ANNUAL REVIEWS Further
Quick links to online content

Ann. Rev. Biophys. Bioeng. 1977. 6:151–76
Copyright © 1977 by Annual Reviews Inc. All rights reserved

Annu. Rev. Biophys. Bioeng. 1977.6:151-176. Downloaded from arjournals.annualreviews.org
by University of Texas - Austin on 09/28/09. For personal use only.

AREAS, VOLUMES, PACKING, AND PROTEIN STRUCTURE

❖9091

Frederic M. Richards
Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06520

INTRODUCTION

The protein folding problem remains today one of the most intriguing and fundamental questions in biochemistry. For peptide chains the covalent connectivity, the general nature of the interatomic forces, and, in many cases, the final average structures are known, yet the details of the folding process remain a mystery. Some of the experimental work and differing ideas on theory are described in recent reviews (1–4).

When the structure of double-stranded DNA was first proposed there was widespread excitement, not only because of the feeling of triumph accompanying a problem solved but also because of the apparent simplicity of the structure, one whose principal features could be characterized by a relatively small number of parameters. Great excitement also attended the first solution of the structure of a protein, myoglobin. However, this was accompanied not by rejoicing over simplicity but by wonder at the remarkable complexity of the answer. The structure did not immediately yield to simplifying descriptions. In the ensuing years a great deal of time and effort has been spent in trying to find useful descriptions.

Homopolymers, or heteropolymers composed of short repeating sequences, tend to form extended structures, i.e. polyamino acids, collagen, silk, etc. The globular proteins are invariably formed from heteropolymers with nonrepeating sequences. Either type of polymer may form α-helical or β structures, but the folding of such units into a compact particle is always associated with chains of nonrepeating sequence. Although globular protein structures can be described in terms of helices, extended chains, turns, and nonregular regions, this useful conformational description by itself has not yet led to an understanding of the structures. Some additional and different points of view are needed.

This review is concerned with the packing of groups of atoms in proteins and with the area of solvent-protein interfaces. As far as possible the emphasis is on simple

Euclidean geometry. The concepts of geometrical packing have played a substantial role in both the description of and theoretical attack on the properties of condensed phases in nonprotein systems. Of the various thermodynamic properties of a system, volume and area appear the easiest to grasp intuitively, and an attempt is made to extend these macroscopic concepts to the molecular level.

Static vs Dynamic Structures

The product of an X-ray crystallographic determination of the structure of a protein is an atomic model that invariably relies on the input of some non-crystallographic data such as the covalent connectivity of the peptide chain. The refinement and establishment of the reliability of such models, or static structures, is the subject of intensive efforts in many laboratories at this time. A review of these problems has been given by Jensen (5) and a general review of protein structure studies has been given by Matthews (6). The atomic model as a list of the three-dimensional coordinates of the atoms, without reliability estimates, is the normal starting point for others wishing to work with this data. In fact, the quality of the model varies from one part of a structure to another, but just the existence of the list of numbers tends to produce a false sense of confidence, particularly in users who are not crystallographers themselves.

Along with the mean atomic positions, the X-ray measurements can provide estimates of root mean square displacements of the atoms or of groups that arise either from disorder or, more interestingly, from actual motion of the groups involved. These parameters are a reflection of the dynamic structure of the molecules as they exist in the crystal lattice. The X-ray models referred to below are based on mean positions averaged over both space and time and represent the static not the dynamic structures.

A protein molecule is actually undergoing substantial fluctuations in the relative positions of its constituent atoms. Thus, the instantaneous accessible areas and volumes of individual atoms and of the structure as a whole (discussed below) vary with time. These fluctuations may be crucial to biological and chemical reactivity and to such physical probes as NMR or hydrogen exchange. As a result, without knowledge of at least the standard deviations, any interpretation of the mean atomic coordinates must be approached very cautiously. The increasing activity in the study of internal motion of proteins is referred to only very briefly at the end of this review. The concepts to be discussed have not yet been extended into a consideration of the time-based fluctuations of the dynamic structures.

The van der Waals Envelope

Because of the diffuse radial distribution of the electron density surrounding any atomic center, the apparent position of the surface of a molecule will depend on the technique used to examine it. For chemically bonded atoms the distribution is not spherically symmetric nor are the properties of such atoms isotropic. In spite of all this the use of the hard sphere model has a venerable history and an enviable record in explaining a variety of different observable properties. As applied specifically to proteins, the work of G.N. Ramachandran and his colleagues has provided much

Annu. Rev. Biophys. Bioeng. 1977.6:151-176. Downloaded from arjournals.annualreviews.org
by University of Texas - Austin on 09/28/09. For personal use only.

Annu. Rev. Biophys. Bioeng. 1977.6:151-176. Downloaded from arjournals.annualreviews.org
by University of Texas - Austin on 09/28/09. For personal use only.

of our present thinking about permissible peptide chain conformations (reviewed in 7). Different approaches using more realistic models, complex mathematics, and even quantum mechanical approximations have improved the details but have not altered the basic outline provided by the hard sphere approximation. The steepness of the repulsive term in the potential function for nonbonded interactions is responsible for the success of "hard" in the hard sphere.

Within limits the choice of van der Waals radii is arbitrary, with each author having his favorite list for the different atoms (see, for example, 7–10). The most appropriate values for successful predictions may vary with the problem. For covalently bonded atoms the spheres are normally truncated by a plane perpendicular to the interatomic bond. The exact position of this plane again is somewhat arbitrary but it is normally chosen to cut the bond into two segments proportional to the covalent radii of the atoms involved. The complex surface that results from linking a number of spheres together in this way is referred to as the van der Waals surface. This surface has a defined area and it encloses a defined volume. Although the construction is easy to visualize and is logically consistent, it should be recognized that no chemical procedure ever directly measures either this area or this volume.

In the lowest order approximation specific consideration of most hydrogen atoms is excluded. The heavy atoms, C, N, O, and S, are expanded into a series of groups with zero, one, two, or three hydrogen atoms attached as appropriate, and each one of these chemical groups is considered to be spherically symmetrical. The assigned radius for each group, centered at the heavy atom nucleus, attempts to account for the actual size contribution of the hydrogen atom(s). The groups are referred to as carbon, or nitrogen, atoms regardless of the number of hydrogens attached. This approximation is used in this review, and it is normally clear from the context what the real chemical group is.

AREA

Definitions

On the molecular scale any conceivable probe has dimensions comparable to the features of the surface being examined. Consider the cross-section of part of the surface of the hypothetical macromolecule shown in Figure 1. The trace of the van der Waals envelope of some of the atoms of the structure is shown. A spherical probe of radius R_1 is allowed to roll on the outside while maintaining contact with the van der Waals surface. It will never contact atoms 3, 9, or 11. Such atoms are considered not to be part of the surface of the molecule and are referred to as interior atoms. The question of how to define and quantitate the surface is a matter of convenience. One straightforward procedure is simply to use the continuous sheet defined by the locus of the center of the probe, the *accessible surface.* Another alternative would be to consider the *contact surface,* those parts of the molecular van der Waals surface that can actually be in contact with the surface of the probe. This would provide a series of disconnected patches. The *reentrant surface* is also a series of patches defined by the interior-facing part of the probe when it is simultaneously

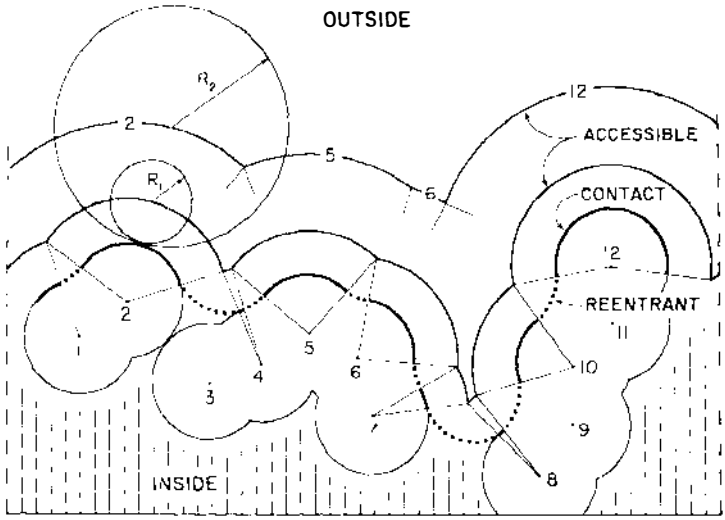


Figure 1 Schematic representation of possible molecular surface definitions. A section through part of the van der Waals envelope of a hypothetical protein is shown with the atom centers numbered.

in contact with more than one atom. Considered together the contact and reentrant surfaces represent a continuous sheet, which might be called the *molecular surface.*

By the nature of the geometrical construction there are no reentrant sections of the accessibility surface, i.e. viewed from the molecule each spherical segment is convex. This does entail a possible loss of information as the ratio of contact to reentrant surface may be a useful measure of molecular surface roughness. This can be seen qualitatively by inspecting Figure 1. The molecular surface also has the advantage that the area approaches a finite limiting value as the size of the probe increases. (It does have the curious property that, by the definitions given, the contact area approaches a very small value and the reentrant area approaches the limiting total as the probe radius goes to infinity.) In spite of these advantages, to my knowledge only the accessible surface has been calculated and reported in literature (11, 12). [A different approach based on the faces of Voronoi polyhedra (see below) has been initiated by Finney (personal communication).]

With any of the surface definitions the actual numbers derived will depend on the radius chosen for the probe. An example of the change that is produced by probe size is shown in Figure 1. In going from R_1 to R_2 the number of noncontact or interior atoms increases from three to eight. The accessible surface becomes much smoother (as does the molecular surface, not shown); there is only a slight dimple replacing the deep crevice revealed by the R_1 probe. The appearance of deeply convoluted features or actual holes in the interior of the protein becomes very sensitive to the choice of probe radius. The smaller the probe the larger the number

Annu. Rev. Biophys. Bioeng. 1977.6:151-176. Downloaded from arjournals.annualreviews.org
by University of Texas - Austin on 09/28/09. For personal use only.

