aliphatic C—H carbon	0.08	0.06

22.4.5.10. *Modelling applications that use CSD data*

Predicting binding modes of ligands at protein binding sites is a problem of paramount importance in drug design. One approach to this problem is to attempt to dock the ligand directly into the binding site. There are several protein–ligand docking programs available, *e.g. DOCK* (see Kuntz *et al.*, 1994), *GRID* (Goodford, 1985), *FLExX* and *FLExS* (Rarey *et al.*, 1996; Lemmen & Lengauer, 1997), and *GOLD* (Jones *et al.*, 1995, 1997). The docking program *GOLD*, developed by the University of Sheffield, Glaxo Wellcome and the CCDC, and which has the high docking success rate of 73%, uses a small torsion library, based on the data from the CSD, to explore the conformational space of the ligand. Its hydrogen-bond geometries and fitness functions are also partly based on CSD data. In the future, we intend to create a more direct link between the crystallographic data and the docking program, *via* IsoStar and the developing torsion library.

Another approach to the prediction of binding modes is to calculate the energy fields for different probes at each position of the binding site, for instance using the *GRID* program (Goodford, 1985). The resulting maps can be displayed as contoured surfaces which can assist in the prediction and understanding of binding modes of ligands. CCDC is developing a program called *SuperStar* (Verdonk *et al.*, 1999) which uses a similar approach to that of the *X-SITE* program (Singh *et al.*, 1991; Laskowski *et al.*, 1996). However, *SuperStar* uses non-bonded interaction data from the CSD rather than the protein side chain–side chain interaction data employed in *X-SITE*. Thus, for a given binding site and contact group (probe), *SuperStar* selects the appropriate scatter plots from the IsoStar library, superimposes the scatter plots on the relevant functional groups in the binding site, and transforms them into one composite probability map. Such maps can then, for example, be used to predict where certain functional groups are likely to interact with the binding site. The strength of *SuperStar* is that it is based entirely on experimental data (although this is also the cause of some limitations). The fields simply represent what has been observed in crystal strucures. We are currently verifying *SuperStar* on a test set of more than 100 protein–ligand complexes from the PDB and preliminary results are encouraging.

Finally, CSD data are used in several *de novo* design programs. These types of programs, *e.g. LUDI* (Böhm, 1992*a*,*b*), predict novel ligands that will interact favourably with a given protein and use hydrogen-bond geometries from the CSD (indirectly) to position their structural fragments in the binding site.

22.4.6. Conclusion

This chapter has summarized the vast range of structural knowledge that can be derived from the small-molecule data contained in the CSD. We have attempted to show that much of this knowledge is directly transferable and applicable to the protein environment. Far from being discrete, structural studies of small molecules and proteins have a natural synergy which, if exploited creatively, will lead to significant advances in both areas. It is therefore unsurprising that some of these CSD studies have been prompted by initial observations made on proteins.

As a result of this activity, it is now very clear that software access to the information stored in the CSD and the PDB must be at two levels: a raw-data level and a derived-knowledge level. The onward development of structural knowledge bases from the underlying data provides for the preservation and storage of the results of data-mining experiments, thus avoiding repetition of standard experiments and providing instant access to complex derivative information. Most importantly, a suitably structured knowledge base can be acted on by software tools that are designed to solve complex problems in structural chemistry (see *e.g.* Thornton & Gardner, 1989; Allen *et al.*, 1990; Bruno *et al.*, 1997; Jones *et al.*, 1997). The availability of knowledge bases derived from experimental observations is likely to be a crucial factor in the solution of those two analogous, and currently intractable, problems in the small-molecule and protein-structure domains: crystal structure and polymorph prediction on the one hand, and protein folding on the other.

References

22.1

Acharya, R., Fry, E., Logan, D., Stuart, D., Brown, F., Fox, G. & Rowlands, D. (1990). *The three-dimensional structure of foot-and-mouth disease virus. New aspects of positive-strand RNA viruses*, edited by M. A. Brinton & S. X. Heinz, pp. 319–327. Washington DC: American Society for Microbiology.

Arnold, E. & Rossmann, M. G. (1990). *Analysis of the structure of a common cold virus, human rhinovirus 14, refined at a resolution of 3.0 Å. J. Mol. Biol.* **211**, 763–801.

Baker, E. N. & Hubbard, R. E. (1984). *Hydrogen bonding in globular proteins. Prog. Biophys. Mol. Biol.* **44**, 97–179.

Bernal, J. D. & Finney, J. L. (1967). *Random close-packed hard-sphere model II. Geometry of random packing of hard spheres. Discuss. Faraday Soc.* **43**, 62–69.

Blake, C. C. F., Koenig, D. F., Mair, G. A., North, A. C. T., Phillips, D. C. & Sarma, V. R. (1965). *Structure of hen egg-white lysozyme, a three-dimensional Fourier synthesis at 2 Å resolution. Nature (London)*, **206**, 757–761.

Bondi, A. (1964). *van der Waals volumes and radii. J. Phys. Chem.* **68**, 441–451.

Bondi, A. (1968). *Molecular crystals, liquids and glasses.* New York: Wiley.

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983). *CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. J. Comput. Chem.* **4**, 187–217.

Chandler, D., Weeks, J. D. & Andersen, H. C. (1983). *van der Waals picture of liquids, solids, and phase transformations. Science*, **220**, 787–794.

Chapman, M. S. (1993). *Mapping the surface properties of macromolecules. Protein Sci.* **2**, 459–469.

Chapman, M. S. (1994). *Sequence similarity scores and the inference of structure/function relationships. Comput. Appl. Biosci. (CABIOS)*, **10**, 111–119.

Chothia, C. (1975). *Structural invariants in protein folding. Nature (London)*, **254**, 304–308.

Chothia, C. (1976). *The nature of the accessible and buried surfaces in proteins. J. Mol. Biol.* **105**, 1–12.

Chothia, C. & Janin, J. (1975). *Principles of protein–protein recognition. Nature (London)*, **256**, 705–708.

Connolly, M. (1986). *Measurement of protein surface shape by solid angles. J. Mol. Graphics*, **4**, 3–6.

Connolly, M. L. (1983). *Analytical molecular surface calculation. J. Appl. Cryst.* **16**, 548–558.

Connolly, M. L. (1991). *Molecular interstitial skeleton. Comput. Chem.* **15**, 37–45.

Diamond, R. (1974). *Real-space refinement of the structure of hen egg-white lysozyme. J. Mol. Biol.* **82**, 371–391.

Dunfield, L. G., Burgess, A. W. & Scheraga, H. A. (1979). *J. Phys. Chem.* **82**, 2609.

Edelsbrunner, H., Facello, M. & Liang, J. (1996). *On the definition and construction of pockets in macromolecules*, pp. 272–287. Singapore: World Scientific.

Edelsbrunner, H., Facello, M., Ping, F. & Jie, L. (1995). *Measuring proteins and voids in proteins. Proc. 28th Hawaii Intl Conf. Sys. Sci.* pp. 256–264.

Edelsbrunner, H. & Mucke, E. (1994). *Three-dimensional alpha shapes. ACM Trans. Graphics*, **13**, 43–72.

Eisenberg, D. & McLachlan, A. D. (1986). *Solvation energy in protein folding and binding. Nature (London)*, **319**, 199–203.

Fauchere, J.-L. & Pliska, V. (1983). *Hydrophobic parameters π of amino-acid side chains from the partitioning of N-acetyl-amino-acid amides. Eur. J. Med. Chem. Chim. Ther.* **18**, 369–375.

Finkelstein, A. (1994). *Implications of the random characteristics of protein sequences for their three-dimensional structure. Curr. Opin. Struct. Biol.* **4**, 422–428.

Finney, J. L. (1975). *Volume occupation, environment and accessibility in proteins. The problem of the protein surface. J. Mol. Biol.* **96**, 721–732.

Finney, J. L., Gellatly, B. J., Golton, I. C. & Goodfellow, J. (1980). *Solvent effects and polar interactions in the structural stability and dynamics of globular proteins. Biophys. J.* **32**, 17–33.

Fritz-Wolf, K., Schnyder, T., Wallimann, T. & Kabsch, W. (1996). *Structure of mitochondrial creatine kinase. Nature (London)*, **381**, 341–345.

Gelin, B. R. & Karplus, M. (1979). *Side-chain torsional potentials: effect of dipeptide, protein, and solvent environment. Biochemistry*, **18**, 1256–1268.

Gellatly, B. J. & Finney, J. L. (1982). *Calculation of protein volumes: an alternative to the Voronoi procedure. J. Mol. Biol.* **161**, 305–322.

Gerstein, M. (1992). *A resolution-sensitive procedure for comparing surfaces and its application to the comparison of antigen-combining sites. Acta Cryst.* A**48**, 271–276.

