A Set of van der Waals and Coulombic Radii of Protein Atoms for Molecular and Solvent-Accessible Surface Calculation, Packing Evaluation, and Docking

Ai-Jun Li[1] **and Ruth Nussinov**[2,3]*
[1]*Laboratory of Experimental and Computational Biology, NCI-FCRDC, Frederick, Maryland*
[2]*Laboratory of Experimental and Computational Biology, IRSP, SAIC, NCI-FCRDC, Frederick, Maryland*
[3]*Sackler Institute of Molecular Medicine, School of Medicine, Tel Aviv University, Tel Aviv, Israel*

ABSTRACT **We analyze the contact distance distributions between nonbonded atoms in known protein structures. A complete set of van der Waals (VDW) radii for 24 protein atom types and for crystal-bound water is derived from the contact distance distributions of these atoms with a selected group of apolar atoms. In addition, a set of Coulombic radii for polar atoms is derived from their contacts with water. The contact distance distributions and the two sets of radii are derived in a systematic and self-consistent manner using an iterative procedure. The Coulombic radii for polar atoms are, on average, 0.18 Å smaller than their VDW radii. The VDW radius of water is 1.7 Å, which is 0.3 Å larger than its Coulombic radius. We show that both the VDW and the Coulombic radii of polar atoms are needed in calculating the molecular and solvent-accessible surfaces of proteins. The VDW radii are needed to generate the apolar portions of the surface and the Coulombic radii for the polar portions. The fact that polar atoms have two apparent sizes implies that a hydrophobic cavity has to be larger than a polar cavity in order to accommodate the same number of water molecules. Most surface area calculations have used only one radius for each polar atom. As a result, unreal cavities, grooves, or pockets may be generated if the Coulombic radii of polar atoms are used. On the other hand, if the VDW radii of polar atoms are used, the details of the polar regions of the surface may be lost. The accuracy of the molecular and the solvent-accessible surfaces of proteins can be improved if the radii of polar atoms are allowed to change depending on the nature of their contacting neighbors. The surface of a protein at a protein–protein interface differs from that in solution in that it has to be generated using at least two kinds of probes, one representing a typical apolar atom and the other a typical polar atom. This observation has important implications for docking, which relies on surface complementarity at the interface. Proteins 32:111–127, 1998.** © **1998 Wiley-Liss, Inc.**

Key words: **van der Waals radius; Coulombic radius; docking; molecular surface; solvent-accessible surface; protein–protein interface; protein–water boundary**

INTRODUCTION

To give a quantitative measure of the size of atoms and molecules, Pauling[1] defined a quantity called the van der Waals (VDW) radius. The VDW radii are determined in terms of the contact distances between pairs of atoms that interact only through the VDW interactions. That is, there are no net charges and covalent bonds between the pairs of atoms. Atom pairs that form hydrogen bonds are also excluded. The size and shape of a molecule is related to its electronic structure, i.e., its electron density distribution,[1,2] which is, in turn, related to the electronic structure of the constituent atoms. For convenience, it is usually assumed that atoms are all spherical and each can be assigned a radius. Thus, the shape of a molecule can be represented by a set of intersecting spheres.

In proteins, hydrogen bonds and charge interactions are numerous. When the VDW radii are used to describe atoms that form favorable polar interactions, these radii overlap with each other. This is expected because favorable polar interactions cause tighter packing. In many applications, however, we

Grant sponsor: National Cancer Institute, DHHS; Grant number: 1-CO-74102; Grant sponsor: Israel-U.S. Binational Science Foundation; Grant number: 95–00208; Grant sponsor: Rekanati Foundation; Grant sponsor: Israeli Ministry of Industry and Commerce (Maget); Grant sponsor: Tel Aviv University.

The opinions expressed in this article are those of the authors and do not reflect the views or policies of the DHHS or the U.S. Government. The mention of trade names, commercial products, or organizations does not imply endorsement by the U.S. Government.

*Correspondence to: Ruth Nussinov, Laboratory of Experimental and Computational Biology, IRSP, SAIC, NCI-FCRDC, Building 469, Room 151, Frederick, MD 21701. E-mail: ruthn@ncifcrf.gov

Received 10 October 1997; Accepted 3 February 1998

would like to know the most probable separation between atoms that form hydrogen bonds and salt bridges, in addition to the tightness of packing. Thus, it is convenient to define another radius for polar atoms that reflects their apparent sizes when they are involved in favorable polar interactions. In recognition of the Coulombic nature of such interactions, we refer to this radius as the Coulombic radius.

It turns out that defining two radii for each polar atom is not just a convenience—it is a necessity. In particular, two sets of radii are needed for polar atoms in generating the molecular and solvent-accessible surfaces of proteins. In an analysis of buried waters and internal cavities, Williams et al.[3] have noted the need for two radii for polar atoms. They defined, for each polar atom, a characteristic polar radius and a characteristic apolar radius. The need for two or more radii for polar atoms has also been pointed out by Gerstein et al.,[4] Gerstein and Chothia,[5] and Kocher et al.[6] Based on a molecular dynamics simulation study of the volumes of protein surface atoms, Gerstein et al.[4] have suggested that three different radii be used for water corresponding to interactions with apolar, polar, and highly charged atoms. In an analysis of the packing of crystal-bound waters against protein atoms in 22 high-resolution crystal structures, Gerstein and Chothia[5] have tested the effect of different VDW radii sets on the volume of protein surface atoms and have discovered that, by increasing the radius of oxygen (including water) from 1.4 Å to 1.6 Å, the observed volume increase of surface atoms decreased from 2.5% to 0.6%. Kocher et al.[6] increased the Coulombic radius of oxygen atoms in recognition of the increase in size of this atom type when it interacts with apolar neighbors.

To understand why two sets of radii are needed in calculating the molecular and solvent-accessible surfaces, it is necessary to examine the definitions of these quantities. The solvent-accessible surface of a molecule is defined by Lee and Richards[7] as the envelope traced out by the center of a solvent probe when it is rolled over the molecule. Let us consider that the molecule is a protein and the solvent is water, which is polar. In addition, we view water as a spherical united atom. When a water is in contact with an apolar atom on the protein, the location of its center is determined by the VDW radii of the two atoms. When a water is in contact with a polar atom, however, the most probable separation is now given by the sum of their Coulombic radii. When a water is in contact with both polar and apolar atoms, the two radii of water are needed simultaneously to specify its center location. Thus, to calculate the solvent-accessible surface of a protein, both the VDW and Coulombic radii of water are needed. As for atoms in the protein, the VDW radii are needed for apolar atoms and the Coulombic radii are needed for polar atoms. Williams et al.[3] are probably the first to use two radii for water to generate the solvent-accessible surfaces of proteins.

The molecular surface of a molecule is defined by Richards[8] as the envelope traced out by the contacting sphere of the probing molecule. The sphere of the probe is not allowed to overlap the spheres on the target at all times. Using the protein–water example again, when water probes a group of polar atoms on the protein, it is clear that the only way to satisfy this no-overlap requirement is to use the Coulombic radii to represent the size of these polar atoms. On the other hand, when water is probing an apolar portion of the proteins surface, its size and the sizes of the protein atoms are all determined by the VDW radii. Thus, the choice of radii is exactly the same as that for generating the solvent-accessible surfaces. That is, the VDW radii are to be used for apolar atoms in the protein and for water when it is in contact with apolar atoms, and the Coulombic radii are to be used for polar atoms in the protein and for water when it is in contact with polar atoms. At the polar–apolar boundary, the shape of the probe is no longer spherical, which complicates the surface calculation. A practical solution to this problem is suggested later.

In the above, we did not distinguish the polarity of the polar atoms on the protein because water can interact favorably with both positive and negative polarity atoms. If the probe has a specified polarity, then the polarity of polar atoms in the protein has to be distinguished. It is relatively straightforward to determine the appropriate radii to use for other types of probes, such as apolar, hydrogen bond donor or acceptor. The important thing to realize is that the molecular and solvent-accessible surfaces are not properties of a molecule itself. These surfaces are defined together with the probing atom.[7,8] That is, the surface of a molecule may change when different probes are used.

The dependence of a protein's surface on the nature of the probe complicates the calculation of its molecular surface at a protein–protein interface, since the probe is not a single atom, but another protein that has both polar and apolar atoms. In principle, the molecular surface at the interface, the interfacial molecular surface, is undefined before the complex is formed, as there is no a priori knowledge of which probe to choose for a particular region. In practice, however, polar atoms are almost always paired with polar neighbors and apolar atoms are almost always paired with apolar neighbors. Thus, from a practical point of view, one could use two kinds of probes to generate the interfacial molecular surface, one representing a typical polar atom to probe the polar regions and the other representing a typical apolar atom to probe the apolar regions. As far as calculations are concerned, one can still think that there is one probe, but the VDW and the Coulombic radius of the probe differ from water.

In the past, most surface area calculations have used only one radius for each polar atom. If the smaller Coulombic radius of water (1.4 Å) is used to generate the molecular and solvent-accessible surfaces of proteins, apolar portions of the surface may contain extra and unreal cavities, grooves, or pockets. On the other hand, if the VDW radius of water is used, the real fine structure of some polar regions may be lost and regions that are accessible to water may appear forbidden. The same argument also applies to the calculation of protein–water boundary, as it is often chosen as the molecular surface of the protein.[9] Considering the important role played by the molecular and solvent-accessible areas in protein folding,[10–12] stability,[13,14] and protein–protein recognition,[15–18] it is very important that they are calculated accurately.

It should be pointed out that only the VDW radii are needed for the evaluation of protein volumes, interior packing, and packing at the protein–water interface, as these are quantities related to the electron density distribution[19] of the molecules.

The VDW radius of each element varies slightly with its covalent bonding environment. In addition, since hydrogen atoms are usually not observed in crystal structures, heavy atoms and their covalently bonded hydrogen atoms are often considered united atoms.[20] Thus, it is often necessary to divide each element into several types according to its covalent bonding environment. There is no consensus in the literature regarding the number of atom types and the radius of each type. The first calculation of the solvent-accessible surface of proteins by Lee and Richards[7] has used five atom types whose VDW radii are based on a compilation by Bondi[21] of contact distances in crystals of small molecules. Richards[22] has defined two types of carbons, three types each of oxygens and nitrogens, and one type of sulfur in the calculation of protein internal packing. Chothia[13] has used a set of VDW radii (six atom types) derived from crystal structures of amino acids. Williams et al.[3] defined five atom types, including water. Other often quoted values of VDW radii are given by Pauling[1] and Gavezzotti,[23] which are also based on contact distances in crystals of small molecules.