Annu. Rev. Biophys. Bioeng. 1977.6:151-176. Downloaded from arjournals.annualreviews.org
by University of Texas - Austin on 09/28/09. For personal use only.

of features that will be revealed. About the smallest physically reasonable probe is a water molecule considered as a sphere of radius of 1.4 or 1.5 Å.

The accurate calculation of even the static surface area is a complex geometrical problem. Various approximate methods have been used whose accuracy is adequate for present purposes. The procedure reported by Lee & Richards (12) developed from a program used to graphically portray the van der Waals surface of a protein. Two-dimensional sections through the surface were drawn. By using pseudo-atoms, whose van der Waals radius was increased by the radius of the probe, a section of the accessible surface was obtained (Figure 2). The lengths of the individual arcs assigned to each atom on the full series of such planes were then multiplied by the section separation and added to get a value for the patch area in square angstroms. [The dimensionless quantity *accessibility,* defined as the percentage of the total solid angle around the atom center subtended by the accessible surface patch (12), in general is less useful than the actual surface area.] Shrake & Rupley (13) approx-

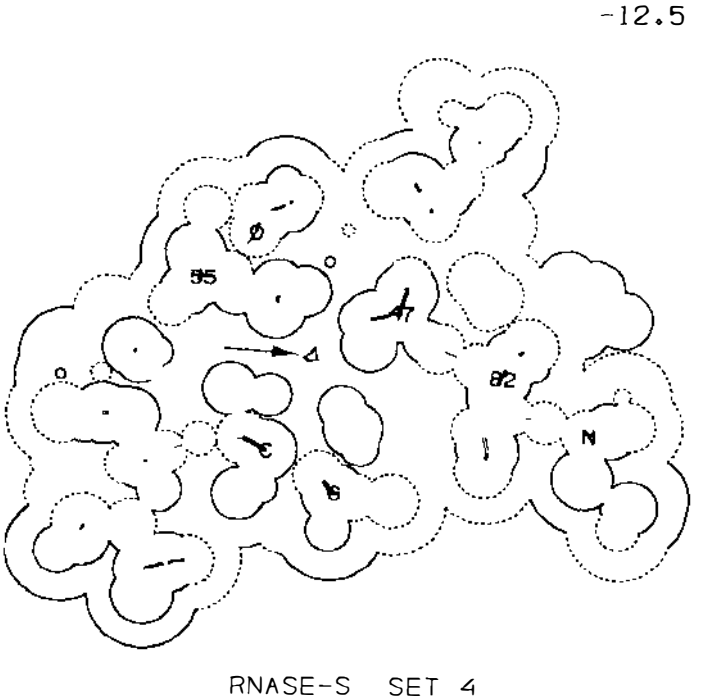


Figure 2 Superposition of sections through the van der Waals' envelope and the accessibility surface of ribonuclease S. The arrow indicates a cavity inside the molecule large enough to accommodate a solvent molecule with a radius of 1.4 Å, although it appears to be unfilled in the electron density map. In places the accessible surface is controlled by atoms above or below the section shown. The dashed outline is the surface of N or O atoms, the solid outline C or S atoms. [Reprinted with permission from (12).]

imated the same expanded surface around each atom by a series of 92 points distributed in space. Points in overlapping volumes were eliminated, and the number remaining provided an estimate of the surface area. Although computationally different, these procedures provide very similar numbers for the accessible areas. In a non-computational approach, balls of the appropriate size are packed as closely as possible around an atomic model of the substance in question (14, 15). The total count of balls in the final shell provides an estimate of the maximum number of simultaneously contacting solvent molecules. From the known radius of the ball the volume of this shell and its area can be approximated.

Annu. Rev. Biophys. Bioeng. 1977.6:151-176. Downloaded from arjournals.annualreviews.org
by University of Texas - Austin on 09/28/09. For personal use only.

Calculation of Areas in Proteins

TOTAL AREA The total accessible areas of 15 proteins have been calculated by Chothia (10) using a probe radius of 1.4 Å. The area of the smooth inertial ellipsoids, which approximate the overall shape of six of the proteins, have been calculated from the same coordinate lists (16). The ratio of the accessible area to the smooth area is about 1.7 ± 0.2. The principal source of the difference in these areas is the atomic level roughness of the molecular surface, which is not expected to vary much from one protein to another. Larger scale grooves and protuberances frequently associated with active sites also contribute to the increment in accessible area but to a lesser extent than the roughness. The significance of this roughness factor in connection with the protein solvent interface is discussed later.

Each residue in a fully extended chain has a maximum accessible area in a conformation where steric hindrance of the probe is minimal. The estimate of this area is only slightly dependent on side chain conformation except possibly for the most flexible groups, lysine and arginine. The mean area expressed as $Å^2$/dalton is 1.46, with a maximum range of 1.36 to 1.69 (except for cystine at 0.9). Averaging over the actual amino acid composition of 15 different proteins, the mean is found to be 1.45 $Å^2$/dalton. The total area of a fully extended chain can be expressed as a linear function of molecular weight within ±3%.

Lee & Richards (12) noted that, during the folding of a fully extended chain to give the compact native structure for proteins with a molecular weight of 15,000, the area decreases to about one-third of its maximum value. For constant shape, one would expect the area to vary as the two-thirds power of the molecular weight. From Chothia's area data (10), the exponent from a least squares fit of log A vs log M is 0.70 ± 0.10. Useful approximate equations are listed in Table 1.

AREAS BY ATOM TYPE One can ascribe other characteristics such as polarity to each area patch, carbon and sulfur considered as nonpolar and nitrogen and oxygen as polar. Of the accessible areas of the native structures, roughly half represents polar atoms and half nonpolar atoms (12, 13). Thus the "grease" is by no means all "buried." In the folding process there are roughly equivalent decreases in the accessibility of both the polar and nonpolar groups. The relevant forces and the final structure require more careful definition than is implied by the common feeling that inside equals nonpolar and outside equals polar.

Table 1 Approximate equations relating volume and surface areas to molecular weight

Relations derived from X-ray coordinate data on 12 proteins	Standard error
1.[a] Total area of extended chain = $A_t = 1.45\ M$	417 Å^2
2.[a] Total accessible surface area = $A_s = 11.12\ M^{2/3}$	401 Å^2
3. Ratio of folded to extended area = $A_s/A_t = 7.67\ M^{-1/3}$	
4. Total buried surface area = $A_b = A_t - A_s$	
5.[a] Packing volume[b] = $V_p = 1.27\ M$	401 Å^3
6.[a] Surface to volume ratio = $A_s/V_p = 8.77\ M^{-1/3}$	0.015 Å^{-1}
Equivalent spheres	
7. Area of sphere of volume $V_p = A_{s'} = 5.67\ M^{2/3}$	
8.[a] Radius from packing volume = $R_v = 0.672\ M^{1/3}$	
9.[a] Radius from accessible area = $R_s = 0.940\ M^{1/3}$	
Surface roughness indices	
10.[a] From equivalent sphere: $R_s/R_v = 1.4$	
11. From equivalent sphere: $A_s/A_{s'} = 2.0$	
12. From inertial ellipsoid: $A_s/A_{s''} = 1.7$[c]	

[a] Taken from Teller (17; see also 120). M, Molecular weight.

[b] V_p is derived from composition data and the mean internal residue volumes given in Table 2. Note that in units of cubic centimeters per gram the factor 1.27 becomes 0.764, the specific volume. This figure should not be used as an estimate of partial specific volume in hydrodynamic measurements. No correction for electrostriction by charged residues is included and uncertainty surrounds the geometrical definition of the protein solvent interface.

[c] The shape of the inertial ellipsoid, and thus its area A_s'', varies for each protein. No general relation to molecular weight can be given. The ratio A_s/A_s'', however, is found to be approximately constant.

Atoms and groups carrying formal charges are almost invariably accessible to the solvent. Even here the accessible area is substantially lower in the native protein (40–60%) than in the extended chain. This fact must affect the mean dielectric constant seen by these groups.

Polar but uncharged atoms show much less bias between internal and external positions. The division of individual residues by type into buried and exposed in most instances does not appear to be correct. Polar groups that are removed from contact with water almost invariably are found to form hydrogen bonds with an appropriate partner within the protein. Stated in another way, a given hydrogen bonded donor-acceptor pair can be treated as a nonpolar unit, can be removed from the solvent and packed inside, and can contribute to the hydrophobic bonding in the same sense as a carbon chain or a benzene ring.

AREAS AND TRANSFER-FREE ENERGY After Kauzmann's original discussion of the hydrophobic bond (18), Nozaki & Tanford (19) derived values for the free energy of transfer of certain amino acid side chains from organic solvents to water from solubility data. Chothia (20) has plotted these values against the maximum

Annu. Rev. Biophys. Bioeng. 1977.6:151-176. Downloaded from arjournals.annualreviews.org
by University of Texas - Austin on 09/28/09. For personal use only.

accessible surface area for each side chain. When grouped into nonpolar and polar categories, the values fall onto two straight lines (Figure 3). The difference between the two groups is roughly constant and related to the hydrogen-bonding characteristics of the single polar group. The clear implication is that each unit decrease in interface area between a solute molecule and water is associated with a constant decrement in free energy, which is independent of the detailed structure of the solute. This approach has been developed also by Hermann (11), Harris et al (14), and Reynolds et al (15). For a detailed consideration of the structure of the solvent in this process, consult the publications of Sinanoğlu (15a, 15b). Whether this is a valid representation of everything that is implied by the term hydrophobic bond is an arguable point. [Klapper's excellent review of this whole problem should be consulted (21), along with Tanford's monograph (22).] The observation, however, does provide a very useful approach to handling a major component of the energy of the solvent-protein interaction. The calculated area change in any given reaction is multiplied by the appropriate factor to get the solvent-squeezing contribution to