Gerstein, M. & Chothia, C. (1996). *Packing at the protein–water interface. Proc. Natl Acad. Sci. USA*, **93**, 10167–10172.

Gerstein, M., Lesk, A. M., Baker, E. N., Anderson, B., Norris, G. & Chothia, C. (1993). *Domain closure in lactoferrin: two hinges produce a see-saw motion between alternative close-packed interfaces. J. Mol. Biol.* **234**, 357–372.

Gerstein, M., Lesk, A. M. & Chothia, C. (1994). *Structural mechanisms for domain movements. Biochemistry*, **33**, 6739–6749.

Gerstein, M. & Lynden-Bell, R. M. (1993*a*). *Simulation of water around a model protein helix. 1. Two-dimensional projections of solvent structure. J. Phys. Chem.* **97**, 2982–2991.

22.1 (*cont.*)

Gerstein, M. & Lynden-Bell, R. M. (1993*b*). *Simulation of water around a model protein helix. 2. The relative contributions of packing, hydrophobicity, and hydrogen bonding. J. Phys. Chem.* **97**, 2991–2999.

Gerstein, M. & Lynden-Bell, R. M. (1993*c*). *What is the natural boundary for a protein in solution? J. Mol. Biol.* **230**, 641–650.

Gerstein, M., Sonnhammer, E. & Chothia, C. (1994). *Volume changes on protein evolution. J. Mol. Biol.* **236**, 1067–1078.

Gerstein, M., Tsai, J. & Levitt, M. (1995). *The volume of atoms on the protein surface: calculated from simulation, using Voronoi polyhedra. J. Mol. Biol.* **249**, 955–966.

Grant, J. A. & Pickup, B. T. (1995). *A Gaussian description of molecular shape. J. Phys. Chem.* **99**, 3503–3510.

Greer, J. & Bush, B. L. (1978). *Macromolecular shape and surface maps by solvent exclusion. Proc. Natl Acad. Sci. USA*, **75**, 303–307.

Harber, J., Bernhardt, G., Lu, H.-H., Sgro, J.-Y. & Wimmer, E. (1995). *Canyon rim residues, including antigenic determinants, modulate serotype-specific binding of polioviruses to mutants of the poliovirus receptor. Virology*, **214**, 559–570.

Harpaz, Y., Gerstein, M. & Chothia, C. (1994). *Volume changes on protein folding. Structure*, **2**, 641–649.

Hermann, R. B. (1977). *Use of solvent cavity area and number of packed solvent molecules around a solute in regard to hydrocarbon solubilities and hydrophobic interactions. Proc. Natl Acad. Sci. USA*, **74**, 4144–4195.

Hubbard, S. J. & Argos, P. (1994). *Cavities and packing at protein interfaces. Protein Sci.* **3**, 2194–2206.

Hubbard, S. J. & Argos, P. (1995). *Evidence on close packing and cavities in proteins. Curr. Opin. Biotechnol.* **6**, 375–381.

Kapp, O. H., Moens, L., Vanfleteren, J., Trotman, C. N. A., Suzuki, T. & Vinogradov, S. N. (1995). *Alignment of 700 globin sequences: extent of amino acid substitution and its correlation with variation in volume. Protein Sci.* **4**, 2179–2190.

Kauzmann, W. (1959). *Some factors in the interpretation of protein denaturation. Adv. Protein Chem.* **14**, 1–63.

Kelly, J. A., Sielecki, A. R., Sykes, B. D., James, M. N. & Phillips, D. C. (1979). *X-ray crystallography of the binding of the bacterial cell wall trisaccharaide NAM-NAG-NAM to lysozymes. Nature (London)*, **282**, 875–878.

Kim, K. H., Willingmann, P., Gong, Z. X., Kremer, M. J., Chapman, M. S., Minor, I., Oliveira, M. A., Rossmann, M. G., Andries, K., Diana, G. D., Dutko, F. J., McKinlay, M. A. & Pevear, D. C. (1993). *A comparison of the anti-rhinoviral drug binding pocket in HRV14 and HRV1A. J. Mol. Biol.* **230**, 206–227.

Kleywegt, G. J. & Jones, T. A. (1994). *Detection, delineation, measurement and display of cavities in macromolecular structures. Acta Cryst.* D**50**, 178–185.

Kocher, J. P., Prevost, M., Wodak, S. J. & Lee, B. (1996). *Properties of the protein matrix revealed by the free energy of cavity formation. Structure*, **4**, 1517–1529.

Kraulis, P. J. (1991). *MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. J. Appl. Cryst.* **24**, 946–950.

Kuhn, L. A., Siani, M. A., Pique, M. E., Fisher, C. L., Getzoff, E. D. & Tainer, J. A. (1992). *The interdependence of protein surface topography and bound water molecules revealed by surface accessibility and fractal density measures. J. Mol. Biol.* **228**, 13–22.

Lee, B. & Richards, F. M. (1971). *The interpretation of protein structures: estimation of static accessibility. J. Mol. Biol.* **55**, 379–400.

Leicester, S. E., Finney, J. L. & Bywater, R. P. (1988). *Description of molecular surface shape using Fourier descriptors. J. Mol. Graphics*, **6**, 104–108.

Levitt, M., Hirshberg, M., Sharon, R. & Daggett, V. (1995). *Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution. Comput. Phys. Comm.* **91**, 215–231.

Lewis, M. & Rees, D. C. (1985). *Fractal surfaces of proteins. Science*, **230**, 1163–1165.

Lim, V. I. & Ptitsyn, O. B. (1970). *On the constancy of the hydrophobic nucleus volume in molecules of myoglobins and hemoglobins. Mol. Biol. (USSR)*, **4**, 372–382.

Madan, B. & Lee, B. (1994). *Role of hydrogen bonds in hydrophobicity: the free energy of cavity formation in water models with and without the hydrogen bonds. Biophys. Chem.* **51**, 279–289.

Matthews, B. W., Morton, A. G. & Dahlquist, F. W. (1995). *Use of NMR to detect water within nonpolar protein cavities.* (Letter.) *Science*, **270**, 1847–1849.

Merritt, E. A. & Bacon, D. J. (1997). *Raster3D: photorealistic molecular graphics. Methods Enzymol.* **277**, 505–525.

Molecular Structure Corporation (1995). *Insight II user guide.* Biosym/MSI, San Diego.

Nemethy, G., Pottle, M. S. & Scheraga, H. A. (1983). *Energy parameters in polypeptides. 9. Updating of geometrical parameters, nonbonded interactions and hydrogen bond interactions for the naturally occurring amino acids. J. Phys. Chem.* **87**, 1883–1887.

Nicholls, A. (1992). *GRASP: graphical representation and analysis of surface properties.* New York: Columbia University.

Nicholls, A. & Honig, B. (1991). *A rapid finite difference algorithm, utilizing successive over-relaxation to solve the Poisson–Boltzmann equation. J. Comput. Chem.* **12**, 435–445.

Nicholls, A., Sharp, K. & Honig, B. (1991). *Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. Proteins*, **11**, 281–296.

Olson, N., Kolatkar, P., Oliveira, M. A., Cheng, R. H., Greve, J. M., McClelland, A., Baker, T. S. & Rossmann, M. G. (1993). *Structure of a human rhinovirus complexed with its receptor molecule. Proc. Natl Acad. Sci. USA*, **90**, 507–511.

O'Rourke, J. (1994). *Computational geometry in C.* Cambridge University Press.

Palmenberg, A. C. (1989). *Sequence alignments of picornaviral capsid proteins.* In *Molecular aspects of picornavirus infection and detection*, edited by B. L. Semler & E. Ehrenfeld, pp. 211–241. Washington DC: American Society for Microbiology.

Pattabiraman, N., Ward, K. B. & Fleming, P. J. (1995). *Occluded molecular surface: analysis of protein packing. J. Mol. Recognit.* **8**, 334–344.

Pauling, L. (1960). *The nature of the chemical bond*, 3rd ed. Ithaca: Cornell University Press.

Peters, K. P., Fauck, J. & Frommel, C. (1996). *The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. J. Mol. Biol.* **256**, 201–213.

Petitjean, M. (1994). *On the analytical calculation of van der Waals surfaces and volumes: some numerical aspects. J. Comput. Chem.* **15**, 1–10.

Pontius, J., Richelle, J. & Wodak, S. J. (1996). *Deviations from standard atomic volumes as a quality measure for protein crystal structures. J. Mol. Biol.* **264**, 121–136.

Procacci, P. & Scateni, R. (1992). *A general algorithm for computing Voronoi volumes: application to the hydrated crystal of myoglobin. Int. J. Quant. Chem.* **42**, 151–152.