Unlike small molecules (including individual amino acids), a protein molecule has a unique three-dimensional structure. The covalent link of amino acids affects the environment of each atom and it is difficult to find a one-to-one correspondence in small molecules for every atom in proteins. In addition, the arrangement of atoms in a protein is strongly affected by its secondary and tertiary structure, while in crystals of small molecules the arrangement is periodic. Thus, it has to be verified whether values derived from crystals of small molecules apply to proteins. Historically, values derived from small molecules were used because the number of protein structures was too small to allow an accurate determination of radii from atom contacts in proteins. Such a task is now possible owing to the availability of a large number of protein crystal structures.

Here, we investigate the contact distance distributions between atoms in proteins and between protein atoms and water, using a representative list of 1,405 protein structures given by Hobohm and Sander.[24] A complete set of VDW radii for 24 protein atom types and crystal-bound water are derived from the most probable contact distances of these atoms with a selected group of apolar atoms. In addition, a set of Coulombic radii for polar atoms are derived from their contact with water. In a study of buried waters and internal cavities in monomeric proteins, Williams et al.[3] have used the shortest atom–atom distance distributions from 75 high-resolution (≤2.5 Å) protein structures to derive the characteristic polar and apolar (referred here as the Coulombic and the VDW) radii of four protein atom types and water. Here, we define 25 atom types and use a dataset of 1,169 structures, 82% of which have resolution equal or better than 2.5 Å. In addition, we calculate the contact distance distributions using an iterative procedure. First, a starting set of VDW and Coulombic radii is assumed and used to define a contact between each pair of atoms. Next, a new set of VDW and Coulombic radii is derived from the resulting distance distributions. Then, the newly derived radii are used to define a contact. This process continues until the distance distributions, and hence the VDW and Coulombic radii, converge. This way, the radii of atoms are derived in a self-consistent fashion. The derived radii are compared with those derived from crystals of small molecules and with the force field parameter r_0 used in several macromolecular simulation packages.

The newly derived radii are used to obtain contact distance distributions across protein–protein interfaces using a list of interfaces given by Tsai et al.[25] We find that atom contact distance distributions across protein–protein interfaces are very similar to those in protein monomers. This result indicates that interfaces are well packed, in agreement with Walls and Sternberg.[26] Other applications and implications of the VDW and Coulombic radii are also discussed.

MATERIALS AND METHODS

Protein Structure Dataset

The protein structures used were those selected by Hobohm and Sander[24] with a sequence identity of less than 75% between any two structures. The list used was updated in March 1997 and contains 1,405 entries, of which 214 are solved by nuclear magnetic resonance (NMR) and 22 have no coordinates for sidechain atoms. The reason for choosing such a high threshold of sequence identity was that our main focus was the contact distance distribution, not the secondary structure of proteins. A difference of sev-

TABLE I. The 25 Atom Types

Number	Atom or Group	Symbol	Notes
1	$>\mathbf{C}\mathrm{HR}$	CA	Main-chain α-carbon (excluding α-carbon of Gly)
2	$>\mathbf{C}{=}\mathrm{O}$	C	Main-chain carbonyl carbon
3	$>\mathbf{C}\mathrm{H}—$	CH	Side-chain aliphatic carbon with one hydrogen (C^{β} of Ile, C^{γ} of Leu, C^{β} of Thr, C^{β} of Val)
4	$>\mathbf{C}\mathrm{H}_2$	CH2	Side-chain aliphatic carbon with two hydrogens, except those at β-position and those next to a charged group (C^{γ} of Arg, $C^{\gamma1}$ of Ile, C^{γ} and C^{δ} of Lys, C^{γ} of Met, C^{γ} and C^{δ} of Pro)
5	$>\mathbf{C}\mathrm{H}_2^{\beta}$	CH2b	Side-chain aliphatic carbon with two hydrogens at β-position (C^{β} of Arg, Asn, Asp, Cys, Gln, Glu, His, Leu, Lys, Met, Phe, Pro, Ser, Trp, Tyr)
6	$>\mathbf{C}\mathrm{H}_2^{\mathrm{ch}}$	CH2ch	Side-chain aliphatic carbon next to a charged group (C^{δ} of Arg, C^{γ} of Glu, C^{ϵ} of Lys)
7	$-\mathbf{C}\mathrm{H}_3$	CH3	Side-chain aliphatic carbon with three hydrogens (C^{β} of Ala, $C^{\gamma2}$ and $C^{\delta1}$ of Ile, $C^{\delta1}$ and $C^{\delta2}$ of Leu, $C^{\gamma2}$ of Thr, $C^{\gamma1}$ and $C^{\gamma2}$ of Val)
8	$-\mathbf{C}\mathrm{H}{=}$	CHar	Aromatic carbon with one hydrogen (carbon atoms on the rings of Phe, Trp and Tyr)
9	$>\mathbf{C}{=}$	Car	Aromatic carbon with no hydrogen (C^{γ} of Phe, C^{γ} and $C^{\epsilon2}$ of Trp, C^{γ} of Tyr)
10	$-\mathbf{C}\mathrm{H}{=}$	CHim	C^{δ} and C^{ϵ} on the imidazole side-chain of His
11	$>\mathbf{C}{=}\mathrm{O}$	Cco	Side-chain carbonyl carbon (C^{γ} of Asn, C^{δ} of Gln)
12	$-\mathbf{C}\mathrm{OO}^-$	Ccoo	Side-chain carboxyl carbon (C^{γ} of Asp, C^{δ} of Glu)
13	$-\mathbf{S}\mathrm{H}$	SH	S on Cys
14	$-\mathbf{S}—$	S	S on Met
15	$>\mathbf{N}\mathrm{H}$	N	Main-chain amide nitrogen
16	$>\mathbf{N}\mathrm{H}$	NH	Side-chain nitrogen with one hydrogen ($N^{\epsilon1}$ of Trp)
17	$>\mathbf{N}\mathrm{H}n^+$	NH+	$N^{\delta2}$ and $N^{\epsilon1}$ of His ($n = 0$ or 1; may be partially charged)
18	$-\mathbf{N}\mathrm{H}_2$	NH2	Side-chain neutral nitrogen with two hydrogen ($N^{\delta2}$ of Asn, $N^{\epsilon2}$ of Gln)
19	$-\mathbf{N}\mathrm{H}_2^+$	NH2+	Side-chain partially charged nitrogen on Arg
20	$-\mathbf{N}\mathrm{H}_3^+$	NH3+	Side-chain nitrogen on Lys
21	$>\mathrm{C}{=}\mathbf{O}$	O	Main-chain carbonyl oxygen
22	$>\mathrm{C}{=}\mathbf{O}$	Oco	Side-chain carbonyl oxygen ($O^{\delta1}$ of Asn, $O^{\epsilon1}$ of Gln)
23	$-\mathrm{C}\mathbf{OO}^-$	Ocoo	Side-chain carboxyl oxygen ($O^{\delta1}$ and $O^{\delta2}$ of Asp, $O^{\epsilon1}$ and $O^{\epsilon2}$ of Glu)
24	$-\mathbf{O}\mathrm{H}$	OH	Side-chain hydroxyl oxygen (O^{γ} of Ser, $O^{\gamma2}$ of Thr, O^{η} of Tyr)
25	$\mathrm{H}_2\mathbf{O}$	H2O	Water oxygen

eral residues in a hydrophobic core, for example, is enough to alter the distance distributions between atoms such that the data are not statistically redundant. In addition, the majority of the structures have a sequence identity of less than 65% (the number of entries in the sequence identity range 65–75% is only 80). The entries that do not have coordinates for sidechains and those that are solved by NMR were ignored, leaving 1,169 structures actually being used. Of the 1169 structures, 954 have resolution ≤2.5 Å. The worst resolution is 3.5 Å. In the list given by Hobohm and Sander,[24] one chain is selected from each entry. Here, we used all the chains in each entry since there might be crystal-bound waters at the chain–chain interface. Thus, the distance distributions contained a small contribution from contacts at the chain–chain interface. Comparison of these distributions with those obtained using one chain from each entry did not reveal any significant difference.

The list of protein–protein interfaces given by Tsai et al.[25] contains 795 entries at level A. Some of these interfaces are from real protein–protein complexes, while others are due to crystal contacts. In the interface calculation, contacts within each monomer were ignored and only those across the interface were counted.

Twenty-Five Atom Types

Since hydrogen atoms are generally not observed in the crystallographic determination of protein structures, only heavy atoms were considered. Thus, for those with covalently bonded hydrogens, the contact radii are actually for the united atoms. From a practical point of view, one would like the number of atom types to be as small as possible. However, the number of atom types must be large enough so that the variation in contact radius is reflected. Here, we classify atoms into 25 types (Table I). The first 14 types are carbon and sulfur atoms with a varying number of bonded hydrogens and/or different covalent bonding environments. These atoms can be considered as apolar or hydrophobic. The last 11 types are nitrogen and oxygen atoms, with a varying number of bonded hydrogens, that are either polar or charged. The placement of the boundary between polar and apolar is somewhat arbitrary. Some of the apolar atoms, for example the carbon atoms that are covalently bonded to either polar or charged atoms, may have substantial polar character. The covalent bonding pattern for each atom is shown in the second column of Table I, with the atom under consideration in bold typeface. In the third column, we designate a

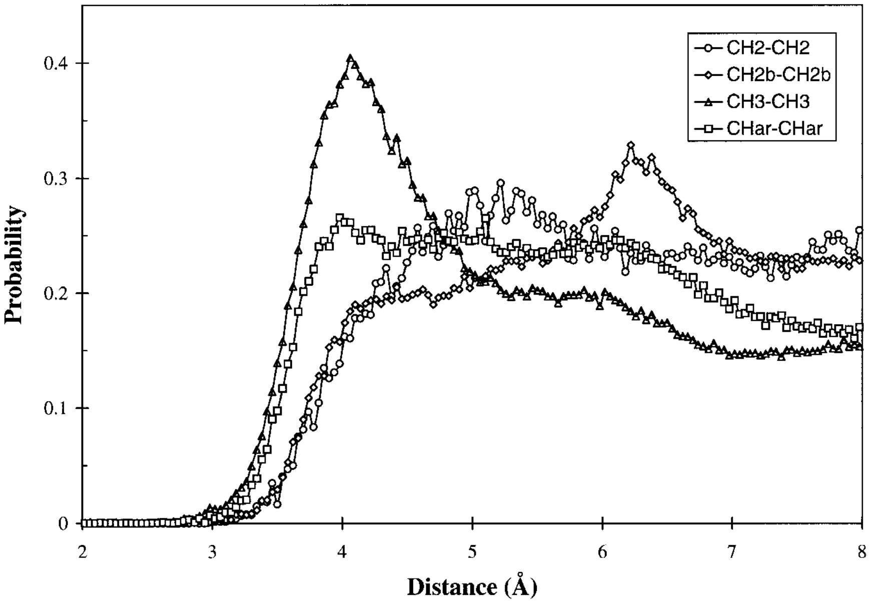


Fig. 1. Distance distributions obtained by simply binning the interatomic distances. Only atoms separated by four or more residues in sequence are considered. The count in each bin is scaled by r_n^2 first, where r_n is the distance of the center of bin n, the curves are then normalized to have unit area.

symbol for each atom type. The classification is quite detailed. For example, mainchain carbonyl carbon and oxygen are distinguished from their sidechain counter parts; sidechain aliphatic carbons with two hydrogens were divided into three types, since atoms at the β-position are likely to be affected by the secondary structure of the proteins and those near a charged group by charge–charge interactions. Our intention here is to define as many atom types as possible and the users can consolidate some of the types according to their own needs.