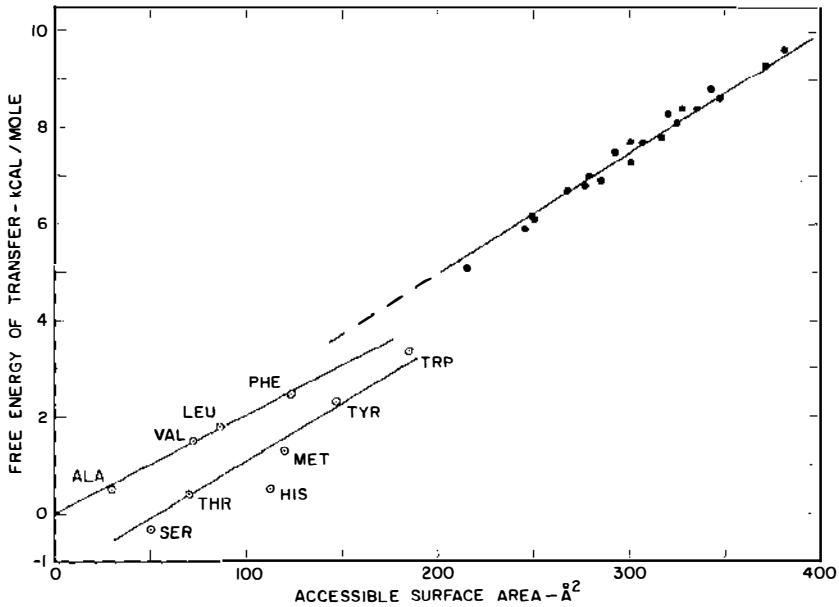


Figure 3 Hydrophobicity, expressed as the free energy of transfer from 100% organic solvent to water plotted as a function of molecular surface area. The black dots and line shown in the upper right, extrapolating back to the origin with a slope of 25 cal/Å², are adapted from Figure 1, reference 15, which in turn quotes the data of reference (11) on a series of hydrocarbons. The data shown in the lower left refer to the side chains of the indicated amino acids [adapted from Figure 1, reference 2]. The line passing through Ala, Val, Leu, Phe has a slope of 22 cal/Å².

Annu. Rev. Biophys. Bioeng. 1977.6:151-176. Downloaded from arjournals.annualreviews.org
by University of Texas - Austin on 09/28/09. For personal use only.

Annu. Rev. Biophys. Bioeng. 1977.6:151-176. Downloaded from arjournals.annualreviews.org
by University of Texas - Austin on 09/28/09. For personal use only.

the total free energy change. The very large area changes occurring during the protein folding process made this term the major contribution to the overall energy change, as many authors have pointed out. This component of the total energy will always push towards structures with minimum accessible areas. Chain entropy effects will favor larger molecular volumes and thus larger areas. The requirements of specific polar interactions, in general, will be unrelated in any specific way to overall molecular volume or area. The actual structure represents, of course, a balance of the various components of the total energy.

The important point in Figure 3 is the linear relation between free energy and area. The appropriate value of the proportionality constant is less certain. The data shown yield 24–26 cal/Å^2/mole (see also 15) and reflect the partitioning of the solute between two liquid phases. A liquid hydrocarbon is probably not a good model for the interior of a protein, although after Kauzmann's famous article it has been widely considered such. One can use the solubilities of the crystalline amino acids in water on a plot equivalent to Figure 3. Omitting cystine and tyrosine, there is an acceptable correlation with a slope of 13–15 cal/Å^2/mole. The 11-cal/Å^2 difference in part is related to the entropy difference between the amino acid in the crystal and in the organic solvent. However, a crystal of a single amino acid also is not a very good model for the protein interior since a given side chain is not commonly in contact with other residues of the same type, and there is more motion in a protein molecule than in most simple organic crystals. At the moment there is, in fact, no good model. However, 20 cal/Å^2/mole has a fair chance of being an appropriate value for rough estimates of the squeezing effect. This number will no doubt be refined in the future.

VOLUME

Definitions

Packing efficiency is frequently discussed in terms of *packing density* (8), a dimensionless ratio of the minimum (or actual) volume of an object to the volume of space that it occupies. For crystals of many organic molecules the packing density falls in a rather narrow range of 0.70 to 0.78, with the mean close to 0.74, the theoretical value for close packed spheres. This narrow range is observed in spite of the wide variety of molecular sizes and shapes that are being packed in the different crystals. The *minimum volume* of a molecule is taken as the volume of space enclosed by the van der Waals surface (23–25).

For irregular objects the occupied volume is not always intuitively obvious. A theorem from geometry due to Voronoi (26, 27) now becomes particularly useful. Consider any extended set of points in space that are restricted in no way except that each has a fixed position identified by three known space coordinates. If planes are now drawn as perpendicular bisectors of the vectors between each pair of points, these planes will intersect to form a uniquely defined, irregular polyhedron around each of the points. The volume of each polyhedron is easily calculated and is assigned to the enclosed point. The polyhedra do not overlap nor are there any unaccounted for voids. Thus, the sum of the polyhedral volumes is exactly equal

to the total space occupied by the points. If the set of points is finite, the statements made above are true only for those members of the set that are surrounded by other points. The points on the surface of the finite set will have an inappropriate or undefined volume.

Both the utility and problems of this form of volume partition are immediately evident. All substances can be represented as a collection of points taken to be the centers of the atoms of which the material is composed. Those that are interior in the sense just discussed will occupy a uniquely defined volume. However, in general, the atoms are not identical and differ intrinsically in size. Apart from hydrogen atoms the size differences, in fact, are not very large, especially if one considers only carbon, nitrogen, oxygen, sulfur, and phosphorus. The Voronoi polyhedra will tend to assign too much volume to the smaller atoms and too little to the larger ones, but for comparative purposes this is not serious and the systematic error is much smaller than the deviations within a given class of atoms, which will be of interest here.

For crystals of substances where the total contents and exact atomic positions are known, the surface problem does not arise. The unit cells of ordinary crystals are automatically surrounded by an infinite set of points. Protein crystals as a class unfortunately represent an exception. The contents of the unit cell are not accurately known because the exact nature and amount of the nonprotein components are uncertain. The positions of the atoms in the solvent part of the cell for the most part are not well defined, if seen at all. Thus the volume calculations on the protein atoms suffer from the surface problem of a finite set. However, there will be a number of atoms whose volume is totally determined by the position of other protein atoms.

Calculated Volumes in Proteins

After discussions with John Finney concerning this general approach and his earlier work with Bernal on the structure of liquids (28–32), Richards (33) applied these procedures to the reported X-ray structures of ribonuclease S and lysozyme. This program was used by Chothia (10) on a larger set of proteins, and the method was examined in detail and improved by Finney (34).

VOLUMES OF INDIVIDUAL ATOM TYPES Restrict consideration for the moment to those atoms classified as internal. Finney found that for a given atom type the standard deviations of most of the distributions are about 10–15% of the mean atom volume. He noted that this value is larger than the value found in glasses formed from identical atoms (~5%), and this may be caused in part by the differences in size of the protein atoms and in part by the covalent structure and directional bonding that introduce factors other than hard sphere packing. Even so, the distributions are sufficiently narrow that the mean volumes could be useful as test criteria for evaluating proposed tertiary structures in the protein folding problem. A trial structure for an essentially solvent-free compact molecule would be considered unacceptable if the mean atom volumes were significantly outside the ranges indicated.

Annu. Rev. Biophys. Bioeng. 1977.6:151-176. Downloaded from arjournals.annualreviews.org
by University of Texas - Austin on 09/28/09. For personal use only.

Annu. Rev. Biophys. Bioeng. 1977.6:151-176. Downloaded from arjournals.annualreviews.org
by University of Texas - Austin on 09/28/09. For personal use only.

MEAN RESIDUE VOLUMES The volumes of individual atoms may not be the most appropriate summary of the data since the atoms cannot be moved, and thus packed, independently. Finney (34) summed the component atom values to get individual residue volumes for ribonuclease S, and Chothia (10) did the same for the interior residues of nine different proteins (Table 2). The standard deviations show improvement over those for single atoms with values 10% or less of the mean residue volume. A reference volume for each residue was calculated by taking the measured molecular volume of the amino acid from crystallographic data on the pure amino acids and subtracting a constant volume for the one oxygen and two hydrogen atoms by which the free amino acid differs from the residue. Chothia's estimate for this constant volume was 11.1 Å^3. There is a remarkable correspondence between these reference volumes in column 7 and the mean occupied volumes given in column 4 (see Table 2). The sum of the reference volumes and the sum of the corresponding mean occupied volumes differ by less than 0.4%. With only two exceptions the differences for individual residue types differ by less than 6%. The conclusion appears inescapable: the mean packing density of interior protein atoms is essentially identical with that found in crystals of small organic molecules.

When this observation is combined with the fact that proteins contain few if any interior water molecules, it can be seen that the packing is remarkably effective, even

Table 2 Volume occupied by residues in the interior of nine proteins[a](10)

Residue	Total no. in proteins	No. buried[b]	Average volume (V_R) of buried residues (Å^3)	Standard deviation of V_R Å^3	Standard deviation of V_R % V_R	Residue crystal volume equal to amino acid volume less than 11.1 Å^3[c]
Val	163	91	141.7	8.4	5.9	143.4
Ala	186	71	91.5	6.7	7.3	96.6
Ile	106	69	168.8	9.8	5.8	169.7
Gly	160	60	66.4	4.7	7.1	66.5
Leu	138	57	167.9	10.2	6.1	—
Ser	190	46	99.1	7.4	7.5	102.2
Thr	128	32	122.1	6.7	5.5	124.3
Phe	60	29	203.4	10.3	5.1	—
Asp	117	17	124.5	7.7	6.2	122.0
Cys	34	16	105.6	6.0	5.7	108.7
Pro	67	16	129.3	7.3	5.6	124.4
Met	28	14	170.8	8.9	5.2	176.1
Tyr	98	13	203.6	9.6	4.7	201.7
Glu	65	13	155.1	11.4	7.4	143.9
Asn	116	12	135.2	10.1	7.5	—
Trp	39	9	237.6	13.6	5.3	—
His	43	8	167.3	7.4	4.4	166.3
Lys	119	5	171.3	6.8	4.0	-
Gln	80	5	161.1	13.0	8.1	148.0
Cyh	10	4	117.7	4.9	4.2	123.1
Arg	63	0	—	—		—

[a]The nine proteins are calcium binding protein, ribonuclease S, lysozyme, papain, α-chymotrypsin, subtilisin, carboxypeptidase, thermolysin, and lactate dehydrogenase.