Rashin, A. A., Iofin, M. & Honig, B. (1986). *Internal cavities and buried waters in globular proteins. Biochemistry*, **25**, 3619–3625.

Reynolds, J. A., Gilbert, D. B. & Tanford, C. (1974). *Empirical correlation between hydrophobic free energy and aqueous cavity surface area. Proc. Natl Acad. Sci. USA*, **71**, 2925–2927.

Richards, F. M. (1974). *The interpretation of protein structures: total volume, group volume distributions and packing density. J. Mol. Biol.* **82**, 1–14.

Richards, F. M. (1977). *Areas, volumes, packing, and protein structure. Annu. Rev. Biophys. Bioeng.* **6**, 151–176.

Richards, F. M. (1979). *Packing defects, cavities, volume fluctuations, and access to the interior of proteins. Including some general comments on surface area and protein structure. Carlsberg Res. Commun.* **44**, 47–63.

22.1 (*cont.*)

Richards, F. M. (1985). *Calculation of molecular volumes and areas for structures of known geometry. Methods Enzymol.* **115**, 440–464.

Richards, F. M. & Lim, W. A. (1994). *An analysis of packing in the protein folding problem. Q. Rev. Biophys.* **26**, 423–498.

Richmond, T. J. (1984). *Solvent accessible surface area and excluded volume in proteins: analytical equations for overlapping spheres and implications for the hydrophobic effect. J. Mol. Biol.* **178**, 63–89.

Richmond, T. J. & Richards, F. M. (1978). *Packing of alpha-helices: geometrical constraints and contact areas. J. Mol. Biol.* **119**, 537–555.

Rossmann, M. G. (1989). *The canyon hypothesis. J. Biol. Chem.* **264**, 14587–14590.

Rossmann, M. G. & Palmenberg, A. C. (1988). *Conservation of the putative receptor attachment site in picornaviruses. Virology*, **164**, 373–382.

Rowland, R. S. & Taylor, R. (1996). *Intermolecular nonbonded contact distances in organic crystal structures: comparison with distances expected from van der Waals radii. J. Phys. Chem.* **100**, 7384–7391.

Sgro, J.-Y. (1996). *Virus visualization.* In *Encyclopedia of virology plus* (CD-ROM version), edited by R. G. Webster & A. Granoff. San Diego: Academic Press.

Sharp, K. A., Nicholls, A., Fine, R. F. & Honig, B. (1991). *Reconciling the magnitude of the microscopic and macroscopic hydrophobic effects. Science*, **252**, 107–109.

Sherry, B., Mosser, A. G., Colonno, R. J. & Rueckert, R. R. (1986). *Use of monoclonal antibodies to identify four neutralization immunogens on a common cold picornavirus, human rhinovirus 14. J. Virol.* **57**, 246–257.

Sherry, B. & Rueckert, R. (1985). *Evidence for at least two dominant neutralization antigens on human rhinovirus 14. J. Virol.* **53**, 137–143.

Shrake, A. & Rupley, J. A. (1973). *Environment and exposure to solvent of protein atoms. Lysozyme and insulin. J. Mol. Biol.* **79**, 351–371.

Sibbald, P. R. & Argos, P. (1990). *Weighting aligned protein or nucleic acid sequences to correct for unequal representation. J. Mol. Biol.* **216**, 813–818.

Singh, R. K., Tropsha, A. & Vaisman, I. I. (1996). *Delaunay tessellation of proteins: four body nearest-neighbor propensities of amino acid residues. J. Comput. Biol.* **3**, 213–222.

Sreenivasan, U. & Axelsen, P. H. (1992). *Buried water in homologous serine proteases. Biochemistry*, **31**, 12785–12791.

Tanford, C. (1997). *How protein chemists learned about the hydrophobicity factor. Protein Sci.* **6**, 1358–1366.

Tanford, C. H. (1979). *Interfacial free energy and the hydrophobic effect. Proc. Natl Acad. Sci. USA*, **76**, 4175–4176.

Ten Eyck, L. F. (1977). *Efficient structure-factor calculation for large molecules by the fast Fourier transform. Acta Cryst.* **A33**, 486–492.

Tsai, J., Gerstein, M. & Levitt, M. (1996). *Keeping the shape but changing the charges: a simulation study of urea and its isosteric analogues. J. Chem. Phys.* **104**, 9417–9430.

Tsai, J., Gerstein, M. & Levitt, M. (1997). *Estimating the size of the minimal hydrophobic core. Protein Sci.* **6**, 2606–2616.

Tsai, J., Taylor, R., Chothia, C. & Gerstein, M. (1999). *The packing density in proteins: standard radii and volumes. J. Mol. Biol.* **290**, 253–266.

Tsai, J., Voss, N. & Gerstein, M. (2001). *Voronoi calculations of protein volumes: sensitivity analysis and parameter database. Bioinformatics*. In the press.

Voronoi, G. F. (1908). *Nouvelles applications des paramétres continus à la théorie des formes quadratiques. J. Reine Angew. Math.* **134**, 198–287.

Williams, M. A., Goodfellow, J. M. & Thornton, J. M. (1994). *Buried waters and internal cavities in monomeric proteins. Protein Sci.* **3**, 1224–1235.

Wodak, S. J. & Janin, J. (1980). *Analytical approximation to the accessible surface areas of proteins. Proc. Natl Acad. Sci. USA*, **77**, 1736–1740.

Xie, Q. & Chapman, M. S. (1996). *Canine parvovirus capsid structure, analyzed at 2.9 Å resolution. J. Mol. Biol.* **264**, 497–520.

Zhou, G., Somasundaram, T., Blanc, E., Parthasarathy, G., Ellington, W. R. & Chapman, M. S. (1998). *Transition state structure of arginine kinase: implications for catalysis of bimolecular reactions. Proc. Natl Acad. Sci. USA*, **95**, 8449–8454.

22.2

Adman, E., Watenpaugh, K. D. & Jensen, L. H. (1975). *N—H···S hydrogen bonds in Peptococcus aerogenes ferredoxin, Clostridium pasteurianum rubredoxin and Chromatium high potential iron protein. Proc. Natl Acad. Sci. USA*, **72**, 4854–4858.

Alber, T., Dao-pin, S., Wilson, K., Wozniak, J. A., Cook, S. P. & Matthews, B. W. (1987). *Contributions of hydrogen bonds of Thr 157 to the thermodynamic stability of phage T4 lysozyme. Nature (London)*, **330**, 41–46.

Artymiuk, P. J. & Blake, C. C. F. (1981). *Refinement of human lysozyme at 1.5 Å resolution. Analysis of non-bonded and hydrogen-bonded interactions. J. Mol. Biol.* **152**, 737–762.

Baker, E. N. (1995). *Solvent interactions with proteins as revealed by X-ray crystallographic studies.* In *Protein–solvent interactions*, edited by R. B. Gregory, pp. 143–189. New York: Marcel Dekker Inc.

Baker, E. N. & Hubbard, R. E. (1984). *Hydrogen bonding in globular proteins. Prog. Biophys. Mol. Biol.* **44**, 97–179.

Blundell, T., Barlow, D., Borkakoti, N. & Thornton, J. (1983). *Solvent-induced distortions and the curvature of α-helices. Nature (London)*, **306**, 281–283.

Bordo, D. & Argos, P. (1994). *The role of side-chain hydrogen bonds in the formation and stabilization of secondary structure in soluble proteins. J. Mol. Biol.* **243**, 504–519.

Burley, S. K. & Petsko, G. A. (1986). *Amino–aromatic interactions in proteins. FEBS Lett.* **203**, 139–143.

Cate, J. H., Gooding, A. R., Podell, E., Zhou, K., Golden, B. L., Kundrot, C. E., Cech, T. R. & Doudna, J. A. (1996). *Crystal structure of a group I ribozyme domain: principles of RNA packing. Science*, **273**, 1678–1685.

Derewenda, Z. S., Derewenda, U. & Kobos, P. (1994). *(His)Cε—H···O=C< hydrogen bond in the active site of serine hydrolases. J. Mol. Biol.* **241**, 83–93.

Derewenda, Z. S., Lee, L. & Derewenda, U. (1995). *The occurrence of C—H···O hydrogen bonds in proteins. J. Mol. Biol.* **252**, 248–262.

Edwards, R. A., Baker, H. M., Whittaker, M. M., Whittaker, J. W. & Baker, E. N. (1998). *Crystal structure of Escherichia coli manganese superoxide dismutase at 2.1 Å resolution. J. Biol. Inorg. Chem.* **3**, 161–171.

Fersht, A. R. & Serrano, L. (1993). *Principles in protein stability derived from protein engineering experiments. Curr. Opin. Struct. Biol.* **3**, 75–83.