Several atoms are not classified into any of the 25 types in Table I. These are C^{ζ} of Arg, C^{γ} of His, and N of Pro. The number of these atoms in the dataset is too small for each of them to be considered a separate type. In obtaining the final results, the C_{α} atom of Gly is excluded from the CA atom type. Tests show that the distance distributions involving CA atoms barely change when C_{α} of Gly is included. The radii for the above atoms can be estimated from atoms with similar covalent bonding environments.

To simplify the description, below we use the symbols in the third column of Table I to refer to each atom type. Thus, for the purpose of this article, we have 25 distinct atoms, i.e., the CA atom, the CH atom, etc. Also, polar and charged atoms are sometimes collectively referred to as polar.

Distance Distribution Calculation

Atom–atom contacts in a homogeneous liquid (or crystal) of small molecules are mainly intermolecular and distance distribution calculation in this case is relatively straightforward. One simply divides the distance between two atoms by a predefined bin size and increments the counts in the appropriate bin. After a volume scaling, the first peak in the distribution defines the position of the first shell neighbors, hence the contact radii for the two atom types. For proteins, the contacts of interest are mainly intramolecular. In addition, the calculation of distance distributions is complicated by several factors. First, there are many atom types with large variations in sizes. Second, the spatial distribution of atoms is far from homogeneous. Instead, it is strongly affected by the secondary and tertiary structure of the protein, and by the covalent link of the peptide chain. Third, an atom in a liquid or crystal of small molecules can be considered embedded in an infinite medium, while proteins have finite sizes and surface atoms, which face solvent on one side, constitute a large fraction of the total. Because of these factors, simply binning the atom–atom distances does not produce distributions that clearly show the position of the first shell neighbors, as shown in Figure 1.

In this figure, only the CH3–CH3 distance distribution shows a clear peak at short distance, the others only show the rising edge.

To overcome the above difficulties, we confined the distance distribution calculations to only those atoms that are in direct contact. Two atoms, *i* and *j*, are in direct contact if, one, they are within a cutoff distance of 5.2 Å and, two, there is no other atom whose sphere-to-sphere distances to atoms *i* and *j* are both shorter than the sphere-to-sphere distance between *i* and *j*, that is, there are no other atoms that lie in between them. In addition, they must be separated by four or more residues in sequence (i.e., Res#(*i*) - Res#(*j*) $\geq$ 5) so that the distance is not affected by the covalent link and the helical structure of the peptide chain. There is no way to avoid the effect of β-sheets on the contact distance distributions. However, we use mainly sidechain atoms as reference to measure the contact radius of each atom and hence the effect of β-sheets is minimized. The 5.2-Å cutoff is chosen because at such distance the probability for two atoms to be in direct contact is approximately zero.

We emphasize that here direct contact is not defined by a fixed distance separation. For example, it is possible for one N and O pair to be separated by 5.1 Å and still be considered to be in direct contact, while a pair separated by 3.5 Å to be considered not in direct contact. The probability for the former, however, is very small, as we show later. Using such a definition of direct contact, self-consistency is achieved without imposing arbitrary cutoff distances.

If atom *j* satisfies the criteria (described in detail below) of being in direct contact with atom *i*, then its distance to atom *i* is registered in bin number $n = r_{ij}/\Delta r_{\text{bin}}$ of the distribution $F_{ti,tj}(n)$, where *ti* and *tj* represent the atom type of atoms *i* and *j*, respectively, and Δr_{bin} is the bin size, chosen to be 0.04 Å. Finally, after all the entries in the database are analyzed, the distributions are normalized to obtain the probability distributions

$$P_{ti,tj}(r_{\text{n}}) = \frac{F_{ti,tj}(n)/r_{\text{n}}^2}{\sum_{m=1}^{N} (\Delta r_{\text{bin}} \cdot F_{ti,tj}(m)/r_{\text{m}}^2)}$$

where *N* is the total number of bins, r_{n} is the distance of the center of bin *n*, and $P_{ti,tj}(r_{\text{n}}) \cdot \Delta r$ is the probability, within the cutoff distance, of observing atom types *ti* and *tj* in a distance range Δr at r_{n}.

In order to determine if two atoms are in direct contact, we need to know the radii of all the atoms involved. This is not possible since these are exactly the quantities we are trying to determine. To resolve this dilemma, we use an iterative procedure. First, we assume that all the atoms are of the same size and obtain a set of contact distance distributions, from which a preliminary set of VDW and Coulombic radii are derived (described below). Next, the preliminary radii are used to determine if two atoms are in direct contact and a new set of distance distributions is obtained. This procedure is continued until the distributions, and hence the radii, converge. In each of the iterations, the same definition of direct contact is used, while the radii of the atoms change.

Figure 2 illustrate the criteria for direct contact between two atoms for the first and subsequent iterations. To determine if atoms *i* and *j* are in direct contact in the first iteration, we draw two arcs centered on the two atoms using the distance between them as a radius (Fig. 2a). For atoms *i* and *j* to be in direct contact, no third atom should be allowed to enter the area enclosed by the two arcs, defined as the excluding zone. In Figure 2a, we draw equal-sized circles around each atom using an arbitrary radius. It is apparent that, in the case drawn, the contacts atom *k* makes with both atom *i* and atom *j* are better than that between atoms *i* and *j*. Therefore, atoms *i* and *j* are separated by atom *k* and the distance r_{ij} is discarded. While these criteria may seem too stringent, they ensure that atom *j* is the closest to atom *i* in the area (the excluding zone) associated with the direction $\vec{r}_{ij}$.

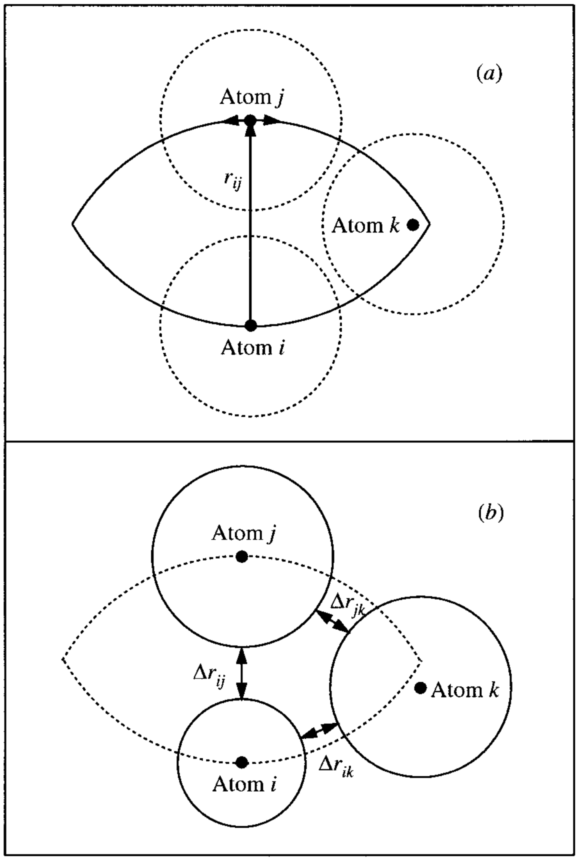


Fig. 2. Conditions for two atom, *i* and *j*, to be in direct contact in (**a**) the initial and (**b**) the subsequent iterations. In the initial iteration, all atoms are assumed to have the same size, and the area enclosed by the two arcs is excluded to a third atom if atoms *i* and *j* are to be in direct contact. In subsequent iterations, the radii obtained in the previous run is used and atom *k* is said to separate atoms *i* and *j* if the following conditions are satisfied: $\Delta r_{ik} < \Delta r_{ij}$ and $\Delta r_{jk} < \Delta r_{ij}$. In Figure 2b, atom *k* is moved out of the area enclosed by the two arcs, but it still interferes with the direct contact of atoms *i* and *j* due to atom size changes.

In subsequent iterations, the excluding zone (drawn with dotted lines in Fig. 2b) is no longer used. Instead, the VDW and Coulombic radii obtained in the first iteration are utilized to determine if two

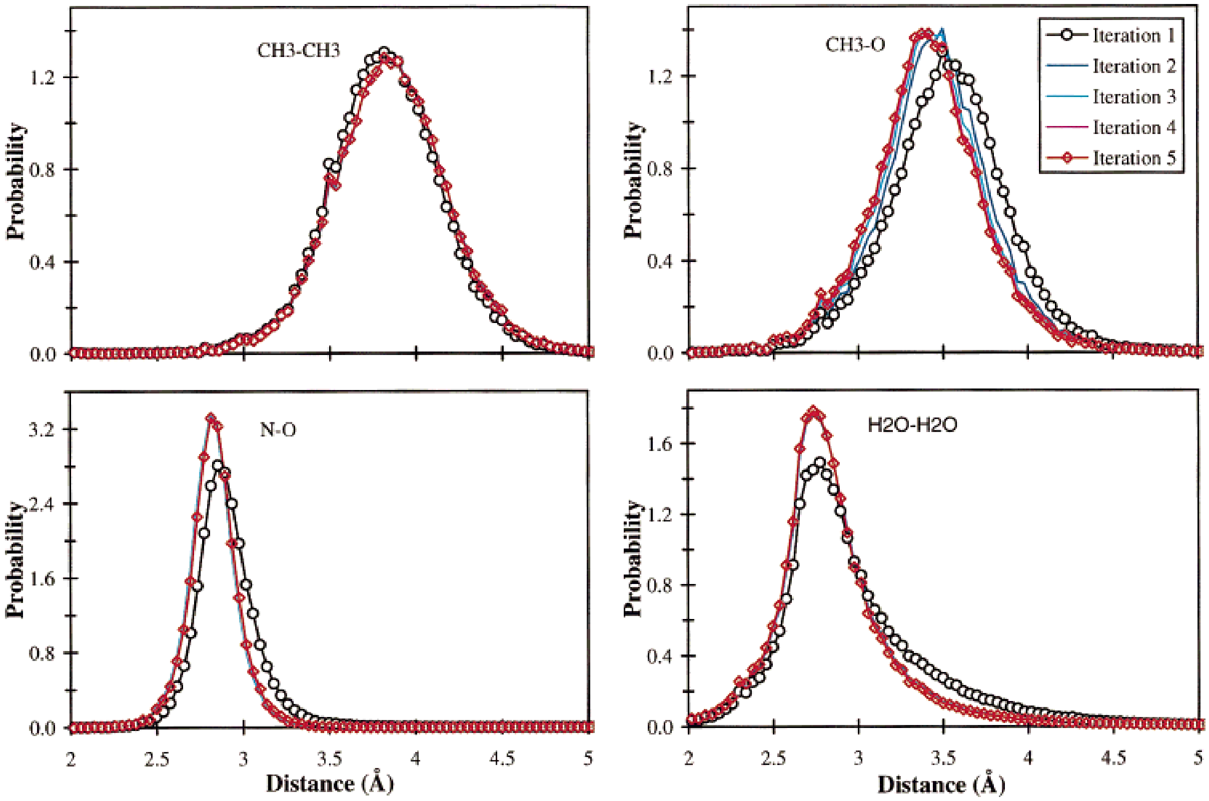


Fig. 3. Some typical contact distance distributions from each iteration showing the process of convergence. The change from the first to the second iteration is the largest. Subsequent iterations cause only small changes to the distributions. Most distributions converge at the second iteration. As a result, the lines representing iterations 2–4 are barely visible in three of the four panels.