[b]A residue is defined as buried if 5% or less of its potential accessible surface area is available to solvent contact.

[c]11.1 Å is the volume lost by an amino acid on becoming a residue. The value used here was found by comparing the crystal volumes of glycerine and glycylglcine. 11.0 Å^3, determined by solution studies. No accurate unit cell dimensions were found for Leu, Phe, Asn, Trp, and Lys.

with all of the atoms covalently connected through the polymer chain. The interior of a protein molecule is not an oil drop but resembles rather a molecular crystal. This aspect of the protein interior was explicitly described by Klapper (36) before the present detailed statistics were available. Quite independently, the significance of volume and packing effects in hemoglobin and myoglobin was estimated and discussed by Lim & Ptitsyn (37, 38). The constancy of the interior volume over a number of different species was noted. Wyckoff has discussed the paired replacements in ribonucleases from different organisms that apparently maintain a constant space-filling condition (39). From energy calculations in nucleic acids, Motherwell and associates (40, 41) have concluded that packing and van der Waals interactions are the principal factors involved in these structures as well.

It is likely that for a given peptide chain any solvent-free structure other than the observed one would have a lower packing density. It is also unlikely that any structure could be proposed that would have a higher average packing density. The latter would require the chain to pack more efficiently than independent packing units of comparable covalent structure. Thus the final structure assumed by a folding protein chain probably represents a thermodynamic minimum and not a kinetic cul-de-sac on the side of some steep global gradient. This statement is certainly true with regard to packing criteria. If, in fact, there are a number of possible structures with the same packing density, then the choice must be made on other grounds, and the kinetic accessibility of the various structures could still be an issue. The final choice would then be made on the basis of energy terms not directly related to hard sphere packing geometry.

DEVIATIONS FROM THE MEAN VOLUMES The discussion so far has centered on the mean volumes taken over a large number of groups. However very substantial deviations from these mean values occur throughout a protein molecule when the averages are taken over smaller volumes containing relatively few atoms. When calculated in adjacent cubes with an edge length of 5.6 Å (usually five to ten atoms), the packing density in ribonuclease S or lysozyme varies from <0.6 to >0.85 in different parts of the structure (33). Shifting one or two atoms between adjacent cubes does not significantly alter the observations. The packing differences are real and may be connected with enzyme function. Packing defects could reflect structural flexibility and actually control permitted motions during substrate binding and catalysis. Thus the active site region in ribonuclease S has a fairly low packing density and is surrounded by a region of high density, suggesting a flexible, horseshoe-shaped clamp.

Kauzmann and colleagues (42) have made mass density, rather than packing density, calculations on a number of proteins and also find a non-uniformity in the structures characterized grossly as a low-density core and high-density periphery. More recently Schultz (43) has expanded these calculations on a much more detailed level, concentrating on the packing above and below peptide groups in extended chains by using only regions containing interior atoms. Substantial variations again were found which became larger as the sampled volume became smaller, as might be expected. She was concentrating on side chain packing where van der Waal's

Annu. Rev. Biophys. Bioeng. 1977.6:151-176. Downloaded from arjournals.annualreviews.org
by University of Texas - Austin on 09/28/09. For personal use only.

interactions should predominate. The very interesting conclusion was reached: where backbone hydrogen bonding was regular and fully developed, the side chain packing was worse (i.e. the packing density was lower) than in regions of less regular polar interactions, such as bends where the side chain packing appeared to be better. This would agree with Finney's general conclusion (34) that there will usually be competition between close packing and directed bonding effects in real structures containing a variety of chemical groups.

The Surface Problem

The Voronoi volumes of surface atoms are only defined if there are yet other atoms external to them. The only candidates are solvent molecules. These are not clearly located for the most part, and even in the most detailed crystal structure determinations much of the solvent will represent time-fluctuating assemblies that are amenable only to statistical treatment. Attempts have been made by Richards (33) and by Finney (34) to circumvent this difficulty by specifying hypothetical arrays of solvent molecules around the protein. The sole function of these arrays is to provide calculatable volumes for the protein atoms, and the volumes depend on the details of the array chosen. Richards assumed a simple cubic lattice into which the protein was inserted. All cells next to the protein were filled with solvent and the vectors from the protein atoms to the center of these cells were used in the Voronoi construction. Depending on the details of how these cells were used, the calculated protein atom volumes were either equal to or greater than those of the interior atoms.

Finney also used various treatments of a hypothetical solvent shell. One gave minimal protein atom volumes and employed many more solvent atoms than could actually surround any atom at one instant. Another used a shell of non-overlapping solvent molecules and lead to higher, but perhaps more realistic, volume estimates. Considering the approximations of Finney and Richards together, the packing densities of the surface atoms can be made equal to, slightly greater than, or slightly less than that of the internal protein atoms. In the absence of better factual data on the real solvent interface, the detailed packing of the surface atoms can only be considered uncertain at this time. The best rule of thumb at the moment seems to be that all protein atoms, surface and interior, have the same mean packing density, which closely approximates that of the same groups in molecular crystals.

Kauzmann et al (42) have arrived at the conclusion of denser packing at the protein surface than in the interior based on mass density calculations. The surface problem was not adequately handled in this study. Direct comparison of these calculations with the Voronoi procedure is involved and it is not clear at this time why there is such a discrepancy in the conclusions.

THE PROTEIN-SOLVENT INTERFACE

The hydration of macromolecules is the subject of constant investigation and conjecture [see reviews by Kuntz and others (43a–45)]. Observations with many different techniques can be conveniently described by referring to three distinct types of water. Type I is bulk water with all of its known characteristics, especially a

Annu. Rev. Biophys. Bioeng. 1977.6:151-176. Downloaded from arjournals.annualreviews.org
by University of Texas - Austin on 09/28/09. For personal use only.

rotational relaxation time of the order of 10^{-11} sec. Type II is bound water with a relaxation time of about 10^{-9} sec and an altered freezing point brought about by interaction with the macromolecular surface. (Most techniques give estimates of 0.3 to 0.5 g of H_2O per g of protein for water of this type, but it is not always clear that the different techniques are sensitive to the same group of solvent molecules, even though the estimates of mass may be similar.) Type III is irrotationally bound water with the relaxation time characteristic of the macromolecule and generally in the range 10^{-5} to 10^{-7} sec. (The amount of this material is usually quite small, 0.05 g/g or less.) NMR data on water signals from protein crystals obtained by Bryant (46, 47) also fall roughly into these three classes. Considerable solvent appears to be bulk water even though it is contained in channels in the crystal lattice, which are sufficiently small that most of the water is in direct contact with a protein surface (48). The same conclusion was drawn from earlier studies of the chemical properties of proteins in the crystal lattice (49, 50) as well as the direct measure of diffusion rates of small solutes in these water-filled channels (51). A detailed analysis of the solvent in protein crystals has recently been made by Scanlon & Eisenberg (52). The comments in this section are restricted to the packing behavior of the water layer immediately adjacent to the protein surface.

Annu. Rev. Biophys. Bioeng. 1977.6:151-176. Downloaded from arjournals.annualreviews.org
by University of Texas - Austin on 09/28/09. For personal use only.

Non-Ideal Mixing of Hard Spheres

In a two-component ideal solution, partial specific volumes are constant and the specific volume is a linear function of the weight fraction composition, terminating at the specific volumes of the pure components. Even in hypothetical hard sphere fluids, volume additivity requires mixing spheres of identical size. An analysis of this problem for aqueous solutions of small molecules has been given by Assarsson & Eirich (53) and discussed by Bøje & Hvidt (54). Proteins in water represent an extreme example.

In dilute solution it is commonly found that the apparent specific volume of a protein is constant over the concentration range of measurement, generally 1–5 wt% protein. All the water of the solution behaves as though it had the same partial specific volume as the pure solvent.

The other end of the composition range can be examined in protein crystals. Such crystals can be dried slowly in stages in an atmosphere of controlled relative humidity. Paired composition and density measurements can be plotted as shown in Figure 4, taken from some early work of Low & Richards (55). Within the accuracy of the data the relationship is linear and represents an extension of the dilute solution measurements. Extrapolation to 0% water gives the normally accepted value for the partial specific volume of the protein in solution. Extrapolation to 0% protein gives the partial specific volume of pure water.

The unit cell of the crystal is shrinking constantly in the region 50–20 wt% water. Abruptly at 10–15 wt% water the cell shrinkage ceases and no further volume change occurs during removal of the rest of the water. In this region the partial specific volume of the protein appears to be much higher and $\bar{v}_{H_2O} = 0$. The surfaces of the protein molecules are now firmly in contact. Their approximate ellipsoidal shape and large size leaves voids of very substantial dimensions on the scale of a

Annu. Rev. Biophys. Bioeng. 1977.6:151-176. Downloaded from arjournals.annualreviews.org
by University of Texas - Austin on 09/28/09. For personal use only.