Fersht, A. R., Shi, J.-P., Knill-Jones, J., Lowe, D. M., Wilkinson, A. J., Blow, D. M., Brick, P., Carter, P., Waye, M. M. Y. & Winter, G. (1985). *Hydrogen bonding and biological specificity analysed by protein engineering. Nature (London)*, **314**, 235–238.

Flocco, M. M. & Mowbray, S. L. (1995). *Strange bedfellows: interactions between acidic side-chains in proteins. J. Mol. Biol.* **254**, 96–105.

Gregoret, L. M., Rader, S. D., Fletterick, R. J. & Cohen, F. E. (1991). *Hydrogen bonds involving sulfur atoms in proteins. Proteins Struct. Funct. Genet.* **9**, 99–107.

Hagler, A. T., Huler, E. & Lifson, S. (1974). *Energy functions for peptides and proteins. I. Derivation of a consistent force field including the hydrogen bond from amide crystals. J. Am. Chem. Soc.* **96**, 5319–5327.

Harper, E. T. & Rose, G. D. (1993). *Helix stop signals in proteins and peptides: the capping box. Biochemistry*, **32**, 7605–7609.

Huggins, M. L. (1971). *50 years of hydrogen bonding theory. Angew. Chem. Int. Ed. Engl.* **10**, 147–208.

22.2 (*cont.*)

Ippolito, J. A., Alexander, R. S. & Christianson, D. W. (1990). *Hydrogen bond stereochemistry in protein structure and function. J. Mol. Biol.* **215**, 457–471.

Jeffrey, G. A. & Saenger, W. (1991). *Hydrogen bonding in biological structures.* New York: Springer-Verlag.

Kauzmann, W. (1959). *Some factors in the interpretation of protein denaturation. Adv. Protein Chem.* **14**, 1–64.

Legon, A. C. & Millen, D. J. (1987). *Directional character, strength, and nature of the hydrogen bond in gas-phase dimers. Acc. Chem. Res.* **20**, 39–45.

Levitt, M. & Perutz, M. F. (1988). *Aromatic rings act as hydrogen bond acceptors. J. Mol. Biol.* **201**, 751–754.

McDonald, I. K. & Thornton, J. M. (1994*a*). *Satisfying hydrogen bonding potential in proteins. J. Mol. Biol.* **238**, 777–793.

McDonald, I. K. & Thornton, J. M. (1994*b*). *The application of hydrogen bonding analysis in X-ray crystallography to help orientate asparagine, glutamine and histidine side chains. Protein Eng.* **8**, 217–224.

Matthews, B. W. (1972). *The γ turn. Evidence for a new folded conformation in proteins. Macromolecules*, **5**, 818–819.

Mitchell, J. B. O., Nandi, C. L., McDonald, I. K., Thornton, J. M. & Price, S. L. (1994). *Amino/aromatic interactions in proteins: is the evidence stacked against hydrogen bonding? J. Mol. Biol.* **239**, 315–331.

Nemethy, G. & Printz, M. P. (1972). *The γ turn, a possible folded conformation of the polypeptide chain. Comparison with the β turn. Macromolecules*, **5**, 755–758.

Pauling, L. (1960). *The nature of the chemical bond*, 3rd ed. Ithaca: Cornell University Press.

Pauling, L. & Corey, R. B. (1951). *Configurations of polypeptide chains with favoured orientations around single bonds: two new pleated sheets. Proc. Natl Acad. Sci. USA*, **37**, 729–740.

Pauling, L., Corey, R. B. & Branson, H. R. (1951). *The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. Proc. Natl Acad. Sci. USA*, **37**, 205–211.

Pley, H. W., Flaherty, K. M. & McKay, D. B. (1994). *Three-dimensional structure of a hammerhead ribozyme. Nature (London)*, **372**, 68–74.

Presta, L. G. & Rose, G. D. (1988). *Helix signals in proteins. Science*, **240**, 1632–1641.

Richardson, J. S., Getzoff, E. D. & Richardson, D. C. (1978). *The β-bulge: a common small unit of nonrepetitive protein structure. Proc. Natl Acad. Sci. USA*, **75**, 2574–2578.

Richardson, J. S. & Richardson, D. C. (1988). *Amino acid preferences for specific locations at the ends of α-helices. Science*, **240**, 1648–1652.

Savage, H. J., Elliott, C. J., Freeman, C. M. & Finney, J. L. (1993). *Lost hydrogen bonds and buried surface area: rationalising stability in globular proteins. J. Chem. Soc. Faraday Trans.* **89**, 2609–2617.

Schellman, C. (1980). *The alpha-L conformation at the ends of helices.* In *Protein folding*, edited by R. Jaenicke, pp. 53–61. Amsterdam: Elsevier.

Stickle, D. F., Presta, L. G., Dill, K. A. & Rose, G. D. (1992). *Hydrogen bonding in globular proteins. J. Mol. Biol.* **226**, 1143–1159.

Sutor, D. J. (1962). *The C—H$\cdots$O hydrogen bond in crystals. Nature (London)*, **195**, 68–69.

Taylor, R., Kennard, O. & Versichel, W. (1983). *Geometry of the N—H$\cdots$O=C hydrogen bond. 1. Lone pair directionality. J. Am. Chem. Soc.* **105**, 5761–5766.

Thanki, N., Thornton, J. M. & Goodfellow, J. M. (1988). *Distribution of water around amino acids in proteins. J. Mol. Biol.* **202**, 637–657.

Thanki, N., Umrania, Y., Thornton, J. M. & Goodfellow, J. M. (1991). *Analysis of protein main-chain solvation as a function of secondary structure. J. Mol. Biol.* **221**, 669–691.

Wahl, M. C. & Sundaralingam, M. (1997). *C—H$\cdots$O hydrogen bonding in biology. Trends Biochem. Sci.* **22**, 97–102.

Williams, M. A., Goodfellow, J. M. & Thornton, J. M. (1994). *Buried waters and internal cavities in monomeric proteins. Protein Sci.* **3**, 1224–1235.

22.3

Antosiewicz, J., McCammon, J. A. & Gilson, M. K. (1994). *Prediction of pH-dependent properties of proteins. J. Mol. Biol.* **238**, 415–436.

Åqvist, J., Luecke, H., Quiocho, F. A. & Warshel, A. (1991). *Dipoles localized at helix termini of proteins stabilize charges. Proc. Natl Acad. Sci. USA*, **88**, 2026–2030.

Bacquet, R. & Rossky, P. (1984). *Ionic atmosphere of rodlike polyelectrolytes. A hypernetted chain study. J. Phys. Chem.* **88**, 2660.

Bashford, D. & Karplus, M. (1990). *pK$_a$'s of ionizable groups in proteins: atomic detail from a continuum electrostatic model. Biochemistry*, **29**, 10219–10225.

Beroza, P., Fredkin, D., Okamura, M. & Feher, G. (1991). *Protonation of interacting residues in a protein by a Monte-Carlo method. Proc. Natl Acad. Sci. USA*, **88**, 5804–5808.

Bharadwaj, R., Windemuth, A., Sridharan, S., Honig, B. & Nicholls, A. (1994). *The fast multipole boundary-element method for molecular electrostatics – an optimal approach for large systems. J. Comput. Chem.* **16**, 898–913.

Bruccoleri, R. E., Novotny, J., Sharp, K. A. & Davis, M. E. (1996). *Finite difference Poisson–Boltzmann electrostatic calculations: increased accuracy achieved by harmonic dielectric smoothing and charge anti-aliasing. J. Comput. Chem.* **18**, 268–276.

Davis, M. E. (1994). *The inducible multipole solvation model – a new model for solvation effects on solute electrostatics. J. Chem. Phys.* **100**, 5149–5159.

Davis, M. E. & McCammon, J. A. (1989). *Solving the finite difference linear Poisson Boltzmann equation: comparison of relaxational and conjugate gradient methods. J. Comput. Chem.* **10**, 386–395.

Gilson, M. (1993). *Multiple-site titration and molecular modeling: 2. Rapid methods for computing energies and forces for ionizable groups in proteins. Proteins Struct. Funct. Genet.* **15**, 266–282.

Gilson, M., Davis, M., Luty, B. & McCammon, J. (1993). *Computation of electrostatic forces on solvated molecules using the Poisson–Boltzmann equation. J. Phys. Chem.* **97**, 3591–3600.

Gilson, M. & Honig, B. (1986). *The dielectric constant of a folded protein. Biopolymers*, **25**, 2097–2119.

Gilson, M., McCammon, J. & Madura, J. (1995). *Molecular dynamics simulation with continuum solvent. J. Comput. Chem.* **16**, 1081–1095.

Gilson, M., Sharp, K. A. & Honig, B. (1988). *Calculating the electrostatic potential of molecules in solution: method and error assessment. J. Comput. Chem.* **9**, 327–335.