atoms are in direct contact. In Figure 2b, circles are drawn around each atom according to its radius. Three distances, Δr_{ij}, Δr_{ik}, and Δr_{jk}, are calculated. Atom k interferes with the direct contact between atoms i and j only if both Δr_{ik} and Δr_{jk} are shorter than Δr_{ij}. In case two of the three atoms are polar and have opposite polarity, say j and k, their Coulombic radii are used to calculate Δr_{jk} and their VDW radii are used to calculate Δr_{ij} and Δr_{ik}. By using the refined radii to define a direct contact, a larger atom can interfere with the direct contact between two smaller ones even if it lies outside the excluding zone defined previously. On the other hand, a smaller atom can enter the excluding zone without affecting the direct contact between two larger ones. Thus, the determination of direct contact becomes more accurate.

By assuming that atoms are of the same size in the first iteration, the contact radii of smaller atoms are overestimated and, vice versa, those of larger atoms are underestimated. As iteration continues, each distribution gradually converges to a constant peak position. This is illustrated in Figure 3 for several distributions. Most of the distributions converge fairly quickly. The exception is the CH3–O distribution, which does not converge until the fourth iteration (the distributions from the fifth iteration are almost indistinguishable from those of the fourth iteration). The reason for this is not clear at present.

Figure 4 shows some typical distance distributions from the fifth (the last) iteration. The distributions are fairly wide, especially the apolar–apolar and apolar–polar ones, reflecting the different packing environments atoms experience in proteins. Most of the distributions involving sidechain apolar atoms are nearly Gaussian-shaped. For opposite-polarity polar atom pairs, the rising edge is steeper than the falling edge, reflecting the long-range nature of the polar interaction and the fact that these interactions occur mostly at the surface (with the exception of the N–O pair). Most of the apolar–apolar and apolar–polar distance distributions can be fitted by two Gaussians, one is narrower and the other is wider and shifted slightly outwards. In this study, we simply took the peak position of each distribution as the most probable contact distance for the corresponding atom pair. No assumptions regarding the shape of the distributions were made and the fitting was simply used to locate the peak position. Unless specified, the distributions used are all from the last iteration.

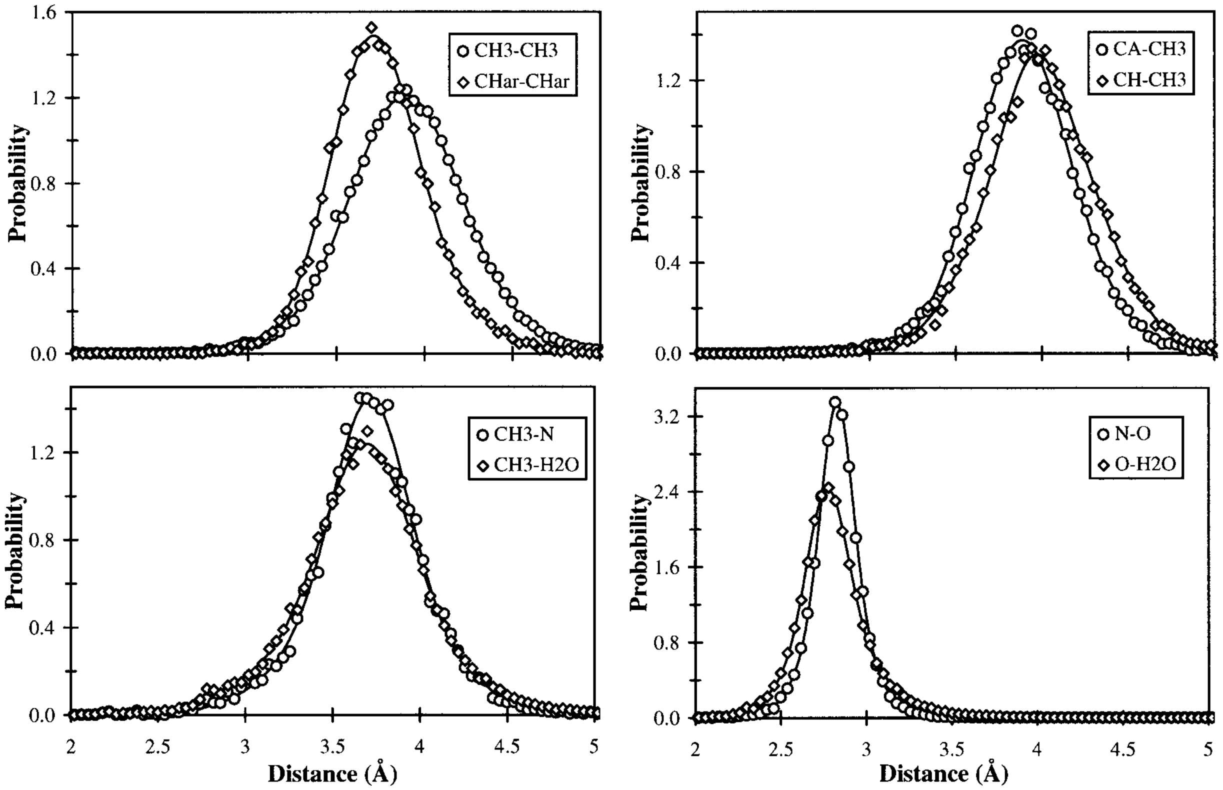


Fig. 4. Some typical distance distributions. The distributions are normalized as described in the text. The solid curves are fit to the data using two Gaussians.

Determination of van der Waals and Coulombic Radii

The interactions between apolar–apolar and apolar–polar atom pairs are mainly van der Waals in nature. Thus, these distributions can be used to derive the VDW radius of atoms. Since pair contact frequencies and the statistics of the corresponding distributions vary considerably, we choose to determine the VDW radii of CH2, CH2b, CH3, and CHar (see Table I for atom symbols) first, and then use these as references to measure the VDW radii of the rest of atoms. These four atoms have relatively high contact frequency among themselves and with other atoms (see below). For the Coulombic radius, we chose H2O as a reference since it can interact favorably with atoms of both positive and negative polarities.

RESULTS AND DISCUSSION

Atom Contact Frequencies

Table II lists the relative contact frequencies for the total of 300 possible atom pairs. The most frequent contact occurs between water molecules, with a total number of occurrence of 3.3 X 10^6, followed by CH3–CH3, H2O–O, CA–O, H2O–CH2b, and N–O. Since we only consider contacts between atoms separated by four or more residues (i.e., *i* → *i* + 5), regular *i* → *i* + 4 N–O contacts in helices are not counted. In general, pairs involving H2O and CH3 have high contact frequencies. It is interesting to note that even though the number of N and O atoms in a protein is roughly the same, the total number of contacts involving O is much larger than N (excluding the contacts between N and O and adding up the rest of the contact frequencies involving N gives 0.188, while the same number for O is 1.6). One reason for this difference is that O has a single link (though a double bond) with the peptide chain and is less buried by covalently bonded neighbors, while N has a double link and is more buried. Also, oxygens contact frequency with water is 10 times larger than that for nitrogen. A ratio of about 2 is expected since O can form two hydrogen bonds with water. This suggests that there are more main-chain oxygen atoms exposed to water than amide nitrogens, at least for the crystallized waters in the database and using the current definition of direct contact (different definitions of contact may result in different relative contact frequencies).

van der Waals and Coulombic Radii

Table III lists the peak positions of the contact distance distributions among the apolar reference atoms. Initially, the radius for each atom was chosen

TABLE II. Relative Contact Frequencies for the 300 Atom Pairs

	CA	C	CH	CH2	CH2b	CH2ch	CH3	CHar	Car	CHim	Cco	Ccoo	SH	S	N	NH	NH+	NH2	NH2+	NH3	O	Oco	Ocoo	OH	H2O
CA	0.010	0.002	0.005	0.017	0.034	0.004	0.077	0.053	0.005	0.003	0.002	0.002	0.003	0.005	0.007	0.001	0.002	0.008	0.010	0.002	0.353	0.014	0.027	0.032	0.219
C		0.001	0.001	0.005	0.009	0.001	0.019	0.015	0.001	0.001	0.000	0.000	0.001	0.001	0.002	0.001	0.001	0.003	0.005	0.002	0.007	0.002	0.002	0.007	0.052
CH			0.005	0.008	0.023	0.002	0.044	0.020	0.002	0.001	0.001	0.001	0.002	0.001	0.002	0.000	0.001	0.003	0.003	0.001	0.031	0.003	0.008	0.009	0.050
CH2				0.022	0.051	0.007	0.093	0.057	0.008	0.003	0.002	0.004	0.004	0.005	0.006	0.001	0.002	0.006	0.009	0.001	0.072	0.008	0.021	0.020	0.120
CH2b					0.111	0.014	0.240	0.129	0.013	0.008	0.005	0.005	0.009	0.010	0.015	0.003	0.007	0.015	0.021	0.005	0.163	0.018	0.039	0.048	0.320
CH2ch						0.003	0.023	0.016	0.002	0.001	0.001	0.001	0.001	0.001	0.002	0.000	0.001	0.003	0.005	0.002	0.045	0.005	0.026	0.011	0.087
CH3							0.543	0.198	0.015	0.012	0.007	0.006	0.012	0.015	0.023	0.003	0.007	0.015	0.023	0.004	0.131	0.014	0.021	0.040	0.161
CHar								0.101	0.008	0.007	0.005	0.004	0.005	0.006	0.018	0.002	0.005	0.010	0.016	0.003	0.082	0.008	0.016	0.019	0.084
Car									0.000	0.001	0.000	0.000	0.000	0.000	0.002	0.000	0.001	0.001	0.002	0.000	0.002	0.001	0.000	0.001	0.006
CHim										0.001	0.001	0.001	0.001	0.001	0.002	0.000	0.001	0.001	0.002	0.001	0.018	0.002	0.008	0.006	0.028
Cco											0.000	0.000	0.000	0.000	0.002	0.000	0.001	0.001	0.001	0.001	0.003	0.001	0.001	0.002	0.019
Ccoo												0.000	0.000	0.000	0.008	0.001	0.003	0.003	0.016	0.010	0.002	0.000	0.001	0.007	0.050
SH													0.055	0.000	0.002	0.000	0.001	0.001	0.001	0.000	0.004	0.000	0.001	0.001	0.005
S														0.000	0.001	0.000	0.000	0.001	0.001	0.000	0.002	0.000	0.001	0.001	0.004
N															0.003	0.001	0.002	0.008	0.005	0.001	0.279	0.010	0.013	0.007	0.043
NH																0.000	0.001	0.000	0.000	0.000	0.008	0.001	0.003	0.002	0.006
NH+																	0.004	0.001	0.002	0.001	0.007	0.001	0.006	0.003	0.013
NH2																		0.003	0.005	0.002	0.043	0.006	0.009	0.007	0.050
NH2+																			0.007	0.001	0.057	0.006	0.043	0.006	0.056
NH3+																				0.001	0.017	0.002	0.011	0.002	0.025
O																					0.036	0.015	0.012	0.031	0.459
Oco																						0.003	0.006	0.006	0.047
Ocoo																							0.014	0.022	0.144
OH																								0.011	0.082
H2O																									1.000