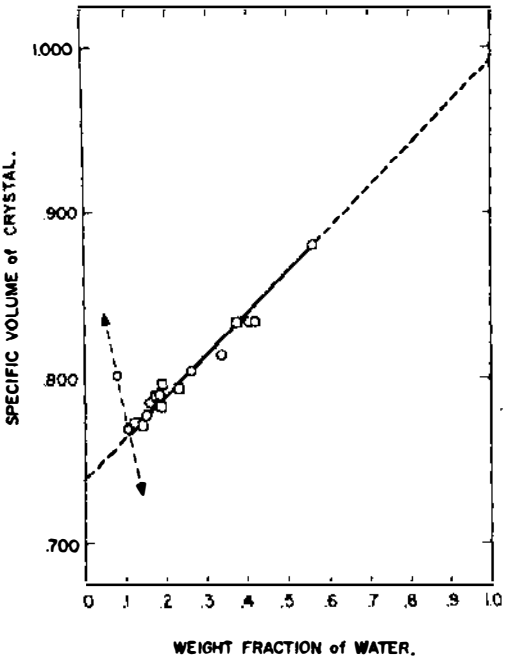


Figure 4 Specific volume of crystals of the dimer form of human serum albumin as a function of water content. ○, Determined during drying; □, determined during rehydration. Extrapolated intercepts of solid line yield $\bar{v}_p = 0.734$ cc g^{-1} and $\bar{v}_{H_2O} = 0.995$ cc g^{-1}. [Adapted with permission from (55).]

water molecule. Emptying or filling these voids produces no macroscopic volume change regardless of the actual volume of the water molecules involved. The system thus behaves just as one would expect for the mixing of rigid particles of very different size. Nothing can be said from such data about the actual volume of the water in the interprotein voids. However, once these voids are filled, additional water starts separating the protein molecules from each other. At this stage the water behaves as though its mean specific volume is identical to that of bulk water even though each molecule is in direct and intimate contact with at least one protein surface.

For a protein of 15,000 molecular weight, a monolayer of water around its surface will require 600–1000 molecules (33, 34). This layer will have about the same mass as the protein itself. At least as a statistical average, it appears that the interface between a protein and the solvent is characterized by an abrupt change from the packing density of the protein to the packing density of bulk water.

The nearest neighbor oxygen-oxygen distance in ice is about 2.8 Å, leading to the common characterization of a water molecule as a sphere of radius 1.4 Å. Tetrahedral bonding of such a molecule in the hexagonal ice I lattice leads to the low packing density of 0.34 or about 0.38 for liquid water after correction for the

Annu. Rev. Biophys. Bioeng. 1977.6:151-176. Downloaded from arjournals.annualreviews.org
by University of Texas - Austin on 09/28/09. For personal use only.

contraction on melting. The molecules in ice or water are not in van der Waals contact. The strong hydrogen bonds bring a neighboring oxygen atom well inside the van der Waals surface of the H atom on a neighboring molecule. Thus a sphere of 1.4-Å radius (volume, 11.5 Å^3) is a substantial underestimate of the volume of a water molecule and leads to an unrealistically low estimate of packing density. From the dimensions of the water molecule (r_{vdwH} = 1.20 Å, r_{vdwO} = 1.40 Å, $r_{covalent\ H}$ = 0.30 Å, $r_{covalent\ O}$ = 0.66 Å), the van der Waals volume as normally calculated would be 17.4 Å^3 giving a 1.6-Å-radius equivalent sphere. With these numbers the packing density of liquid water would be 0.58, substantially higher than 0.38 but still low in comparison to molecular crystals. [As pointed out by Finney (34, 56), water is a liquid whose structure is not controlled by packing considerations.]

At the protein surface some water will be hydrogen bonded, with an effective radius of 1.4 Å, to the accessible polar groups. Where these groups have formal charges electrostriction of the solvent occurs. These solvation shells have a much smaller volume than bulk water, but they do not represent a large fraction of the total surface layer. (The latter statement may not be true for highly charged macromolecules such as nucleic acids.) About half the surface of an average protein is nonpolar yet in contact with water. In this area the water should pack against the protein in the hard sphere sense with an effective radius of about 1.6 Å. This layer will have a slightly larger specific volume than bulk water. The volume excess will be numerically very close to the predicted electrostrictive decrement and thus the mean volume for the whole monolayer would be expected to be very close to that of bulk water as observed.

Volume Change on Denaturation

Tanford (57, 58) has given an extensive discussion of the general process of protein denaturation from the native protein to the random coil. The question of how much and how many intermediate structures there may be has been the subject of considerable argument. [See the excellent paper by Tsong et al (59) on this complex issue.]

The overall volume change during denaturation has been particularly puzzling. There is general agreement that the measured apparent volume change is always very small (60–66). For proteins in the molecular-weight range of 10,000 to 40,000 the number is usually in the range –30 to –300 cc/mole, which corresponds to a change in specific volume of 0.005 cc/g or less.

In considering unit processes, the largest volume changes are normally found in the appearance or disappearance of charged groups (67). In proteins the charged groups are almost invariably exposed to the solvent. The added exposure occurring on denaturation may permit the development of a more complete hydration shell, but most of the electrostrictive effect is probably present around the groups in the native structure. The added contraction from this cause during denaturation is expected to be small. In his original discussion of the hydrophobic bond, Kauzmann (18) pointed out the large volume changes that occur on transferring small nonpolar molecules from an organic solvent to water (see also 68). Since major changes in accessibility of nonpolar groups occur during denaturation, one was led to suspect

that large volume changes should be observed during this reaction. Indeed the absence of such effects has been used to call into question the concept of hydrophobic bonding as an important component of protein structure (61, 62).

This problem may have been resolved by Bφje & Hvidt (69–71). They questioned the choice of reference model compounds, suggesting that simple hydrocarbons are not appropriate. Measurements on a series of alcohols, ketones, amides, and ethers, substances more closely related to the components of a protein, have shown that the volume effects are very different than for pure hydrocarbons. Solutions of many of these substances were studied over the complete composition range. The results show marked concentration effects reflecting the non-ideal mixing of molecules of different size.

Hvidt has noted that dilute solution is never an appropriate reference condition for comparisons between small molecules and polymer segments. The reason can be seen in some recent calculations by Fixman on polymethylene chains (72; personal communication). Under conditions in which the interaction energy of the chain segments with themselves and with solvent molecules is the same, the concentration of other chain segments in the neighborhood of a particular interior segment is of the order of 7 to 8 M, or weight fraction of >0.1. This number would increase somewhat in a poor solvent, where the chain contracts, and decrease somewhat in a good solvent, where expansion occurs, but it would never approach dilute solution. Hvidt concluded that the expected volume change on denaturation due to uncharged species should be small and perhaps even positive. This component combined with the possible small negative electrostrictive contribution would surely lead to the observed small net volume changes.

Slip at the Protein Surface

The common hydration estimate of 0.3 g of H_2O per g of protein corresponds to about 250 molecules of water for a protein of 15,000 molecular weight. The roughness of the molecular surface discussed earlier requires a minimum of 600 water molecules to form a complete monolayer around a protein of this size. All techniques detect hydration as water whose properties are significantly different than bulk water. Conversely water whose properties are not so affected is considered to be bulk solvent. Over half of the water in contact with the protein surface thus is indistinguishable from the pure solvent. For hydrodynamic effects in such regions, the slip surface thus is between the water monolayer and the protein and not outside of the first monolayer, as is commonly assumed in the derivation of the Stokes frictional coefficient.

The structure of the solvent in protein crystals is now coming under intense investigation as part of the X-ray refinement problem. Recent observations on lysozyme are described by Moult et al (73), and on rubredoxin by Watenpaugh et al (74). Finney summarized the available data on these and other proteins (56). The details are not in but a general picture is emerging. There are a number of water molecules that are sufficiently well localized to appear as significant electron density peaks in the maps. These are usually found within hydrogen bonding distance of one or more polar groups on the protein. Second shell water molecules are rarely

Annu. Rev. Biophys. Bioeng. 1977.6:151-176. Downloaded from arjournals.annualreviews.org
by University of Texas - Austin on 09/28/09. For personal use only.

seen. The sum of all identified water peaks is far less than the total amount of water present in the crystals. The unobserved water is presumed to have a high mobility. This mobility may or may not be equivalent to bulk water. The X-ray data provide no information on this point. NMR relaxation data indicate a slow exchange between much of the surface water and the bulk solvent, but this trapped water appears to be in rapid motion. (All of these observations are affected to a greater or lesser extent by uncertainties over the distribution of other components of the solvent such as salt ions.) In the light of conclusions reached earlier, it may be significant that the mean Voronoi volume of the water identified in the lysozyme structure is identical to that of bulk water, although the volume variation among individual water molecules is considerable (33).

Annu. Rev. Biophys. Bioeng. 1977.6:151-176. Downloaded from arjournals.annualreviews.org
by University of Texas - Austin on 09/28/09. For personal use only.

CHAIN FOLDING

When the ϕ and ψ main chain dihedral angles are fixed for each residue in a chain, 67% of all the atoms (N,C,O,C_α,C_β) will have been positioned for any globular protein of average composition. Since these atoms are common to all proteins, the actual structure assumed by any given protein is controlled by the 33% of the atoms not fixed by ϕ and ψ. Specification of the side chain dihedral angle χ_1, rotation about the alpha-beta bond, raises the total number of atoms fixed to over 80%, and with the addition of χ_2 to over 93%. The remaining angles refer only to such flexible side chains as glutamic acid, lysine, and arginine, which are almost invariably on the protein surface and whose detailed specification may not be required until late in the folding process, if then.

Secondary Structure

If any ϕ-ψ pair is repeated sequentially along the main chain, a helix will result. The hard sphere, or Ramachandran map, shows that for amino acids other than glycine only three general regions are available for such pairs, the right- and left-handed α-helices and the extended β structure (7, 75–77). Observed dihedral angles fall almost entirely in these areas. Close packing requirements explain why larger helices with central holes are never found. The 3_{10} helix is a bit too tightly packed apparently for a stable extended structure, but a few residues in this conformation frequently occur at the C terminal end of an α-helix or individually in turns (see discussion in 78, 79). In the β region there is more latitude in ϕ-ψ choices and the somewhat irregular β sheets that are frequently observed are compatible with this. The importance of the ϕ-ψ area distribution can be seen in the twist that β sheets normally have, as pointed out and analyzed by Chothia (80). Only occasional single residues are found in the left-handed α-helix region. Thus it is not surprising that α-helices and β chains are the only indefinitely extendable secondary structural elements so far found. It is unlikely that future completed protein structures will reveal any others.

Compact particles require folding the chain back on itself and thus bends are the third general aspect of secondary structure so far codified. The specific bend geome-

Annu. Rev. Biophys. Bioeng. 1977.6:151-176. Downloaded from arjournals.annualreviews.org
by University of Texas - Austin on 09/28/09. For personal use only.

try based on two ϕ-ψ pairs was first established by Venkatachalam (81) and subsequently by Lewis et al (82, 83). The pairs are neither identical nor repeating, thus distinguishing bends from α or β structures.