Holst, M., Kozack, R., Saied, F. & Subramaniam, S. (1994). *Protein electrostatics – rapid multigrid-based Newton algorithm for solution of the full nonlinear Poisson–Boltzmann equation. J. Biomol. Struct. Dyn.* **11**, 1437–1445.

Holst, M. & Saied, F. (1993). *Multigrid solution of the Poisson–Boltzmann equation. J. Comput. Chem.* **14**, 105–113.

Jayaram, B., Fine, R., Sharp, K. A. & Honig, B. (1989). *Free energy calculations of ion hydration: an analysis of the Born model in terms of microscopic simulations. J. Phys. Chem.* **93**, 4320–4327.

Jayaram, B., Sharp, K. A. & Honig, B. (1989). *The electrostatic potential of B-DNA. Biopolymers*, **28**, 975–993.

Jean-Charles, A., Nicholls, A., Sharp, K., Honig, B., Tempczyk, A., Hendrickson, T. & Still, C. (1990). *Electrostatic contributions to solvation energies: comparison of free energy perturbation and continuum calculations. J. Am. Chem. Soc.* **113**, 1454–1455.

Langsetmo, K., Fuchs, J. A., Woodward, C. & Sharp, K. A. (1991). *Linkage of thioredoxin stability to titration of ionizable groups with perturbed pK$_a$. Biochemistry*, **30**, 7609–7614.

Lee, F., Chu, Z. & Warshel, A. (1993). *Microscopic and semimicroscopic calculations of electrostatic energies in proteins by the Polaris and Enzymix programs. J. Comput. Chem.* **14**, 161–185.

22.3 (*cont.*)

Lee, F. S., Chu, Z. T., Bolger, M. B. & Warshel, A. (1992). *Calculations of antibody–antigen interactions: microscopic and semi-microscopic evaluation of the free energies of binding of phosphorylcholine analogs to Mcpc603. Protein Eng.* **5**, 215–228.

Misra, V., Hecht, J., Sharp, K., Friedman, R. & Honig, B. (1994). *Salt effects on protein–DNA interactions: the lambda cI repressor and Eco R1 endonuclease. J. Mol. Biol.* **238**, 264–280.

Misra, V. & Honig, B. (1995). *On the magnitude of the electrostatic contribution to ligand–DNA interactions. Proc. Natl Acad. Sci. USA*, **92**, 4691–4695.

Misra, V., Sharp, K., Friedman, R. & Honig, B. (1994). *Salt effects on ligand–DNA binding: minor groove antibiotics. J. Mol. Biol.* **238**, 245–263.

Mohan, V., Davis, M. E., McCammon, J. A. & Pettitt, B. M. (1992). *Continuum model calculations of solvation free energies – accurate evaluation of electrostatic contributions. J. Phys. Chem.* **96**, 6428–6431.

Murthy, C. S., Bacquet, R. J. & Rossky, P. J. (1985). *Ionic distributions near polyelectrolytes. A comparison of theoretical approaches. J. Phys. Chem.* **89**, 701.

Nakamura, H., Sakamoto, T. & Wada, A. (1988). *A theoretical study of the dielectric constant of a protein. Protein Eng.* **2**, 177–183.

Nicholls, A. & Honig, B. (1991). *A rapid finite difference algorithm utilizing successive over-relaxation to solve the Poisson–Boltzmann equation. J. Comput. Chem.* **12**, 435–445.

Oberoi, H. & Allewell, N. (1993). *Multigrid solution of the nonlinear Poisson–Boltzmann equation and calculation of titration curves. Biophys. J.* **65**, 48–55.

Olmsted, M. C., Anderson, C. F. & Record, M. T. (1989). *Monte Carlo description of oligoelectrolyte properties of DNA oligomers. Proc. Natl Acad. Sci. USA*, **86**, 7766–7770.

Olmsted, M. C., Anderson, C. F. & Record, M. T. (1991). *Importance of oligoelectrolyte end effects for the thermodynamics of conformational transitions of nucleic acid oligomers. Biopolymers*, **31**, 1593–1604.

Pack, G., Garrett, G., Wong, L. & Lamm, G. (1993). *The effect of a variable dielectric coefficient and finite ion size on Poisson–Boltzmann calculations of DNA–electrolyte systems. Biophys. J.* **65**, 1363–1370.

Pack, G. R. & Klein, B. J. (1984). *The distribution of electrolyte ions around the B- and Z-conformers of DNA. Biopolymers*, **23**, 2801.

Pack, G. R., Wong, L. & Prasad, C. V. (1986). *Counterion distribution around DNA. Nucleic Acids Res.* **14**, 1479.

Rashin, A. A. (1990). *Hydration phenomena, classical electrostatics and the boundary element method. J. Phys. Chem.* **94**, 725–733.

Record, T., Olmsted, M. & Anderson, C. (1990). *Theoretical studies of the thermodynamics of ion interaction with DNA.* In *Theoretical biochemistry and molecular biophysics.* New York: Adenine Press.

Reiner, E. S. & Radke, C. J. (1990). *Variational approach to the electrostatic free energy in charged colloidal suspensions. J. Chem. Soc. Faraday Trans.* **86**, 3901.

Schaeffer, M. & Frommel, C. (1990). *A precise analytical method for calculating the electrostatic energy of macromolecules in aqueous solution. J. Mol. Biol.* **216**, 1045–1066.

Sharp, K. & Honig, B. (1990). *Calculating total electrostatic energies with the non-linear Poisson–Boltzmann equation. J. Phys. Chem.* **94**, 7684–7692.

Sharp, K. A., Friedman, R., Misra, V., Hecht, J. & Honig, B. (1995). *Salt effects on polyelectrolyte–ligand binding: comparison of Poisson–Boltzmann and limiting law counterion binding models. Biopolymers*, **36**, 245–262.

Simonson, T. & Brooks, C. L. (1996). *Charge screening and the dielectric-constant of proteins: insights from molecular-dynamics. J. Am. Chem. Soc.* **118**, 8452–8458.

Simonson, T. & Brünger, A. (1994). *Solvation free energies estimated from a macroscopic continuum theory. J. Phys. Chem.* **98**, 4683–4694.

Simonson, T. & Perahia, D. (1995). *Internal and interfacial dielectric properties of cytochrome c from molecular dynamics in aqueous solution. Proc. Natl Acad. Sci. USA*, **92**, 1082–1086.

Sitkoff, D., Sharp, K. & Honig, B. (1994). *Accurate calculation of hydration free energies using macroscopic solvent models. J. Phys. Chem.* **98**, 1978–1988.

Slagle, S., Kozack, R. E. & Subramaniam, S. (1994). *Role of electrostatics in antibody–antigen association: anti-hen egg lysozyme/lysozyme complex (HyHEL–5/HEL). J. Biomol. Struct. Dyn.* **12**, 439–456.

Smith, P., Brunne, R., Mark, A. & van Gunsteren, W. (1993). *Dielectric properties of trypsin inhibitor and lysozyme calculated from molecular dynamics simulations. J. Phys. Chem.* **97**, 2009–2014.

Still, C., Tempczyk, A., Hawley, R. & Hendrickson, T. (1990). *Semianalytical treatment of solvation for molecular mechanics and dynamics. J. Am. Chem. Soc.* **112**, 6127–6129.

Takashima, S. & Schwan, H. P. (1965). *Dielectric constant measurements on dried proteins. J. Phys. Chem.* **69**, 4176.

Warshel, A. & Åqvist, J. (1991). *Electrostatic energy and macromolecular function. Annu. Rev. Biophys. Biophys. Chem.* **20**, 267–298.

Warshel, A. & Russell, S. (1984). *Calculations of electrostatic interactions in biological systems and in solutions. Q. Rev. Biophys.* **17**, 283.

Warwicker, J. (1994). *Improved continuum electrostatic modelling in proteins, with comparison to experiment. J. Mol. Biol.* **236**, 887–903.

Warwicker, J. & Watson, H. C. (1982). *Calculation of the electric potential in the active site cleft due to alpha-helix dipoles. J. Mol. Biol.* **157**, 671–679.

Wendoloski, J. J. & Matthew, J. B. (1989). *Molecular dynamics effects on protein electrostatics. Proteins*, **5**, 313.

Yang, A., Gunner, M., Sampogna, R., Sharp, K. & Honig, B. (1993). *On the calculation of pK_a in proteins. Proteins Struct. Funct. Genet.* **15**, 252–265.

Yoon, L. & Lenhoff, A. (1992). *Computation of the electrostatic interaction energy between a protein and a charged suface. J. Phys. Chem.* **96**, 3130–3134.