TABLE III. Peak Positions of the Distance Distributions and van der Waals Radii for the Apolar Reference Atoms

	CH2	CH2b	CH3	CHar	vdW Radii[a]
CH2	3.86	3.84	3.84	3.72	1.924
CH2b		3.82	3.82	3.72	1.907
CH3			3.84	3.76	1.921
CHar				3.66	1.824

[a] Three significant figures are kept here. The final results are rounded off to two significant figures.

to be one-half the distance of the peak in the self-contact distance distribution. Then, a 0.2-Å window centered on the tentative radius of each atom (i.e., $R \pm 0.1$ Å) was given and a systematic search through each window, in a step of 0.001 Å, was made to find the minimum of the residual given by,

$$R^2 = \sum_{i=1}^{4}\sum_{j=1}^{4} (R_{vdW,i} + R_{vdW,j} - D_{ij})^2$$

where D_{ij} is the peak position of the distribution P_{ij}. The initial and final residuals are 0.0023 and 0.0016 Å^2, respectively. The final values for the VDW radii of these reference atoms are listed in the last column of Table III.

The VDW radii for the rest of the atoms were calculated based on the contact distance distributions with the above reference atoms and results are listed in Table IV. The VDW radii of carbon atoms vary from 1.74 to 2.01 Å. The smallest are triangle carbons with no hydrogen (C, Car, CHim, Cco, and Ccoo) and the largest is the CH atom. In general, triangle carbons are smaller than tetrahedral carbons by more than 0.15 Å. It is not clear why the radius of CH is significantly larger. One reason may be that CH has a relatively low polarizability.[20] In addition, CH is covalently bonded to three heavy atoms that may prevent it from making good contacts with other atoms. However, atom CA, having similar covalent bonding environment as CH, has a radius approximately equal to those of CH2 and CH3.

The VDW radii for polar atoms have relatively large uncertainties. The CH2b–N and CH2b–O distributions are not used in determining the VDW radii of N and O since CH2b is right next to the mainchain and these distributions may be affected by the secondary structure of the proteins.[4] On average, the VDW radius for nitrogen is 1.66 ± 0.03 Å and that for oxygen is 1.51 ± 0.02 Å (excluding H2O). H2O has a significantly larger VDW radius than other oxygen atoms, as was also found by Gerstein et al.[4] in their molecular dynamics simulations. The radii of oxygen atoms as determined by CH3 and CHar, especially CHar, are consistently larger than those determined by CH2 and CH2b. The reason for this is not completely clear. It has been suggested that an aromatic ring can act as a hydrogen bond acceptor.[27]

TABLE IV. Peak Positions of the Distance Distributions with Apolar Reference Atoms and the van der Waals Radii for the 25 Atom Types

	Peak Positions				R_{vdW}				
Symbol	CH2	CH2b	CH3	CHA	CH2	CH2b	CH3	CHA	Avg
CA	3.83	3.81	3.84	3.70	1.906	1.903	1.919	1.876	1.90 ± 0.02[a]
C	3.67	3.64	3.68	3.59	1.746	1.733	1.759	1.766	1.75 ± 0.01
CH	3.94	3.94	3.94	3.80	2.016	2.033	2.019	1.976	2.01 ± 0.02
CH2	3.86	3.84	3.84	3.72	1.936	1.933	1.919	1.896	1.92 ± 0.02
CH2b	3.84	3.82	3.82	3.72	1.916	1.913	1.899	1.896	1.91 ± 0.01
CH2ch	3.80	3.82	3.79	3.68	1.876	1.913	1.869	1.856	1.88 ± 0.02
CH3	3.84	3.82	3.84	3.76	1.916	1.913	1.919	1.936	1.92 ± 0.01
CHar	3.72	3.72	3.76	3.66	1.796	1.813	1.839	1.836	1.82 ± 0.02
Car	3.64	3.62	3.68	3.58	1.716	1.713	1.759	1.756	1.74 ± 0.02
CHim	3.68	3.62	3.67	3.58	1.756	1.713	1.749	1.756	1.74 ± 0.02
Cco	3.74	3.72	3.71	3.63	1.816	1.813	1.789	1.806	1.81 ± 0.01
Ccoo	3.71	3.62	3.66	3.62	1.786	1.713	1.739	1.796	1.76 ± 0.04
SH	3.79	3.78	3.80	3.72	1.866	1.873	1.879	1.896	1.88 ± 0.01
S	3.82	3.83	3.87	3.81	1.896	1.923	1.949	1.986	1.94 ± 0.04
N	3.68	3.60	3.59	3.52	1.756	x[c]	1.669	1.696	1.71 ± 0.04
NH	3.58	3.56	3.60	3.48	1.656	1.653	1.679	1.656	1.66 ± 0.01
NH+	3.62	3.50	3.60	3.46	1.696	1.593	1.679	1.636	1.65 ± 0.05
NH2	3.54	3.52	3.54	3.44	1.616	1.613	1.619	1.616	1.62 ± 0.01
NH2+	3.60	3.59	3.60	3.48	1.676	1.683	1.679	1.656	1.67 ± 0.01
NH3+	—[b]	3.55	3.60	3.52	—	1.643	1.679	1.696	1.67 ± 0.03
O	3.36	3.32	3.41	3.37	1.436	x[c]	1.489	1.546	1.49 ± 0.06
Oco	3.42	3.38	3.46	3.40	1.496	1.473	1.539	1.576	1.52 ± 0.05
Ocoo	3.38	3.35	3.44	3.35	1.456	1.443	1.519	1.526	1.49 ± 0.04
OH	3.42	3.43	3.47	3.41	1.496	1.523	1.549	1.586	1.54 ± 0.04
H2O	3.56	3.55	3.62	3.55	1.636	1.643	1.699	1.726	1.68 ± 0.04

[a] Standard deviation calculated using $\sigma = [E(x - x_0)^2/(n - 1)]^{1/2}$, where n = 4 for most cases.
[b] Distribution has poor statistics.
[c] Distributions affected by secondary structure of proteins, data not used.

Thus, CHar may contain some negative-charge character, which would cause unfavorable interactions with negatively charged oxygen atoms.

The Coulombic radius for the polar reference atom, H2O, is determined from the H2O–H2O distance distribution to be 1.37 Å. The Coulombic radii for the rest of the polar atoms are calculated from distance distributions with H2O and the results are listed in Table V, together with the differences, $R_{\text{Coul}} - R_{\text{vdW}}$, from the corresponding VDW radii. We find that the Coulombic radii of polar atoms are, on average, 0.18 Å smaller than their VDW radii. Water shows the largest difference of 0.3 Å. The average Coulombic radii for neutral and charged nitrogens are 1.51 and 1.43 Å, respectively. The average Coulombic radius for oxygen is 1.38 Å.

TABLE V. Peak Positions of the Contact Distance Distributions of Polar Atoms with Water and Their Coulombic Radii

Symbol	Peak (Å)	R_{Coul}	$R_{\text{Coul}} - R_{\text{vdW}}$
N	2.86	1.49	−0.22
NH	2.92	1.55	−0.11
NH+	2.78	1.41	−0.24
NH2	2.86	1.49	−0.13
NH2+	2.86	1.49	−0.18
NH3+	2.77	1.40	−0.27
O	2.78	1.41	−0.08
Oco	2.78	1.41	−0.11
Ocoo	2.72	1.35	−0.14
OH	2.72	1.35	−0.19
H2O	2.74	1.37	−0.31

So far, we have ignored cases of same-polarity polar–polar contacts. The statistics for these distance distributions are, in general, very poor and an accurate determination of contact radii using these distributions is not possible. The estimated peak positions for the N–N and O–O contact distance distributions are 3.5 and 3.0 Å, respectively, roughly twice their VDW radii. These two distributions, however, may be dictated by the secondary structure of proteins and not due to true packing contacts. Nevertheless, since the probability of forming such contacts is low, a separate set of radii for same-polarity polar–polar contacts does not seem to be necessary and the VDW radii can be used in case same-polarity polar contacts occurs.

Examination of radii values listed in Tables IV and V indicates that the number of atom types can be reduced, as some atoms have very similar radii. For example, the tetrahedral (or sp^3) carbon atoms CA, CH3, CH2b, CH2ch, and CH3 can be considered as one type for most practical purposes. A minimum set of atom types that we suggest is listed in Table VI,

TABLE VI. A Minimum Set of Atom Types and the Corresponding vdW and Coulombic Radii

Atom	R_{vdW} (Å)[a]	R_{Coul} (Å)[a]	Notes
CHn (sp^3)	1.92 (1.90)		Tetrahedral carbons with 1 to 3 hydrogens
CH (sp^2)	1.82 (1.80)		Triangle carbons with 1 hydrogen
C (sp^2)	1.74 (1.75)		Triangle carbons with no hydrogen
S	1.92 (1.90)		All sulfur atoms (S and SH)
N	1.66 (1.65)	1.47 (1.50)	All nitrogen atoms with or without covalently bonded hydrogens
O	1.51 (1.50)	1.38 (1.40)	All oxygen atoms except water with or without covalently bonded hydrogens
H2O	1.68 (1.70)	1.37 (1.40)	Water

[a]Values in parentheses are the derived values rounded off to the nearest 0.05, except for the Coulombic radius of N and H2O which are rounded off to the nearest 0.1.

which is consistent with most atom type sets currently in use in the literature. The users, however, can make their own reduced sets based on values listed in Tables IV and V.