Residues that do not fall into any of these categories are frequently referred to as random. This, of course, is a misnomer as they are as fixed in the molecular space as any other residues. They simply do not fall into neat categorizable units. With the exception of proline any single residue can be forced into an α-helix or a β chain. However, there are clearly established preferences.

The organization of helices, β chains, and bends into regular higher-level units has recently been recognized. The name super secondary structure has been proposed by Rossmann for such units (84). The search for and codification of such units in a general sense is only now underway (85–87). As expected, the permitted types appear to be largely controlled by packing considerations (87). There is no evidence yet that they exist outside of completed protein molecules, but some may represent transient and important intermediates in the folding process. The packing of two helices was considered a number of years ago by Crick (88), resulting in the specification of coiled coils in fibrous proteins. Lim (89) has considered this problem in more detail at a higher level of resolution and has shown how the observed orientations improve polar side chain interaction as well as the hydrophobic group packing.

Even longer range order between high-density regions has been found by Federov and his colleagues. Evidence of regular inhomeogeneities has been provided through the use of properly constructed radial distribution functions both in single chain (90) and oligomeric proteins (91).

Statistical Correlations

Over the past decade a number of different groups have attempted to correlate the observed secondary structural elements in X-ray-derived protein models with the amino acid sequence [see review by Fasman (92)]. Many of the correlation procedures have relied solely on the observed structures and the numerical appearance data. The derived α, β, or bend preferences are sufficiently marked that they clearly represent some important chemical aspects of the individual side chain, although such properties were not explicitly assumed in the analysis.

Recently Lim has applied the hard sphere packing approach to this problem in an interesting way (93–95). In an α-helix consider the positions i, i+3, i+4. The center of the triangle so formed will be a potential cavity in the surface of the helix whose shape and dimensions will depend on the nature of the three side chains. For such a unit to actually exist on the inward facing part of a helix, this cavity will have to be filled efficiently by the side chain of some sequentially distant residue from another part of the chain. There will be certain triads for which no side chain can be found that would properly fill the cavity, and which therefore could not exist in α-helical form. Lim identified all such triads by model building. This data and other geometric and packing considerations lead to a prediction scheme thus developed without reference to any actual protein structures (see also 96–98). In application this scheme compares favorably with those involving statistical analysis.

Annu. Rev. Biophys. Bioeng. 1977.6:151-176. Downloaded from arjournals.annualreviews.org
by University of Texas - Austin on 09/28/09. For personal use only.

Prediction

The correlations, noted above, have been inverted to provide rules for predicting a secondary structure from known amino acid sequences. Some recent blind tests have been very interesting. A variety of different predictive procedures were applied to the sequence of adenylate kinase without knowledge of the X-ray structure (99). The mean predictions were in reasonable agreement with the observed structure. However, the same test on phage T4 lysozyme gave results that were considerably less satisfactory (100). The inherent difficulties with prediction schemes have been discussed by Burgess et al (101, 102). The early hope that such schemes would be precise enough to lead quite directly to three-dimensional structures is still unfulfilled.

What the prediction efforts do indicate, however, is very significant and useful in its own right. The statistical correlations have as input the tertiary structure of the proteins with all of the multiple interactions that that implies. Assigned probabilities are based directly on these structures. There is no a priori reason that the assigned probabilities should predict the existence of the same units of secondary structure in the same sequence when in free solution isolated from the rest of the protein. However, this does appear to be the case. The assigned probabilities for α-helix formation correlate well with experimental helicity factors from synthetic homo- or co-polymer systems (103). The properties of a single residue can be quantitated as a useful predictor of secondary structure potential independent of other parts of the sequence or of higher levels of protein structure. The major importance of short-range interactions has been emphasized by Ramachandran & Sasisekharan (7), Scheraga et al (104–107), and Ptitsyn (78). Intrinsic secondary structure may be further stabilized by interactions of residues three or four units away (79).

Step I: Nucleation

Since the early discussions of timing problems, it has been clear that the folding process must consist of efficient nucleation followed by a fairly narrowly defined path of conformational adjustment to yield the compact native structure (1, 59, 108–111). The broad steps have been suggested by several investigators, perhaps most clearly by Ptitsyn (112) (see also 113, 113a). Step I is the formation of fluctuating nuclei (almost certainly helices); step II is the collapse of these nuclei into larger subassemblies (now including bends and β structure). [As alternative extremes, the product of step II may be considered either (*a*) a loose structure that grossly resembles the final structure but is imperfect in detail at the atomic level, or (*b*) a collection of connected supersecondary units that are fully finished except for the final condensation involving no change at the atomic level within these units. In a given real case some intermediate situation seems the most likely.] Step III is the condensation of the larger unit(s) into the final compact form. Essential to this proposal is the concept of retaining structural elements, i.e. all the helices present at the end of step III were present as fluctuating helices at step I, etc. Current prediction schemes are really directed at I and II and not at III. All possible

Annu. Rev. Biophys. Bioeng. 1977.6:151-176. Downloaded from arjournals.annualreviews.org
by University of Texas - Austin on 09/28/09. For personal use only.

secondary structural elements should be predicted for the early steps, even overlapping predictions (see especially 114). Subsequent algorithms for steps II and/or III will then discard irrelevant predictions but will not have to invent secondary structure themselves. The difficulty in precisely predicting where helical segments begin and end becomes unimportant. The step II folding algorithm must be able to handle fuzzy bits of structure because this is undoubtedly exactly what happens in the real situation. Viewed in this light some of the current schemes may be quite useful. The formation of helical nuclei is discussed in detail by Ptitsyn and his associates (78, 115–118).

The most general force encouraging the formation of such nuclei is the hydrophobic or solvent-squeezing effect, as estimated by changes in accessible surface area. As an example, Richards et al (119) calculated the area changes accompanying the formation of ribonuclease S from S protein and the 20-residue S peptide component in two hypothetical steps: (1) conversion of the extended peptide to the two helical turn conformation it would have in the complex (step I and II*b* for S peptide), and (2) association of this preformed unit with S protein to give the final complex (step III for the complex). Of the total area change 60% occurs in (1) and 40% in (2). Chothia has made a much more detailed analysis of six proteins (120). The surface area changes on formation of secondary structure are larger in all cases than the subsequent changes on aggregation of these units. A given residue loses more solvent contact on insertion into a β sheet than into an α-helix. Of the polar groups forming hydrogen bonds within the protein 80% are between donor-acceptor pairs in the same piece of secondary structure. The intrinsic stability of the secondary structural elements thus is larger than that of the next higher level of organization, and therefore the basic assumption of the proposed sequence of folding events is reasonable.

Step II: Collapse

Step II is the least-well-developed part of the folding scheme. Distances are still quite large and the required docking algorithms for association of the secondary structural units have yet to be developed in detail. One would expect packing criteria to be most useful here. Two very interesting papers have appeared recently. Ptitsyn & Rashin (121) took a physical packing model of myoglobin consisting of idealized helices with flexible connections, a perfect step I structure. The helices were then packed together by hand in as many ways as could be devised and energy estimates were made of the products at each stage. There was only one product at the lowest energy and it bore a remarkable resemblance to the actual myoglobin structure. Computerization of this approach is very desirable but very difficult.

Levitt & Warshel (122) idealized a chain to a point permitting computation. Each residue was characterized by a single main chain dihedral angle, a spherical side chain of appropriate size, and a set of modified energy parameters. They allowed the computer to fold an extended chain based on the pancreatic trypsin inhibitor sequence, hopefully avoiding the false minimum problem by thermally kicking the chains at intervals. On occasion the final structure resembled the known structure

but not always. This intriguing approach clearly deserves further development. It is not yet obvious at what stage the idealized chain approximation will have to be abandoned to get reasonable packing structures. Also, this initial attempt essentially combines steps I and II into a single effort in contrast to the approach of Ptitsyn & Rashin (121). That aspect of the strategy deserves detailed consideration. The observations of Wetlaufer et al (123) on the organization of secondary structural units should be useful at this stage.

Step III: Condensation

In principle, the energy minimization procedures already well developed should be able to handle step III for product II*a* to the final structure (124, 125). The actual changes in atomic coordinates produced by these procedures has generally been small, <1 Å, even when the change in energy was very large. This, at least in part, is due to the local minimum problem and may be resolved by modified search algorithms. However, the product of step II*a* in the folding process should not have to be quite so close to the final structure as is implied by the behavior of current minimization procedures.

Condensation from the II*b* product will require the same kind of difficult docking algorithms required for the collapse from I to II.

Random Coil

The starting conformation for step I has always been called the random coil, perhaps a misleading name. Peptide chains are stiff and normally not long enough to behave as true random coils. Further there is quite a significant interaction between the ϕ-ψ angles and χ_1 (16, 126–128), although this is disputed (101). The γ atom is constrained when ϕ-ψ are fixed or, conversely, the side chain conformation influences the ϕ-ψ space available to the main chain. The van der Waals surface of the γ atom becomes of critical importance. In one class of residues this atom is branched, in a second class it is a methylene group. Serine stands out as a unique residue. The small radius of its oxygen atom produces a considerable increase in conformational flexibility.

Since ϕ-ψ-χ_1 together fix 80% of all atoms in the structure, the conformational space available to a random coil is dramatically reduced when these three angles become linked. The formation of step I nuclei in permitted sequences will be accompanied by considerable immobilization of the side chains. This process will occur in a concerted fashion because of the angle linkage, and nucleation will be facilitated.

In a native protein one can calculate the contacts (and energy) as the γ atom of each residue in turn is moved by rotation of χ_1. It is found that the observed χ_1 values correspond to unhindered positions at or close to the energy minimum in over two-thirds of all the residues surveyed (128). It would appear that the steps of the folding process maintain the atom in a minimum in spite of the packing or polar interaction requirements of the rest of the side chain that may develop in the late states of completion of the structure. The significance of these observations for the folding process is pointed out by Gelin & Karplus (128).