Zauhar, R. & Morgan, R. J. (1985). *A new method for computing the macromolecular electric potential. J. Mol. Biol.* **186**, 815–820.

Zhou, H. X. (1994). *Macromolecular electrostatic energy within the nonlinear Poisson–Boltzmann equation. J. Phys. Chem.* **100**, 3152–3162.

22.4

Abola, E. E., Sussman, J. L., Prilusky, J. & Manning, N. O. (1997). *Protein Data Bank archives of three-dimensional macromolecular structures. Methods Enzymol.* **277**, 556–571.

Allen, F. H., Baalham, C. A., Lommerse, J. P. M. & Raithby, P. R. (1998). *Carbonyl–carbonyl interactions can be competitive with hydrogen bonds. Acta Cryst.* **B54**, 320–329.

Allen, F. H., Bird, C. M., Rowland, R. S. & Raithby, P. R. (1997*a*). *Resonance-induced hydrogen bonding at sulfur acceptors in $R_1R_2C{=}S$ and $R_1CS_2^-$ systems. Acta Cryst.* **B53**, 680–695.

Allen, F. H., Bird, C. M., Rowland, R. S. & Raithby, P. R. (1997*b*). *Hydrogen-bond acceptor and donor properties of divalent sulfur. Acta Cryst.* **B53**, 696–701.

Allen, F. H., Davies, J. E., Galloy, J. J., Johnson, O., Kennard, O., Macrae, C. F., Mitchell, E. M., Mitchell, G. F., Smith, J. M. & Watson, D. G. (1991). *The development of versions 3 and 4 of the Cambridge Structural Database system. J. Chem. Inf. Comput. Sci.* **31**, 187–204.

Allen, F. H., Doyle, M. J. & Auf der Heyde, T. P. E. (1991). *Automated conformational analysis from crystallographic data. 6. Principal-component analysis for n-membered carbocyclic rings ($n = 4, 5, 6$): symmetry considerations and correlations with ring-puckering parameters. Acta Cryst.* **B47**, 412–424.

Allen, F. H., Doyle, M. J. & Taylor, R. (1991). *Automated conformational analysis from crystallographic data. 3. Three-dimensional pattern recognition within the Cambridge Structural Database system: implementation and practical examples. Acta Cryst.* **B47**, 50–61.

22.4 (*cont.*)

Allen, F. H., Harris, S. E. & Taylor, R. (1996). *Comparison of conformer distributions in the crystalline state with conformational energies calculated by ab initio techniques. J. Comput.-Aided Mol. Des.* **10**, 247–254.

Allen, F. H., Howard, J. A. K. & Pitchford, N. A. (1996). *Symmetry-modified conformational mapping and classification of the medium rings from crystallographic data. IV. Cyclooctane and related eight-membered rings. Acta Cryst.* B**52**, 882–891.

Allen, F. H., Kennard, O., Watson, D. G., Brammer, L., Orpen, A. G. & Taylor, R. (1987). *Tables of bond lengths determined by X-ray and neutron diffraction. Part 1. Bond lengths in organic compounds. J. Chem. Soc. Perkin Trans.* 2, pp. S1–S19.

Allen, F. H., Lommerse, J. P. M., Hoy, V. J., Howard, J. A. K. & Desiraju, G. R. (1997). *Halogen···O(nitro) supramolecular synthon in crystal engineering: a combined crystallographic database and ab initio molecular orbital study. Acta Cryst.* B**53**, 1006–1016.

Allen, F. H., Motherwell, W. D. S., Raithby, P. R., Shields, G. P. & Taylor, R. (1999). *Systematic analysis of the probabilities of formation of bimolecular hydrogen bonded ring motifs in organic crystal structures. New J. Chem.* **23**, 25–34.

Allen, F. H., Raithby, P. R., Shields, G. P. & Taylor, R. (1998). *Probabilities of formation of bimolecular cyclic hydrogen bonded motifs in organic crystal structures: a systematic database study. Chem. Commun.* pp. 1043–1044.

Allen, F. H., Rowland, R. S., Fortier, S. & Glasgow, J. I. (1990). *Knowledge acquisition from crystallographic databases: towards a knowledge-based approach to molecular scene analysis. Tetrahedron Comput. Methodol.* **3**, 757–774.

Ashida, T., Tsunogae, Y., Tanaka, I. & Yamane, T. (1987). *Peptide chain structure parameters, bond angles and conformation angles from the Cambridge Structural Database. Acta Cryst.* B**43**, 212–218.

Balasubramanian, R., Chidambaram, R. & Ramachandran, G. N. (1970). *Potential functions for hydrogen-bond interactions. II. Formulation of an empirical potential function. Biochim. Biophys. Acta*, **221**, 196–206.

Berkovitch-Yellin, Z. & Leiserowitz, L. (1984). *The role played by C—H···O and C—H···N interactions in determining molecular packing and conformation. Acta Cryst.* B**40**, 159–165.

Berman, H. M., Olson, W. K., Beveridge, D. L., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S.-H., Srinivasan, A. R. & Schneider, B. (1992). *The Nucleic Acid Database. A comprehensive relational database of three-dimensional structures of nucleic acids. Biophys. J.* **63**, 751–759.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *The Protein Data Bank. Nucleic Acids Res.* **28**, 235–242.

Bertolasi, V., Gilli, P., Ferretti, V. & Gilli, G. (1996). *Resonance-assisted O—H···O hydrogen bonding. Its role in the crystalline self-recognition of beta-diketone enols and its structural and IR characterisation. Chem. Eur. J.* **2**, 925–934.

Blundell, T. L., Sibanda, B. L., Sternberg, M. J. E. & Thornton, J. M. (1987). *Knowledge-based prediction of protein structures and the design of novel molecules. Nature (London)*, **326**, 347–352.

Böhm, H.-J. (1992*a*). *The computer program LUDI: a new method for the de novo design of enzyme inhibitors. J. Comput.-Aided Mol. Des.* **6**, 61–78.

Böhm, H.-J. (1992*b*). *LUDI: rule-based automatic design of new substituents for enzyme inhibitors. J. Comput.-Aided Mol. Des.* **6**, 593–606.

Bondi, A. (1964). *van der Waals volumes and radii. J. Phys. Chem.* **68**, 441–452.

Boström, J., Norrby, P.-O. & Liljefors, T. (1998). *Conformational energy penalties of protein-bound ligands. J. Comput.-Aided Mol. Des.* **12**, 383–396.

Bower, M., Cohen, F. E. & Dunbrack, R. L. Jr (1997). *Prediction of protein side-chain rotamers from a backbone-dependent rotamer library. J. Mol. Biol.* **267**, 1268–1282.

Brammer, L., Zhao, D., Ladipo, F. T. & Braddock-Wilking, J. (1995). *Hydrogen bonds involving transition metal centres – a brief review. Acta Cryst.* B**51**, 632–640.

Bruno, I. J., Cole, J. C., Lommerse, J. P. M., Rowland, R. S., Taylor, R. & Verdonk, M. L. (1997). *IsoStar: a library of information about nonbonded interactions. J. Comput.-Aided Mol. Des.* **11**, 525–537.

Bürgi, H.-B. & Dunitz, J. D. (1983). *From crystal statics to chemical dynamics. Acc. Chem. Res.* **16**, 153–161.

Bürgi, H.-B. & Dunitz, J. D. (1988). *Can statistical analysis of structural parameters from different crystal environments lead to quantitative energy relationships? Acta Cryst.* B**44**, 445–451.

Bürgi, H.-B. & Dunitz, J. D. (1994). *Structure correlation.* Weinheim: VCH Publishers.

Carbonell, J. (1989). Editor. *Machine learning – paradigms and methods.* Amsterdam: Elsevier.

Carrell, A. B., Shimoni, L., Carrell, C. J., Bock, C. W., Murray-Rust, P. & Glusker, J. P. (1993). *The stereochemistry of the recognition of nitrogen-containing heterocycles by hydrogen bonding and by metal ions. Receptor*, **3**, 57–76.

Carrell, C. J., Carrell, H. L., Erlebacher, J. & Glusker, J. P. (1988). *Structural aspects of metal ion–carboxylate interactions. J. Am. Chem. Soc.* **110**, 8651–8656.

Ceccarelli, C., Jeffrey, G. A. & Taylor, R. (1981). *A survey of O—H···O hydrogen bond geometries determined by neutron diffraction. J. Mol. Struct.* **70**, 255–271.

Chakrabarti, P. (1990*a*). *Interaction of metal ions with carboxylic and carboxamide groups in protein structures. Protein. Eng.* **4**, 49–56.

Chakrabarti, P. (1990*b*). *Geometry of interaction of metal ions with histidine residues in protein structures. Protein Eng.* **4**, 57–63.