Cross Validation

One way of testing the radii derived above is to see if they can predict the peak positions for distributions that are not used to derive them. Table VII lists the values of $D_{ij} - (R_i + R_j)$, the difference between the peak position of the distance distribution P_{ij} and the sum of the appropriate radii of atom types i and j, for selected distributions that have reasonable statistics. All of the peak positions are within 0.1 Å of those predicted by the corresponding radii, indicating that the radii derived above are self-consistent.

Comparison With Radii Derived by Williams et al.

In a study of buried waters and internal cavities in monomeric proteins, Williams et al.[3] have used the contact distance distributions in 75 high-resolution (≤2.5 Å) protein structures to derive the characteristic polar and apolar (referred here as the Coulombic and the VDW) radii of four protein atom types and water. Comparison of the radii derived here, appropriately averaged, with those of Williams et al. is shown in Table VIII. It can be seen that the VDW and Coulombic radii for water and the VDW radii for sulfur, nitrogen, and oxygen differ significantly. The main source of discrepancy may stem from the different criteria for defining a contact between two atoms. The difference in the number of protein structures used may also affect the results.

TABLE VII. Values of $D_{ij} - (R_i + R_j)$ (in Å) for Selected Atom Pairs That Are Not Used in the Derivation of vdW and Coulombic Radii[a]

a: vdW Radii

	CA	CH2ch	H2O
CA	0.02	0.03	−0.06
C	0.05	−0.03	−0.01
CH	−0.05	—	−0.07
CH2ch	0.03	−0.02	−0.12
Car	0.00	−0.02	0.02
SH	−0.02	—	−0.04
S	−0.04	−0.07	−0.10

b: Coulombic Radii

	O	Oco	Ocoo	OH
N	−0.07	−0.07	−0.01	0.06
NH	−0.04	—	0.00	0.00
NH+	0.00	−0.10	−0.06	−0.06
NH2	−0.04	0.04	0.06	0.02
NH2+	−0.10	−0.06	−0.02	0.06
NH3+	−0.05	−0.07	0.03	0.03
OH	−0.06	−0.06	−0.05	—

TABLE VIII. Comparison of vdW and Coulombic Radii Derived Here with Those of Williams et al.[3]

	This Work[a]		Williams et al.	
Atom Type	R_{vdW} (Å)	R_{Coul} (Å)	R_{vdW} (Å)	R_{Coul} (Å)
Carbon	1.92	—	1.93	—
Sulfur	1.92	—	1.68	—
Nitrogen	1.66	1.47	1.82	1.45
Oxygen	1.51	1.38	1.62	1.35
Water	1.68	1.37	1.97	1.50

[a]All radii are averages, except for water. Carbon: the average of all tetrahedral carbon atoms; Sulfur: average of S and SH; Nitrogen: average of all nitrogen types; Oxygen: average of all oxygen types.

Comparison With Radii Derived From Crystals of Small Molecules

Table IX compares the VDW radii derived in this study with those derived from crystals of small molecules. The Coulombic radii for polar atoms are also listed. The number of atom types defined by Pauling,[1] Bondi,[2] and Chothia[13] is less than the number defined here. For example, all methylene groups have been defined as one type. The same is true for all aromatic carbons, sulfur, oxygen, and some of the nitrogen atoms. The radii for nitrogen and oxygen atoms given by Pauling are their ionic radii. In general, the agreement is better for apolar atoms. As for polar atoms, some of the existing radii agree with R_{vdW}, while others agree with R_{Coul}, indicating possible inconsistencies in the definition and use of VDW radii in the literature.[19]

Comparison With Force Field Parameter r_0

The force field parameter r_0 is related to the potential energy function used to describe the inter-

TABLE IX. Comparison of vdW Radii Derived Here With Those from Small Molecules and With Force Field Parameter r_0

	This work		Previous work (R_{vdW}, Å)			Force field parameter r_0 (Å)			
Atom	R_{vdW} (Å)	R_{Coul} (Å)	Pauling[a]	Bondi[b]	Chothia[c]	OPLS[d]	Charmm[e]	Amber[f]	ENCAD[g]
CA	1.90		2.0	2.0	1.87	2.13	2.37	1.91	2.10
C	1.75		—	—	1.76	2.10	2.10	1.91	2.05
CH	2.01		—	—	1.87	2.16	2.37	1.91	2.10
CH2	1.92		2.0	2.0	1.87	2.19	2.24	1.91	2.10
CH2b	1.91		2.0	2.0	1.87	2.19	2.24	1.91	2.10
CH2ch	1.88		2.0	2.0	1.87	2.19	2.24	1.91	2.10
CH3	1.92		2.0	2.0	1.87	2.19	2.24	1.91	2.10
CHar	1.82		1.7	1.77	1.76	2.10	2.17	1.91	2.05
Car	1.74		1.7	1.77	1.76	2.10	2.10	1.91	2.05
CHim	1.74		—	—	1.76	2.10	2.10	1.91	2.05
Cco	1.81		—	—	1.76	2.10	2.10	1.91	2.05
Ccoo	1.76		—	—	1.76	2.10	2.10	1.91	2.05
SH	1.88		1.85	—	1.85	1.99	1.89	2.00	2.10
S	1.94		1.85	1.80	1.85	1.99	1.89	2.00	2.10
N	1.70	1.49	1.5	1.65	1.65	1.82	1.60	1.82	1.85
NH	1.66	1.55	1.5	1.65	1.65	1.82	1.60	1.82	1.85
NH+	1.65	1.41	1.5	—	1.65	1.82	1.60	1.82	1.91
NH2	1.62	1.49	1.5	1.75	1.50	1.82	1.60	1.82	1.85
NH2+	1.67	1.49	1.5	—	1.50	1.82	1.60	1.82	1.91
NH3+	1.67	1.40	1.5	—	1.50	1.82	1.60	1.82	1.91
O	1.49	1.41	1.4	1.65	1.40	1.66	1.60	1.66	1.50
Oco	1.52	1.41	1.4	1.65	1.40	1.66	1.60	1.66	1.50
Ocoo	1.49	1.35	1.4	—	1.40	1.66	1.60	1.66	1.55
OH	1.54	1.35	1.4	—	1.40	1.72	1.60	1.72	1.50
H2O	1.68	1.37	1.4	—	1.40	1.77	1.70	1.77	1.68

[a] Ref 1. Values are based on contact distances in crystals of small molecules. Aliphatic carbons are treated as united atoms. The radius for aromatic carbons is the half-thickness of the ring molecules.

[b] Ref 2. Values are based on contact distances in crystals of small molecules.

[c] Ref. 13. Values are based on contact distances in crystals of amino acids.

[d] Ref 28. Values are converted from the parameter σ, the position at which the Lennard-Jones potential equals zero, using $r_0 = 2^{1/6}\sigma/2$. Hydrogens are explicitly present.

[e] Ref 29. Aliphatic carbons and covalent bonded hydrogens are treated as united atoms.

[f] Ref. 30. Hydrogens are explicitly present.

[g] Ref 31. Hydrogens are explicitly present.

actions between atoms. In such a force field, dipoles are usually approximated by assigning explicit partial charges to individual atoms and the nonbonded interaction energy between atoms i and j is often approximated by a Lennard-Jones potential plus a Coulombic term as follows:

$$V = \epsilon\left[\left(\frac{d_{ij}}{r_{ij}}\right)^{12} - 2\left(\frac{d_{ij}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{r_{ij}}$$

where ϵ and d_{ij} are the well depth and the position of the minimum, respectively, of the Lennard-Jones potential. d_{ij} is usually defined by an atomic radius parameter r_0, i.e., $d_{ij} = f(r_{0i},r_{0j})$. Two of the often used forms of the function are $d_{ij} = (r_{0i} + r_{0j})$ and $d_{ij} = 2(r_{0i}\cdot r_{0j})^{1/2}$. r_0 is usually referred to as a VDW radius.[20,30,32] It is important to realize the difference between r_0 and the VDW radius studied here. d_{ij} does not, in general, represent the most probable contact distance between atoms i and j. Rather, its value is chosen in such a manner that when the Coulombic interaction between them and the collective interactions with all surrounding neighbors are considered, the equilibrium separation is reproduced. In addition, if the second definition of d_{ij} is used, the r_0 values are no longer additive, especially if atoms i and j differ significantly in size. Gerstein et al.[4] have calculated an effective VDW radius for each type of atoms by setting the Lennard-Jones part of the interaction energy between atoms of the same type to $0.25k_B$T, where k_B is the Boltzmann constant and T = 300 K. The values obtained agree well with those of Chothia.[13]

Values of r_0 from several force fields are summarized in columns 7–10 of Table IX. The derived VDW radii differ, in some cases significantly, from the corresponding force field parameter r_0. In addition, r_0 varies from one force field to another for the same atom. This indicates that using force field parameter r_0 in places where the VDW radius should be used can cause significant errors. Because of the existence of the Coulombic term, force fields can correctly predict the equilibrium contact distances between all types of atoms with only one set of r_0 values. It should be emphasized that the aim of this study is not to derive force field parameters, but to derive VDW and Coulombic radii for packing and surface area calculations.

Effect of Force Field in Structure Determination Software

In some cases, the electron density map obtained in an X-ray diffraction of a protein crystal does not yield atomic resolution for every atom and the structures have to be refined by using certain potential energy function (force field) as constraints.[33] Therefore, the force field in structure determination software has a certain effect on the contact distances between atoms, especially in low-resolution structures. It is natural to ask if it is still meaningful to derive VDW and Coulombic radii from protein structures. To answer this question, we note that there are accurate experimental data for the positions of at least some atoms. In addition, the potential energy functions are extensively tested against cases where accurate experimental information is available. Thus, the contact radii derived in this study reflect, for the most part, the real sizes of protein atoms. Nevertheless, one should always keep in mind that some inaccuracies in protein structures may exist. A related question is the accuracy of the coordinates of water molecules. The coordinates of the water molecules that form strong electrostatic interaction with polar protein atoms are likely to be more accurate than the coordinates of those that are in contact with apolar atoms. In fact, the VDW radius of water is probably overestimated.

Effects of Hydrogens and Protein Structure Environment on Derived Radii

Seventeen of the 25 atom types are united atoms. The effect of hydrogens is to increase the radius of the heavy atoms to which they are covalently bonded. However, the increase is not uniform in all directions. For triangle carbons in particular, such as CHar and CHim, other atoms can get closer to the carbon atom if they approach from the direction perpendicular to the plane of the triangle than from the direction along the C–H bond. Hence, the radius in the direction along the C–H bond is larger due to the presence of the hydrogen and the derived radius is an average of the two. The difference in radii between CHar and Car is probably mainly caused by the hydrogen atom on CHar.