Annu. Rev. Biophys. Bioeng. 1977.6:151-176. Downloaded from arjournals.annualreviews.org
by University of Texas - Austin on 09/28/09. For personal use only.

Annu. Rev. Biophys. Bioeng. 1977.6:151-176. Downloaded from arjournals.annualreviews.org
by University of Texas - Austin on 09/28/09. For personal use only.

Subunit Associations

The development of quaternary structure from peptide chains already folded into their correct tertiary structure has been investigated by Janin & Chothia. They examined the interfaces between the separate chains in the trypsin-pancreatic trypsin inhibitor complex, the insulin dimer, and the α and β chains of hemoglobin (129–131). Volume calculations showed that the packing density of the residues was indistinguishable from that found in the tertiary structure of individual chains, quantitating the degree of complementarity of the two surfaces. The accessible area decreases caused by association gave the hydrophobic contribution to the interaction energy and were correlated with the experimental values. The loss in interface area in going from the deoxy to the oxy form of hemoglobin was suggested as the important energetic factor in the allosteric binding of O_2 rather than the salt bridges previously proposed.

FLEXIBILITY

Very little information is presently available that permits one to distinguish between actual thermal motion and statistical disorder in protein crystals, the predicted effects on a single diffraction pattern being identical. To date differences in mean structure caused by changes in solvent conditions or by the binding of ligands have provided the most definitive evidence of flexibility and motion, but the time parameter is missing from such comparisons. Techniques other than X-ray diffraction, although less definitive in the sense of overall structure, are better adapted to provide the time base of structural fluctuation or change.

Conformational changes within a single protein molecule encompass a large variety of motions involving different fractions of the total structure. The rates of these processes cover an enormous range on the time scale, nanoseconds to days. These motions can be assessed by spectroscopic procedures, such as magnetic resonance (133) or florescence (134), or intrinsically chemical procedures, such as proton exchange (135, 136). These time-base measurements clearly will be a field of intense activity during the next few years. There are some recent examples: the motion of individual tyrosine (137) and phenylalanine residues (128); large-scale motion implied by tryptophan fluorescence quenching by oxygen (138) and by acrylamide (138a, 138b); hydrogen exchange of specific amide groups in a protein (139); the possible importance of the slow *cis-trans* isomerization of proline in protein folding (140); differential accessibility of ends and central portion of an α-helix (141). All of these motions make an important contribution to the entropy of proteins and to their function (142–144). Cooper (145) has made some provocative suggestions on the probable large magnitude of fluctuations in individual molecules based on the thermodynamics of small systems.

A rigorous analysis of normal modes in an object as large as a protein seems unlikely in the near future. An approximate mechanical analysis assisted by the study of packing defects may be useful. To date this does not appear to have been attempted, and further discussion of molecular dynamics is outside the scope of this review.

ACKNOWLEDGMENTS

In addition to my immediate colleagues, T. A. Steitz, T. Richmond, D. LeMaster, and J. Mercer, I would like to especially thank O. B. Ptitsyn, A. Hvidt, and J. Finney for their helpful comments and criticisms.

Literature Cited

1. Anfinsen, C. B., Scheraga, H. A. 1975. *Adv. Prot. Chem.* 29:205–300
2. Baldwin, R. L. 1975. *Ann. Rev. Biochem.* 44:453–75
3. Wetlaufer, D. B., Ristow, S. 1973. *Ann. Rev. Biochem.* 42:135–58
4. Ptitsyn, O. B., Lim, V. I., Finkelstein, A. V. 1972. *Fed. Eur. Biochem. Soc.* 25:421–29
5. Jensen, L. 1974. *Ann. Rev. Biophys. Bioeng.* 3:81–93
6. Matthews, B. W. 1976. *Ann. Rev. Phys. Chem.* 27:493–524
7. Ramachandran, G. N., Sasisekharan, V. 1968. *Adv. Prot. Chem.* 23:284–438
8. Bondi, A. 1968. *Physical Properties of Molecular Crystals, Liquids, and Glasses.* New York: Wiley
9. Bondi, A. 1964. *J. Phys. Chem.* 68: 441–51
10. Chothia, C. 1975. *Nature* 254:304–8
11. Hermann, R. B. 1972. *J. Phys. Chem.* 76:2754–59
12. Lee, B., Richards, F. M. 1971. *J. Mol. Biol.* 55:379–400
13. Shrake, A., Rupley, J. A. 1973. *J. Mol. Biol.* 79:351–71
14. Harris, M. J., Higuchi, T., Rytting, J. H. 1973. *J. Phys. Chem.* 77:2694–703
15. Reynolds, J. A., Gilbert, D. B., Tanford, C. 1974. *Proc. Natl. Acad. Sci. USA* 71:2925–27

15a. Sinanoğlu, O. 1968. In *Molecular Association in Biology,* pp. 427–45. New York: Academic

15b. Sinanoğlu, O., Abdulnur, A. 1965. *Fed. Proc.* 24:Suppl. 15, pp. S12–23

16. LeMaster, D., Richards, F. M. Unpublished data
17. Teller, D. C. 1976. *Nature* 260:729–31
18. Kauzmann, W. 1959. *Adv. Prot. Chem.* 14:1–64
19. Nozaki, Y., Tanford, C. 1971. *J. Biol. Chem.* 246:2211–17
20. Chothia, C. 1974. *Nature* 248:338–39
21. Klapper, M. H. 1973. *Prog. Bioorg. Chem.* 2:55–132
22. Tanford, C. 1973. *The Hydrophobic Effect.* New York: Wiley
23. Bondi, A. 1954. *J. Phys. Chem.* 58:929–39
24. Edward, J. T. 1956. *Chem. Ind.* pp. 774–77
25. Edward, J. T. 1956. *Sci. Proc. R. Dublin Soc.* 27:273–82
26. Voronoi, G. F. 1908. *J. Reine Angew. Math.* 134:198–287
27. Coxeter, H. S. M. 1961. *Introduction to Geometry.* New York: Wiley
28. Bernal, J. D. 1964. *Proc. R. Soc. London Ser. A* 280:299–322
29. Bernal, J. D. 1965. *Liquids: Structure, Properties, Solid Interactions,* ed. T. J. Hughel, pp. 25–50. Amsterdam: Elsevier
30. Bernal, J. D., King, S. V. 1967. *Disc. Faraday Soc.* 43:60–69
31. Finney, J. L. 1970. *Proc. R. Soc. London Ser. A* 319:479–93
32. Finney, J. L. 1970. *Proc. R. Soc. London Ser. A* 319:495–507
33. Richards, F. M. 1974. *J. Mol. Biol.* 82:1–14
34. Finney, J. L. 1975. *J. Mol. Biol.* 96:721–32
35. Deleted in proof
36. Klapper, M. H. 1971. *Biochim. Biophys. Acta* 229:557–66
37. Lim, V. I., Ptitsyn, O. B. 1972. *Biophysics USSR* 17:21–33
38. Lim, V. I., Ptitsyn, O. B. 1970. *Mol. Biol. USSR* 4:372–82
39. Wyckoff, H. W. 1968. *Brookhaven Symp. Biol.* 21:252–57
40. Motherwell, W. D. S., Isaacs, N. W. 1972. *J. Mol. Biol.* 71:231–41
41. Motherwell, W. D. S., Riva di Sanseverino, L., Kennard, O. 1973. *J. Mol. Biol.* 80:405–22
42. Kauzmann, W., Moore, K., Schultz, D. 1974. *Nature* 248:447–49
43. Schultz, D. 1976. *Density in submolecular regions of globular proteins.* Ph.D. thesis. Princeton Univ., Princeton, N.J. 145 pp.

43a. Cooke, R., Kuntz, I. D. 1974. *Ann. Rev. Biophys. Bioeng.* 3:95–126

44. Kuntz, I. D., Kauzmann, W. 1974. *Adv. Prot. Chem.* 28:239–347
45. Pace, N. 1976. *Crit. Rev. Biochem.* In press
46. Bryant, R. G. 1974. *J. Am. Chem. Soc.* 96:297–99

Annu. Rev. Biophys. Bioeng. 1977.6:151-176. Downloaded from arjournals.annualreviews.org by University of Texas - Austin on 09/28/09. For personal use only.

Annu. Rev. Biophys. Bioeng. 1977.6:151-176. Downloaded from arjournals.annualreviews.org
by University of Texas - Austin on 09/28/09. For personal use only.