Chakrabarti, P. & Dunitz, J. D. (1982). *Directional preferences of ether O-atoms towards alkali and alkaline earth cations. Helv. Chim. Acta*, **65**, 1482–1488.

Chatfield, C. & Collins, A. J. (1980). *Introduction to multivariate analysis.* London: Chapman & Hall.

Conklin, D., Fortier, S., Glasgow, J. I. & Allen, F. H. (1996). *Conformational analysis from crystallographic data using conceptual clustering. Acta Cryst.* B**52**, 535–549.

Cremer, D. & Pople, J. A. (1975). *A general definition of ring puckering coordinates. J. Am. Chem. Soc.* **97**, 1354–1358.

Deane, C. M., Allen, F. H., Taylor, R. & Blundell, T. L. (1999). *Carbonyl–carbonyl interactions stabilise the partially allowed Ramachandran conformations of asparagine and aspartic acid. Protein Eng.* **12**, 1025–1028.

Derewenda, Z. S., Lee, L. & Derewenda, U. (1995). *The occurrence of C—H···O hydrogen bonds in proteins. J. Mol. Biol.* **252**, 248–262.

Desiraju, G. R. (1989). *Crystal engineering: the design of organic solids.* New York: Academic Press.

Desiraju, G. R. (1991). *The C—H···O hydrogen bond in crystals. What is it? Acc. Chem. Res.* **24**, 270–276.

Desiraju, G. R. (1995). *Supramolecular synthons in crystal engineering – a new organic synthesis. Angew. Chem. Int. Ed. Engl.* **34**, 2311–2327.

Desiraju, G. R., Kashino, S., Coombs, M. M. & Glusker, J. P. (1993). *C—H···O packing motifs in some cyclopenta[a]phenanthrenes. Acta Cryst.* B**49**, 880–892.

Desiraju, G. R. & Murty, B. N. (1987). *Correlation between crystallographic and spectroscopic properties for C—H···O bonds in terminal acetylenes. Chem. Phys. Lett.* **139**, 360–361.

Donohue, J. (1952). *The hydrogen bond in organic crystals. J. Phys. Chem.* **56**, 502–510.

Donohue, J. (1968). *Selected topics in hydrogen bonding.* In *Structural chemistry and molecular biology*, edited by W. Davidson & E. Rich, pp. 443–465. San Francisco: W. H. Freeman

Dunbrack, R. L. Jr & Karplus, M. (1993). *Backbone-dependent rotamer library for proteins: applications to side-chain prediction. J. Mol. Biol.* **230**, 534–571.

Dunitz, J. D. & Taylor, R. (1997). *Organic fluorine hardly ever accepts hydrogen bonds. Chem. Eur. J.* **3**, 83–90.

22.4 (*cont.*)

Einspahr, H. & Bugg, C. E. (1981). *The geometry of calcium–carboxylate interactions in crystalline complexes. Acta Cryst.* B**37**, 1044–1052.

Engh, R. A. & Huber, R. (1991). *Accurate bond and angle parameters for X-ray protein structure refinement. Acta Cryst.* A**47**, 392–400.

Everitt, B. (1980). *Cluster analysis.* New York: Wiley.

Fortier, S., Castleden, I., Glasgow, J., Conklin, D., Walmsley, C., Leherte, L. & Allen, F. H. (1993). *Molecular scene analysis: the integration of direct-methods and artificial-intelligence strategies for solving protein crystal structures. Acta Cryst.* D**49**, 168–178.

Glusker, J. P. (1980). *Citrate conformation and chelation: enzymatic implications. Acc. Chem. Res.* **13**, 345–352.

Glusker, J. P., Lewis, M. & Rossi, M. (1994). *Crystal structure analysis for chemists and biologists.* Weinheim: VCH Publishers.

Goodford, P. J. (1985). *A computational procedure for determining energetically favourable binding sites on biologically important molecules. J. Med. Chem.* **28**, 849–857.

Gould, R. O., Gray, A. M., Taylor, P. & Walkinshaw, M. D. (1985). *Crystal environments and geometries of leucine, isoleucine, valine and phenylalanine provide estimates of minimum nonbonded contact and preferred van der Waals interaction distances. J. Am. Chem. Soc.* **107**, 5921–5927.

Hall, S. R., Allen, F. H. & Brown, I. D. (1991). *The crystallographic information file (CIF): a new standard archive file for crystallography. Acta Cryst.* A**47**, 655–685.

Hayes, I. C. & Stone, A. J. (1984). *An intermolecular perturbation theory for the region of moderate overlap. J. Mol. Phys.* **53**, 83–105.

Hooft, R. W. W., Vriend, G., Sander, C. & Abola, E. E. (1996). *Errors in protein structures. Nature (London)*, **381**, 272.

Jeffrey, G. A. & Maluszynska, H. (1982). *A survey of hydrogen bond geometries in the crystal structures of amino acids. Int. J. Biol. Macromol.* **4**, 173–185.

Jeffrey, G. A. & Maluszynska, H. (1986). *A survey of the geometry of hydrogen bonds in the crystal structures of barbiturates, purines and pyrimidines. J. Mol. Struct.* **147**, 127–142.

Jeffrey, G. A. & Mitra, J. (1984). *Three-centre (bifurcated) hydrogen bonding in the crystal structures of amino acids. J. Am. Chem. Soc.* **106**, 5546–5553.

Jeffrey, G. A. & Saenger, W. (1991). *Hydrogen bonding in biological structures.* Berlin: Springer Verlag.

Jones, G., Willett, P. & Glen, R. C. (1995). *Molecular recognition of receptor sites using a genetic algorithm with a description of solvation. J. Mol. Biol.* **245**, 43–53.

Jones, G., Willett, P., Glen, R. C., Leach, A. R. & Taylor, R. (1997). *Development and validation of a genetic algorithm for flexible docking. J. Mol. Biol.* **267**, 727–748.

Joris, L., Schleyer, P. & Gleiter, R. (1968). *Cyclopropane rings as proton acceptor groups in hydrogen bonding. J. Am. Chem. Soc.* **90**, 327–336.

Kennard, O. (1962). In *International tables for X-ray crystallography*, Vol. II. Birmingham: Kynoch Press.

Kennard, O. & Allen, F. H. (1993). *The Cambridge Crystallographic Data Centre. Chem. Des. Autom. News*, **8**, 1, 31–37.

Klebe, G. (1994). *The use of composite crystal-field environments in molecular recognition and the de novo design of protein ligands. J. Mol. Biol.* **237**, 212–235.

Klebe, G. & Mietzner, T. (1994). *A fast and efficient method to generate biologically relevant conformations. J. Comput.-Aided Mol. Des.* **8**, 583–594.

Kroon, J. & Kanters, J. A. (1974). *Non-linearity of hydrogen bonds in molecular crystals. Nature (London)*, **248**, 667–669.

Kroon, J., Kanters, J. A., van Duijneveldt-van de Rijdt, J. G. C. M., van Duijneveldt, F. B. & Vliegenthart, J. A. (1975). *O—H···O hydrogen bonds in molecular crystals: a statistical and quantum chemical analysis. J. Mol. Struct.* **24**, 109–129.

Kuntz, I. D., Meng, E. C. & Stoichet, B. K. (1994). *Structure-based molecular design. Acc. Chem. Res.* **27**, 117–122.

Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *PROCHECK: a program to check the stereochemical quality of protein structures. J. Appl. Cryst.* **26**, 283–291.

Laskowski, R. A., Thornton, J. M., Humblet, C. & Singh, J. (1996). *X-SITE: use of empirically derived atomic packing preferences to identify favourable interaction regions in the binding sites of proteins. J. Mol. Biol.* **259**, 175–201.

Lehn, J.-M. (1988). *Perspectives in supramolecular chemistry – from molecular recognition towards molecular information processing and self-organization. Angew. Chem. Int. Ed. Engl.* **27**, 90–112.

Lemmen, C. & Lengauer, T. (1997). *Time-efficient flexible superposition of medium sized molecules. J. Comput.-Aided Mol. Des.* **11**, 357–368.

Levitt, M. & Perutz, M. (1988). *Aromatic rings act as hydrogen bond acceptors. J. Mol. Biol.* **201**, 751–754.

Lommerse, J. P. M., Price, S. L. & Taylor, R. (1997). *Hydrogen bonding of carbonyl, ether and ester oxygen atoms with alkanol hydroxyl groups. J. Comput. Chem.* **18**, 757–780.

Lommerse, J. P. M., Stone, A. J., Taylor, R. & Allen, F. H. (1996). *The nature and geometry of intermolecular interactions between halogens and oxygen or nitrogen. J. Am. Chem. Soc.* **118**, 3108–3116.