The effect of the protein secondary- and tertiary-structure environment on the derived radii is hard to assess. In general, the effect is larger for mainchain atoms. That is, the contact distance distributions involving mainchain atoms may not due to optimal VDW or Coulombic contacts. Since the reference atoms used to measure the VDW and Coulombic radii are either on sidechains or is a small molecule (H2O), we estimate that the effect of secondary and tertiary structure is very small.

Implications and Applications

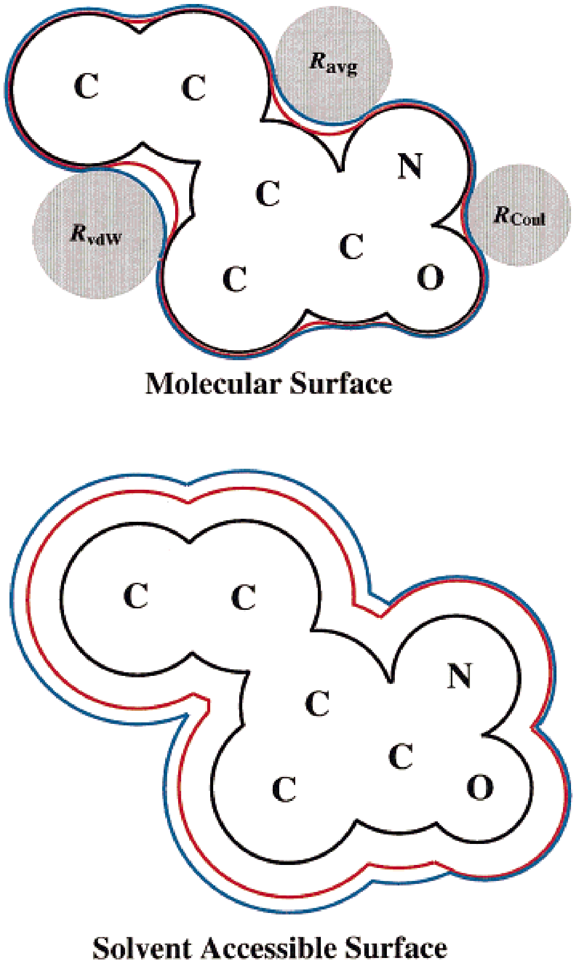
Molecular and solvent-accessible surfaces

Numerous algorithms have been developed and implemented to calculate the molecular and solvent-accessible surfaces of proteins.[7,34–40] Most calculations, however, have used only one radius for the probe and for each protein atom. This is correct if the probe is an apolar molecule, though one has to make sure that the VDW, not the Coulombic, radii of all atoms are used. The situation for the case of a polar probe, such as water, is complicated. For the protein, only one radius for each atom is needed, but the radius for polar atoms should be the Coulombic radius, while that for apolar atoms should be the VDW radius. As for the probing water, both the VDW and the Coulombic radii are needed. The VDW radius is needed for probing apolar regions of the protein surface and the Coulombic radius is needed for probing the polar regions. The dual-size nature of water presents no difficulties for the calculation of the solvent-accessible surfaces, as the center of the water is uniquely specified at all times. It does, however, present some technical difficulties in the calculation of the molecular surfaces as the shape of the probe is no longer spherical when it is in contact simultaneously with a polar and an apolar atom. Thus, approximations have to be made. One practical approximation would be to use the average of the VDW and the Coulombic radius of water to probe the polar–apolar boundary regions (Fig. 5). This way, it is assured that grooves and pockets at the polar–apolar boundary regions are predicted correctly and that no discontinuity of the surface is produced.

Figure 5 illustrates the change of the molecular and solvent-accessible surfaces for a hypothetical planar molecule when the Coulombic radius of a polar probe is used and when the probe radius is allowed to change according to the nature of its contacting neighbors. In this case, the portion of the surface associated with the polar region does not change since the probe radius remains the same. In the apolar region, the surface becomes smoother when the larger VDW radius of the probe is used. It is generally true that using the smaller Coulombic radius to probe the apolar regions may cause unreal grooves and pockets to be generated, while using the correct VDW radius gives smoother, more realistic surfaces.

Internal cavities of proteins

An internal cavity in a protein can be defined by the solvent-accessible surface[3] or by the molecular surface. In any case, since the apparent size of water is larger when interacting with apolar neighbors, an apolar cavity has to be bigger than a polar cavity in



Molecular Surface

Solvent Accessible Surface

Fig. 5. Schematic illustration of the change of the molecular and solvent-accessible surfaces for a hypothetical planar molecule when the Coulombic radius of a polar probe is used and when the probe radius is allowed to change according to the nature of its contacting neighbors. The VDW surface is drawn in black. The surface generated using the Coulombic radius of the probe is in red and that generated using three radii for the probe is in blue. For illustration purpose, in regions that are shared by two or all three surfaces, the latter surfaces are expanded slightly so that all three colors can be seen.

TABLE X. Comparison of Contact Distance Distributions Obtained from Protein Interfaces With Those from Protein Monomers (Values Listed Are: $D_{ij,\text{interface}} - D_{ij,\text{monomer}}$ in Å)

a: Contacts with apolar atoms

	CH2	CH2b	CH3	CHA
CA	0.01	0.06	−0.02	0.07
C	—	—	—	0.05
CH	—	−0.04	0.03	—
CH2	−0.04	−0.02	−0.02	0.01
CH2b	−0.02	0.05	−0.02	−0.01
CH2ch	−0.02	0.01	0.00	0.05
CH3	−0.02	−0.02	−0.03	−0.02
CHar	0.01	−0.01	−0.02	0.01
Car	—	—	0.02	—
CHim	0.09	0.18	0.12	0.10
N	—	0.08	0.09	0.14
NH2	—	0.12	0.04	0.06
NH2+	—	0.01	0.02	—
O	0.14	0.11	0.13	0.08
Oco	0.02	0.08	—	0.20
OH	0.00	0.07	0.01	0.15

b: Contacts between polar atoms

	O	Oco	Ocoo	OH
N	0.03	−0.01	−0.01	0.08
NH2	−0.03	−0.06	—	—
NH2+	—	0.02	−0.02	—
NH3	—	—	−0.02	—
OH	0.00	—	−0.01	—

order to accommodate the same number of water molecules.

Atom contacts across the protein–protein interface

The newly derived VDW and Coulombic radii are used to derived contact distance distributions across protein–protein interfaces using a list of interfaces given by Tsai et al.[25] Table X lists the values of $D_{ij,\text{interface}} - D_{ij,\text{monomer}}$, the difference between the peak position of a distribution derived from protein–protein interfaces and that from protein monomers. With the exception of CHim, the values for apolar–apolar distributions are all less than 0.1 Å, indicating that apolar–apolar contact distance distributions across protein–protein interfaces are very similar to those in protein monomers. The difference for apolar–polar distributions is slightly larger and all positive, indicating that the packing of polar atoms against apolar neighbors across the interface may be slightly looser than in protein monomers. It should be emphasized, however, that the distributions obtained from interfaces are not as accurate as those from protein monomers due to poor statistics. These results demonstrate that protein interfaces are well packed, in agreement with conclusions drawn by Walls and Sternberg[26] based on packing density calculations. It has been shown that the forces involved in protein folding and protein–protein association are similar.[41,42] In fact, structural analysis of protein–protein interfaces has revealed that global features of the architectural motifs present in protein monomers are also present across protein–protein interfaces.[42]

Surface complementarity at the protein–protein interface

Protein–protein complexes exhibit a remarkable surface complementarity at the interface.[15–18] The quality of the complementarity depends on how the surfaces are generated. If the surfaces are generated using the Coulombic radius of a polar probe, there might appear unreal voids at apolar regions of the interface. On the other hand, if the surface is generated using an apolar probe and the VDW radii are used for all atoms, there might appear unreal surface overlaps in polar regions. These unreal voids or overlaps can be reduced or eliminated and the surface complementarity can be further improved if two probes are used to generate the interfacial molecular surface: one probe represents a typical polar atom and the other represents a typical apolar atom. As for the radius, one could use the average Coulombic radius of oxygen and nitrogen atoms (1.43 Å) for the polar probe and the radius of a typical aliphatic carbon, say CH3 (1.92 Å), for the apolar probe. At the polar–apolar boundary, a probe with the average radius of the polar and the apolar probe, 1.68 Å, can be used.

Most automatic docking programs rely on the surface complementarity at the interface to place a ligand onto a receptor.[16,18] Accurate representation of the interfacial molecular surfaces helps to improve the quality of the results from docking. Fischer et al.[18] have realized the difference between the molecular surface in solution and that at the protein–protein interface and used a probing radius of 1.8 Å in order to mimic an average organic atom when generating molecular surfaces for docking. Interfaces are rich in polar interactions.[41] It is important that polar regions of the interfacial molecular surface are also generated correctly. In a computer docking experiment, tens of thousands of candidate complexes are usually generated. One of the important determining factors for discarding bad candidates is the extent of surface overlap. Accurate representation of the interfacial molecular surfaces will also increase the discriminating power of overlap checks.

Protein–water boundary

Where to place the protein–water boundary is somewhat subjective. The simplest way is to place the boundary on the molecular surface of the protein.[9] The fact that two radii have to be used for polar atoms changes the molecular surface, and hence the protein–water boundary. In addition, water molecules are further away from the hydrophobic portions of the boundary than they are from the polar portions. These observations may be of importance in the calculation of the electrostatic contribution to solvation using the continuous solvent model.[9,43]

Protein structure evaluation

Among the 1,169 structures used to derive the radii, 954 have resolutions equal or better than 2.5 Å and the worst resolution is 3.5 Å. Thus, the derived radii can be used to evaluate newly solved protein crystal structures. For this purpose, the minimum radius of each atom is needed. To obtain the minimum radii, we first numerically integrated the distance distributions to identify the minimum distance, D_{min}, below which the probability of finding two atoms in contact is less than 5%. The minimum radii, R_{min}, are calculated following the same procedure used to derive the VDW and Coulombic radii. That is, the R_{min} for four apolar atoms, CH2, CH2b, CH3, and CHar, and water are determined first, then these minimum radii are used as references to calculate the minimum radii of other atoms. The results are listed in Table XI. As the table shows, the minimum VDW and Coulombic radii for polar atoms are approximately the same. This is as expected because at short distances it is the repulsive component of the potential that determines the separation between atoms. When used to check protein structures, the probability of finding atoms whose separation is less than the sum of their minimum radii should be less than 5%. In deriving the minimum radii, atom contacts between residues close in sequence, $i - j < 5$, are not considered. Thus, the minimum radii cannot be used to generate the Ramachandran plots.[45] Our tests show that in order to predict more than 90% of the observed (ϕ, ψ) angles in known protein structures, the current minimum radii have to be scaled by a factor of about 0.9. For comparison, a set of effective VDW radii calibrated for reproducing experimentally observed (ϕ, ψ) angles[44] is listed in the last column of Table XI.