47. Bryant, R. G. 1976. *J. Am. Chem. Soc.* In press
48. Moews, P. C., Kretsinger, R. H. 1975. *J. Mol. Biol.* 91:201–28
49. Rupley, J. 1969. *Structure and Stability of Biological Macromolecules,* ed. S. N. Timasheff, G. D. Fasman, Chap. 4, pp. 291–352. New York: Dekker
50. Bishop, W. H., Richards, F. M. 1968. *J. Mol. Biol.* 33:415–21
51. Bishop, W. H., Richards, F. M. 1968. *J. Mol. Biol.* 38:315–28
52. Scanlon, W. J., Eisenberg, D. 1976. *J. Mol. Biol.* In press
53. Assarsson, P. G., Eirich, F. R. 1968. *J. Phys. Chem.* 72:2710–19
54. Bøje, L., Hvidt, A. 1971. *J. Chem. Thermodyn.* 3:663–73
55. Low, B. W., Richards, F. M. 1954. *J. Am. Chem. Soc.* 76:2511–18
56. Finney, J. L. 1976. *Philos. Trans. Roy. Soc. London.* In press
57. Tanford, C. 1968. *Adv. Prot. Chem.* 23:121–282
58. Tanford, C. 1969. *Adv. Prot. Chem.* 24:1–97
59. Tsong, T. Y., Baldwin, R. L., McPhie, P. 1972. *J. Mol. Biol.* 63:453–75
60. Holcomb, D. N., Van Holde, K. E. 1962. *J. Phys. Chem.* 66:1999–2006
61. Brandts, J. F. 1969. *Structure and Stability of Biological Macromolecules,* ed. S. N. Timasheff, G. D. Fasman, Chap. 3, pp. 213–90. New York: Dekker
62. Brandts, J. F., Oliveira, R. J., Westort, C. 1970. *Biochemistry* 9:1038–47
63. Zipp, A., Kauzmann, W. 1973. *Biochemistry* 12:4217–28
64. Hawley, S. A., Mitchell, R. M. 1975. *Biochemistry* 14:3257–64
65. Li, T. M., Hook, J. W. III, Drickamer, H. G., Weber, G. 1976. *Biophys. J.* 16:TH-AM-F3(Abstr.)
66. Li, T. M., Hook, J. W., Drickamer, H. G., Weber, G. 1976. *Biochemistry* 15: 3205–11
67. Krausz, L. M. 1970. *J. Am. Chem. Soc.* 92:3168–73
68. Friedman, M. E., Scheraga, H. A. 1965. *J. Phys. Chem.* 69:3795–800
69. Bøje, L., Hvidt, A. 1972. *Biopolymers* 11:2357–64
70. Hvidt, A. 1975. *Colloq. Int. CRNS.* In press
71. Hvidt, A. 1975. *J. Theor. Biol.* 50: 245–52
72. Fixman, M., Skolnick, J. Manuscript in preparation
73. Moult, J., Yonath, A., Traub, W., Smilansky, A., Podjarny, A., Rabinovich, D., Saya, A. 1976. *J. Mol. Biol.* 100:179–95
74. Watenpaugh, K. D., Sieker, L. C., Herriott, J. R., Jensen, L. H. 1973. *Acta Crystallogr. Sect. B* 29:943–56
75. Ramachandran, G. N., Ramakrishnan, C., Sasisekharan, V. 1963. *J. Mol. Biol.* 7:95–99
76. Ramachandran, G. N. 1968. *Structural Chemistry and Molecular Biology,* ed. A. Rich, N. Davidson, pp. 77–87. San Francisco: Freeman
77. Ramachandran, G. N., Ramakrishnan, C., Venkatachalam, C. M. 1965. *Biopolymers* 3:591–92
78. Finkelstein, A. V., Ptitsyn, O. B. 1976. *Biopolymers.* In press
79. Ponnuswamy, P. K., Warme, P. K., Scheraga, H. A. 1973. *Proc. Natl. Acad. Sci. USA* 70:830–33
80. Chothia, C. 1973. *J. Mol. Biol.* 75:295–302
81. Venkatachalam, C. M. 1968. *Biopolymers* 6:1425–36
82. Lewis, P. N., Momany, F. A., Scheraga, H. A. 1973. *Biochim. Biophys. Acta* 303:211–29
83. Lewis, P. N., Momany, F. A., Scheraga, H. A. 1971. *Proc. Natl. Acad. Sci. USA* 68:2293–97
84. Rao, S. T., Rossmann, M. G. 1973. *J. Mol. Biol.* 76:241–56
85. Tufty, R. M., Kretsinger, R. H. 1975. *Science* 187:167–69
86. Kuntz, I. D. 1976. *J. Am. Chem. Soc.* In press
87. Levitt, M., Chothia, C. 1976. *Nature* 261:552–58
88. Crick, F. H. C. 1953. *Acta Crystallogr.* 6:689–97
89. Lim, V. I. 1976. *J. Mol. Biol.* In press
90. Fedorov, B. A. 1976. *FEBS Lett.* 62:139–41
91. Damaschun, G., Müller, J. J., Gedicke, Ch., Fedorov, B. A. 1976. *J. Mol. Biol.* 104:735–39
92. Fasman, G. D. 1977. *Ann. Rev. Biochem.* 46:000
93. Lim, V. I. 1972. *Dokl. Akad. Nauk SSSR* 203:103–5
94. Lim, V. I. 1974. *J. Mol. Biol.* 88:857–72
95. Lim, V. I. 1974. *J. Mol. Biol.* 88:873–94
96. Nagano, K. 1973. *J. Mol. Biol.* 75:401–20
97. Nagano, K. 1974. *J. Mol. Biol.* 84:337–72
98. Nagano, K., Hasegawa, K. 1975. *J. Mol. Biol.* 94:257–81
99. Schulz, G. E., Barry, C. D., Friedman, J., Chou, P. Y., Fasman, G. D., Finkelstein, A. V., Lim, V. I., Ptitsyn, O. B.,

Kabat, E. A., Wu, T. T., Levitt, M., Robson, B., Nogano, K. 1974. *Nature* 250:140–42
100. Matthews, B. W. 1975. *Biochim. Biophys. Acta* 405:442–51
101. Burgess, A. W., Scheraga, H. A. 1975. *Proc. Natl. Acad. Sci. USA* 72:1221–25
102. Burgess, A. W., Ponnuswamy, P. K., Scheraga, H. A. 1974. *Isr. J. Chem.* 12:239–86
103. Maxfield, F. R., Alter, J. E., Taylor, G. T., Scheraga, H. A. 1975. *Macromolecules* 8:479–91
104. Scheraga, H. A. 1973. *Pure Appl. Chem.* 36:1–8
105. Lewis, P. N., Momany, F. A., Scheraga, H. A. 1973. *Isr. J. Chem.* 11:121–52
106. Howard, J. C., Ali, A., Scheraga, H. A., Momany, F. A. 1975. *Macromolecules* 8:607–22
107. Burgess, A. W., Scheraga, H. A. 1973. *Biopolymers* 12:2177–83
108. Wetlaufer, D. B. 1973. *Proc. Natl. Acad. Sci. USA* 70:697–701
109. Rose, G. D., Winters, R. H., Wetlaufer, D. B. 1976. *FEBS Lett.* 63:10–16
110. Ralston, E., DeCoen, J.-L. 1974. *J. Mol. Biol.* 83:393–420
111. DeCoen, J. L., Ralston, E. 1973. *The Jerusalem Symposia on Quantum Chemistry and Biochemistry*, ed. E. D. Bergmann, B. Pullman, Vol. 5, pp. 41–48. Jerusalem: Academic
112. Ptitsyn, O. B. 1973. *Dokl. Akad. Nauk SSSR* 210:87–89
113. Tanaka, S., Scheraga, H. A. 1975. *Proc. Natl. Acad. Sci. USA* 72:3802–6
113a. Karplus, M., Weaver, D. L. 1976. *Nature* 260:404–6
114. Lim, V. I. 1975. *Dokl. Akad. Nauk SSSR* 222:1467–69
115. Denesyuk, A. I., Ptitsyn, O. B., Finkelstein, A. V. 1974. *Biofizika* 19:549–61
116. Finkelstein, A. V., Ptitsyn, O. B., Kozitsyn, S. A. *Biopolymers.* In press
117. Finkelstein, A. V. 1976. *Biopolymers.* In press
118. Finkelstein, A. V., Ptitsyn, O. B. 1976. *J. Mol. Biol.* 103:15–24
119. Richards, F. M., Wyckoff, H. W., Carlson, W., Allewell, N. M., Lee, B., Mitsui, Y. 1971. *Cold Spring Harbor Symp. Quant. Biol.* 36:35–43
120. Chothia, C. 1976. *J. Mol. Biol.* 105: 1–14
121. Ptitsyn, O. B., Rashin, A. A. 1973. *Dokl. Akad. Nauk SSSR* 213:473–75
122. Levitt, M., Warshel, A. 1975. *Nature* 253:694–98
123. Wetlaufer, D. B., Rose, G. D., Taaffe, L. 1976. *Biochemistry.* In press
124. Levitt, M. 1974. *J. Mol. Biol.* 82:393–420
125. Warme, P. K., Scheraga, H. A. 1974. *Biochemistry* 13:757–67
126. Ramachandran, G. N., Lakshminarayanan, A. V. 1966. *Biopolymers* 4:495–97
127. Finkelstein, A. V. 1976. *Mol. Biol. USSR* 10:507–13, 879–86
128. Gelin, B., Karplus, M. 1975. *Proc. Natl. Acad. Sci. USA* 72:2002–6
129. Chothia, C., Janin, J. 1975. *Nature* 256:705–8
130. Janin, J., Chothia, C. 1976. *J. Mol. Biol.* 100:197–211
131. Chothia, C., Janin, J., Wodak, S. 1976. *J. Mol. Biol.* In press
132. Deleted in proof
133. Cohen, J. S. 1977. *Ann. Rev. Biophys. Bioeng.* 6:383–417
134. Stryer, L. 1977. *Ann. Rev. Biochem.* 46: In press
135. Englander, S. W., Downer, N. W., Teitelbaum, H. 1972. *Ann. Rev. Biochem.* 41:903–24
136. Hvidt, A. 1973. *Dynamic Aspects of Conformation Changes in Biological Macromolecules*, ed. C. Sadron, D. Reidel, pp. 103–15. Holland: Dordrecht
137. Hull, W. E., Sykes, B. D. 1975. *J. Mol. Biol.* 98:121–53
138. Lakowicz, J. R., Weber, G. 1973. *Biochemistry* 12:4161–70
138a. Eftink, M. R., Ghiron, C. A. 1975. *Proc. Natl. Acad. Sci. USA* 72:3290–94
138b. Eftink, M. R., Ghiron, C. A. 1976. *Biochemistry* 15:672–680
139. Schreier, A. A., Baldwin, R. L. 1976. *J. Mol. Biol.* In press
140. Brandts, J. F., Halvorson, H. R., Brennan, M. 1975. *Biochemistry* 14:4953–63
141. Nakanishi, M., Tsuboi, M., Ikegami, A., Kanehisa, M. 1972. *J. Mol. Biol.* 64:363–78
142. Go, N., Scheraga, H. A. 1969. *J. Chem. Phys.* 51:4751–67
143. Page, M. I., Jencks, W. P. 1971. *Proc. Natl. Acad. Sci. USA* 68:1678–83
144. Jencks, W. 1975. *Adv. Enzymol.* 43: 219–410
145. Cooper, A. 1976. *Proc. Natl. Acad. Sci. USA* 73:2740–41

Annu. Rev. Biophys. Bioeng. 1977.6:151-176. Downloaded from arjournals.annualreviews.org
by University of Texas - Austin on 09/28/09. For personal use only.