Maccallum, P. H., Poet, R. & Milner-White, E. J. (1995*a*). *Coulombic interactions between partially charged main-chain atoms not hydrogen bonded to each other influence the conformations of α-helices and antiparallel β-sheets. J. Mol. Biol.* **248**, 361–373.

Maccallum, P. H., Poet, R. & Milner-White, E. J. (1995*b*). *Coulombic interactions between partially charged main-chain atoms stabilise the right-handed twist found in most β-strands. J. Mol. Biol.* **248**, 374–384.

Miller, J., McLachlan, A. D. & Klug, A. (1985). *Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes. EMBO J.* **4**, 1609–1614.

Murray-Rust, P. & Bland, R. (1978). *Computer retrieval and analysis of molecular geometry. II. Variance and its interpretation. Acta Cryst.* B**34**, 2527–2533.

Murray-Rust, P. & Glusker, J. P. (1984). *Directional hydrogen bonding to sp^2 and sp^3 hybridized oxygen atoms and its relevance to ligand–macromolecule interactions. J. Am. Chem. Soc.* **105**, 1018–1025.

Murray-Rust, P. & Motherwell, S. (1978). *Computer retrieval and analysis of molecular geometry. III. Geometry of the β-1$'$-aminofuranoside fragment. Acta Cryst.* B**34**, 2534–2546.

Nicklaus, M. C., Wang, S., Driscoll, J. S. & Milne, G. W. A. (1995). *Conformational changes of small molecules binding to proteins. Bioorg. Med. Chem.* **3**, 411–428.

Nobeli, I., Price, S. L., Lommerse, J. P. M. & Taylor, R. (1997). *Hydrogen bonding properties of oxygen and nitrogen acceptors in aromatic heterocycles. J. Comput. Chem.* **18**, 2060–2074.

Nyburg, S. C. & Faerman, C. H. (1985). *A revision of van der Waals atomic radii for molecular crystals: N, O, F, S, Cl, Se, Br and I bonded to carbon. Acta Cryst.* B**41**, 274–279.

Orpen, A. G., Brammer, L., Allen, F. H., Kennard, O., Watson, D. G. & Taylor, R. (1989). *Tables of bond lengths determined by X-ray and neutron diffraction. Part II: organometallic compounds and coordination complexes of the d- and f-block metals. J. Chem. Soc. Dalton Trans.* pp. S1–S83.

Pauling, L. (1939). *The nature of the chemical bond.* Ithaca: Cornell University Press.

Pedireddi, V. R. & Desiraju, G. R. (1992). *A crystallographic scale of carbon acidity. Chem. Commun.* pp. 988–990.

Pimentel, G. C. & McClellan, A. L. (1960). *The hydrogen bond.* San Francisco: W. H. Freeman.

Price, S. L., Stone, A. J., Lucas, J., Rowland, R. S. & Thornley, A. E. (1994). *The nature of —Cl···Cl— intermolecular interactions. J. Am. Chem. Soc.* **116**, 4910–4918.

Rappoport, Z., Biali, S. E. & Kaftory, M. (1990). *Application of the structure correlation method to ring-flip processes in benzophenone. J. Am. Chem. Soc.* **112**, 7742–7750.

Rarey, M., Kramer, B., Lengauer, T. & Klebe, G. (1996). *Predicting receptor-ligand interactions by an incremental construction algorithm. J. Mol. Biol.* **261**, 470–481.

22.4 (*cont.*)

Robertson, J. M. (1953). *Organic crystals and molecules*. Ithaca: Cornell University Press.

Rosenfield, R. E., Parthasarathy, R. & Dunitz, J. D. (1977). *Directional preferences of non-bonded atomic contacts with divalent sulphur. 1. Electrophiles and nucleophiles. J. Am. Chem. Soc.* **99**, 4860–4862.

Rosenfield, R. E. Jr, Swanson, S. M., Meyer, E. F. Jr, Carrell, H. L. & Murray-Rust, P. (1984). *Mapping the atomic environment of functional groups: turning 3D scatterplots into pseudo-density contours. J. Mol. Graphics*, **2**, 43–46.

Rowland, R. S. & Taylor, R. (1996). *Intermolecular nonbonded contact distances in organic crystal structures: comparison with distances expected from van der Waals radii. J. Phys. Chem.* **100**, 7384–7391.

Schweizer, W. B. & Dunitz, J. D. (1982). *Structural characteristics of the carboxylic acid ester group. Helv. Chim. Acta*, **65**, 1547–1552.

Singh, J., Saldanha, J. & Thornton, J. M. (1991). *A novel method for the modelling of peptide ligands to their receptors. Protein Eng.* **4**, 251–261.

Steiner, T., Kanters, J. A. & Kroon, J. (1996). *Acceptor directionality of sterically unhindered C—H···O=C hydrogen bonds donated by acidic C—H groups. J. Chem. Soc. Chem. Commun.* **11**, 1277–1278.

Steiner, T. & Saenger, W. (1992). *Geometry of C—H···O hydrogen bonds in carbohydrate crystal structures. Analysis of neutron diffraction data. J. Am. Chem. Soc.* **114**, 10146–10154.

Steiner, T., Starikov, E. B., Amado, A. M. & Teixeira-Dias, J. J. C. (1995). *Weak hydrogen bonding. Part 2. The hydrogen-bonding nature of short C—H··· • contacts. Crystallographic, spectroscopic and quantum mechanistic studies of some terminal alkynes. J. Chem. Soc. Perkin Trans.* 2, pp. 1312–1326.

Strynadka, N. C. J. & James, M. N. G. (1989). *Crystal structures of the helix-loop-helix calcium-binding proteins. Annu. Rev. Biochem.* **58**, 951–998.

Sutton, L. E. (1956). *Tables of interatomic distances and configuration in molecules and ions*. Special Publication No. 11. London: The Chemical Society.

Sutton, L. E. (1959). *Tables of interatomic distances and configuration in molecules and ions (supplement)*. Special Publication No. 18. London: The Chemical Society.

Taylor, R. (1986). *The Cambridge Structural Database in molecular graphics: techniques for the rapid identification of conformational minima. J. Mol. Graphics*, **4**, 123–131.

Taylor, R. & Allen, F. H. (1994). *Statistical and numerical methods of data analysis*. In *Structure correlation*, edited by H.-B. Bürgi & J. D. Dunitz. Weinheim: VCH Publishers.

Taylor, R. & Kennard, O. (1982). *Crystallographic evidence for the existence of C—H···O, C—H···N and C—H···Cl hydrogen bonds. J. Am. Chem. Soc.* **104**, 5063–5070.

Taylor, R. & Kennard, O. (1983). *Comparison of X-ray and neutron diffraction results for the N—H···O=C hydrogen bond. Acta Cryst.* B**39**, 133–138.

Taylor, R., Kennard, O. & Versichel, W. (1983). *Geometry of the N—H···O=C hydrogen bond. 1. Lone-pair directionality. J. Am. Chem. Soc.* **105**, 5761–5766.

Taylor, R., Kennard, O. & Versichel, W. (1984*a*). *Geometry of the N—H···O=C hydrogen bond. 2. Three-centre (bifurcated) and four-centre (trifurcated) bonds. J. Am. Chem. Soc.* **106**, 244–248.

Taylor, R., Kennard, O. & Versichel, W. (1984*b*). *Geometry of the N—H···O=C hydrogen bond. 3. Hydrogen-bond distances and angles. Acta Cryst.* B**40**, 280–288.

Taylor, R., Mullaley, A. & Mullier, G. W. (1990). *Use of crystallographic data in searching for isosteric replacements: composite field environments of nitro and carbonyl groups. Pestic. Sci.* **29**, 197–213.

Thornton, J. M. & Gardner, S. P. (1989). *Protein motifs and database searching. Trends Biochem. Sci.* **14**, 300–304.

Tintelnot, M. & Andrews, P. (1989). *Geometries of functional group interactions in enzyme–ligand complexes: guides for receptor modelling. J. Comput.-Aided Mol. Des.* **3**, 67–84.

Verdonk, M. L. (1998). Unpublished results.

Verdonk, M. L., Cole, J. C. & Taylor, R. (1999). *SuperStar: a knowledge-based approach for identifying interaction sites in proteins. J. Mol. Biol.* **289**, 1093–1108.

Viswamitra, M. A., Radhakrishnan, R., Bandekar, J. & Desiraju, G. R. (1993). *Evidence for O—H···C and N—H···C hydrogen bonding. J. Am. Chem. Soc.* **115**, 4868–4869.

Whitesides, G. M., Simanek, E. E., Mathias, J. P., Seto, C. T., Chin, D. N., Mammen, M. & Gordon, D. M. (1995). *Non-covalent synthesis: using physical-organic chemistry to make aggregates. Acc. Chem. Res.* **28**, 37–43.