TABLE XI. The Minimum Radii of Atoms

	This work		Iijima et al.[a]
Symbol	$R_{\mathrm{min,vdW}}$ (Å)	$R_{\mathrm{min,Coul}}$ (Å)	$R_{\mathrm{min,vdW}}$ (Å)
CA	1.73		1.38
C	1.62		1.42
CH	1.82		1.38
CH2	1.68		1.38
CH2b	1.69		1.38
CH2ch	1.62		1.38
CH3	1.66		1.38
CHar	1.62		1.42
Car	1.62		1.42
CHim	1.50		1.42
Cco	1.67		1.42
Ccoo	1.65		1.42
SH	1.54		—
S	1.67		—
N	1.43	1.31	1.28
NH	1.35	1.36	1.28
NH+	1.38	1.31	1.28
NH2	1.31	1.31	1.28
NH2+	1.34	1.23	1.28
NH3+	1.22	1.23	1.28
O	1.26	1.31	1.18
Oco	1.20	1.27	1.18
Ocoo	1.24	1.23	1.18
OH	1.30	1.23	1.18
H2O	1.36	1.19	—

[a] Iijima et al.[44] Values are averages of two equivalent sets (rows 2 and 3 in their Table III). Six atoms were defined: sp^3 carbon, sp^2 carbon, main-chain O, amide N, aliphatic hydrogen and hydrogen on amide nitrogen.

CONCLUSION

We have shown that, by using an iterative procedure, it is possible to derive a self-consistent set of VDW and Coulombic radii for atoms in proteins based on contact distance distributions in known protein structures. The newly derived radii are in general agreement with those derived from crystal structures of small molecules. The inconsistencies[19] in the VDW radii for polar atoms in the literature are reconciled by the existence of two radii for each polar atom.

Comparison of contact distance distributions within protein monomers with those across protein–protein interfaces indicates that atom packing across protein–protein interfaces is similar to that in protein monomers, in agreement with conclusions drawn by Walls and Sternberg[26] based on packing density calculations. This demonstrates the similarity of the forces involved in protein folding and protein–protein association.[41,42]

Several important applications and implications of the newly derive VDW and Coulombic radii are discussed. It is shown that two radii for polar atoms are needed in the calculation of molecular and solvent-accessible surfaces of proteins. Most earlier calculations have used the Coulombic radius of water (1.4 Å) to generate both the polar and apolar portions of the two surfaces. The accuracy of these surfaces will improve in most cases if the probe radius is allowed to change depending on the nature of its contact neighbor. The molecular surface at the protein–protein interface, the interfacial molecular surface, should be generated using at least two kinds of probes, one representing a polar atom and the other an apolar atom. A more accurate representation of the interfacial molecular surfaces is crucial for protein–protein and protein–small-molecule recognition and interaction, as well as for automatic ligand–receptor docking.

ACKNOWLEDGMENTS

We thank Drs. J. Maizel, S.L. Lin, D. Xu, and C.J. Tsai for helpful discussions and critical reading of the manuscript. The research of R.N. has been sponsored by the National Cancer Institute, DHHS (1-CO-74102), with SAIC. The research of R.N. in Israel has been supported by grants from the Israel–U.S. Binational Science Foundation (95-00208), the Rekanati Foundation, the Israeli Ministry of Industry and Commerce (Maget), and Tel Aviv University.

REFERENCES

1. Pauling, L. "The Nature of the Chemical Bond," 3rd ed. Ithaca: Cornell University Press, 1960.
2. Bondi, A. "Physical Properties of Molecular Crystals, Liquids and Glasses." New York: John Wiley & Sons, Inc., 1968.
3. Williams, M.A., Goodfellow, J.M., Thornton, J.M. Buried waters and internal cavities in monomeric proteins. Prot. Sci. 3:1224–1235, 1994.
4. Gerstein, M., Tsai, J., Levitt, M. The volume of atoms on the protein surface: Calculated from simulation, using Voronoi polyhedra. J. Mol. Biol. 249:955–966, 1995.
5. Gerstein, M., Chothia, C. Packing at the protein–water interface. Proc. Natl. Acad. Sci. U.S.A. 93:10167–10172, 1996.
6. Kocher, J.P., Prevost, M., Wodak, S.J., Lee, B.K. Properties of the protein matrix revealed by the free energy of cavity formation. Structure 4:1517–1529, 1996.
7. Lee, B., Richards, F.M. The interpretation of protein structures: Estimation of static accessibility. J. Mol. Biol. 55:379–400, 1971.
8. Richards, F.M. Areas, volumes, packing and protein structure. Ann. Rev. Biophys. Bioeng. 6:151–176. 1977.
9. Sitkoff, D., Sharp, K.A., Honig, B. Accurate calculation of hydration free energies using macroscopic solvent models. J. Phys. Chem. 98:1978–1988, 1994.
10. Rose, G.D., Geselowitz, A.R., Lesser, G.J. Hydrophobicity of amino acids residues in globular proteins. Science 229:834–838, 1985.
11. Eisenberg, D., McLachlan, A.D. Solvation energy in protein folding and binding. Nature 319:199–203, 1986.
12. Wimly, W.C., Creamer, T.P., White, S.H. Solvation energies of amino acid sidechains and backbone in a family of host–guest pentapeptides. Biochemistry 35:5109–5124, 1996.
13. Chothia, C. Structural invariants in protein folding. Nature 254:304–308, 1975
14. Chothia, C. The nature of the accessible and buried surfaces in proteins. J. Mol. Biol. 105:1–12, 1976.
15. Connolly, M.L. Shape complementarity at the hemoglobin $\alpha_1\beta_1$ subunit interface. Biopolymers 25:1229–1247, 1986.
16. Shoichet, B.K., Kuntz, I.D. Protein docking and complementarity. J. Mol. Biol. 221:327–346. 1991.
17. Norel, R., Lin, S.L., Wolfson, H.J., Nussinov, R. Molecular surface complementarity at protein–protein interfaces: The critical role played by surface normals at well placed, sparse points in docking. J. Mol. Biol. 252:263–273, 1995.
18. Fischer, D., Lin, S.L., Wolfson, H.L., Nussinov, R. A geometry-based suite of molecular docking process. J. Mol. Biol. 248:459–477, 1995.
19. Rellick, L.M., Becktel, W.J. Comparison of van der Waals and semiempirical calculations of the molecular volumes of small molecules and proteins. Biopolymers 42:191–202, 1997.
20. Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., Karplus, M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. J. Comp. Chem. 4:187–217, 1983.
21. Bondi, A. Van der Waals volumes and radii. J. Phys. Chem. 68:441–451, 1964.
22. Richards, F.M. The interpretation of protein structures: Total volume, group volume distributions and packing density. J. Mol. Biol. 82:1–14, 1974.
23. Gavezzotti, A. The calculation of molecular volumes and the use of volume analysis in the investigation of structured media and of solid-state organic reactivity. J. Am. Chem. Soc. 105:5220–5225, 1983.
24. Hobohm, U., Sander, C. Enlarged representative set of protein structures. Prot. Sci. 3:522, 1994.
25. Tsai, C.J., Lin, S.L., Wolfson, H.J., Nussinov, R. A dataset of protein–protein interfaces generated with a sequence-order-independent comparison technique. J. Mol. Biol. 260:604–620, 1996.
26. Walls, P.H., Sternberg, M.J.E. New algorithm to model protein–protein recognition based on surface complementarity: Applications to antibody–antigen docking. J. Mol. Biol. 228:277–297, 1992.
27. Levitt, M. Aromatic rings act as hydrogen bond acceptors. J. Mol. Biol. 201:751–754, 1988.
28. Jorgensen, W.L., Tirado-Rives, J. The OPLS potential function for proteins: Energy minimizations for crystals of cyclic peptides and crambin. J. Amer. Chem. Soc. 110:1657–1666, 1988.
29. Reither, W.E. Ph.D. thesis, Cambridge, MA: Harvard University, 1985.

30. Weiner, S.J., Kollman, P.A., Nguyen, D.T., Case, D.A. An all-atom force field for simulations of proteins and nucleic acids. J. Comp. Chem. 7:230–252, 1986.
31. Levitt, M., Hirshberg, M., Sharon, R., Daggett, V. Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution. Comp. Phys. Comm. 91:215–231, 1995.
32. Hagler, A.T., Huler, E., Lifson, S. 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, 1974.
33. Brunger, A.T. X-PLOR, Verson 3.1. New Haven: Yale University Press, 1992.
34. Shrake, A., Rupley, J.A. Environment and exposure to solvent of proteins. J. Mol. Biol. 79:351–371.
35. Connolly, M.L. Analytical molecular surface calculation. J. Appl. Crystallogr. 16:548–558, 1983.
36. Connolly, M.L.. Solvent-accessible surfaces of proteins and nucleic acids. Science 221:709–713, 1983.
37. Pascual-Ahuir, J.L., Silla, E. GEPOL: An improved description of molecular surfaces. I. Building the spherical surface set. J. Comp. Chem. 11:1047–1060, 1990.
38. Silla, E., Tunon, I., Pascual-Ahuir, J.L. GEPOL: An improved description of molecular surfaces. II. Computing the molecular area and volume. J. Comp. Chem. 12:1077–1088, 1990.
39. Totrov, M., Abagyan, R. The contour-buildup algorithm to calculate the analytical molecular surface. J. Struct. Biol. 116:138–143, 1996.
40. Vorobjev, Y.N., Hermans, J. SIMS: Computation of a smooth invariant molecular surface. Biophys. J. 73:722–732, 1997.
41. Xu, D., Lin, S.L., Nussinov, R. Protein binding versus protein folding: The role of hydrophilic bridges in protein associations. J. Mol. Biol. 265:68–84, 1997.
42. Tsai, C.J., Xu, D., Nussinov, R. Structural motifs at protein–protein interfaces: Protein cores versus two-state and three-state model complexes. Prot. Sci. 6:1793–1805, 1997.
43. Bruccoleri, R.E., Novotny, J., Sharp, K.A., Davis, M.E. Finite difference Poisson-Baltzmann electrostatic calculations: Increased accuracy achieved by harmonic dielectric smoothing and charge antialiasing. J. Comp. Chem. 18:268–276, 1997.
44. Iijima, H., Dunbar, J.B., Marshall, G.R. Calibration of effective van der Waals atomic contact radii. Proteins 2:330–339, 1987.
45. Ramachandran, G.N., Ramakrishnan, C., Sasisekharan, V. Stereochemistry of polypeptide chain configurations. J. Mol. Biol. 7:95–99, 1963.