made the ^{13}C spectra unreliable for the prediction of structure. Often biologically quite different compounds are sufficiently similar chemically and spectroscopically that it is difficult to distinguish them. For example, if given the mass spectrum of a tri- or higher-substituted benzene, one cannot select among the several positional isomers. Nmr, however, can identify such positional isomers. In particular, ^{13}C nmr, because of the larger range of chemical shifts observed and the ease of interpretation of spectra, is ideally suited for the characterization of biological unknowns. If predictability of substituent effects on ^{13}C chemical shifts can be shown to be generally true for other aromatic systems of biological importance, this fact would be of great significance in the characterization of biological unknowns.

Acknowledgments. This investigation was supported by PHS Research Grant NS 09576 from the National Institute of Neurological Diseases and Stroke. The author wishes to thank Dr. Clyde M. Williams for suggesting the compilation of data on aromatic acids and is grateful to Professor W. S. Brey, Jr., for many helpful discussions during this investigation. The author is indebted to Dr. W. Jankowski for obtaining Fourier transform spectra of the disubstituted benzoic acids. The XL-100 nmr spectrometer was purchased in part from funds supplied by the National Science Foundation.

Tertiary Structure in Carboxypeptidase

I. D. Kuntz

Contribution from the Department of Pharmaceutical Chemistry, School of Pharmacy, University of California, San Francisco, San Francisco, California 94122. *Received March 6, 1972*

Abstract: Intramolecular hydrogen bonding and hydrophobic interactions in carboxypeptidase are analyzed. The hydrogen bonding in the helices and the β sheet is well described by the classical models. The ends of each helix and the ends and edges of the β sheet are not hydrogen bonded within the protein molecule. Hairpin turns do not fit the bonding pattern suggested in the literature. Few internal hydrogen bonds are made in the turns, so that the turns appear to be weak links in the peptide chain. For the entire protein about 42% of the potential donor or acceptor sites are not occupied by intramolecular bonds. It is likely that water molecules are bound at most of these sites. The eight helices are held to each other and to the β sheet by a maximum of 17 hydrogen bonds, less than two bonds per helix. This level of interconnection appears insufficient to be the dominant term in determining protein tertiary structure. To explore the role of hydrophobic bonding, we define all side-chain carbon atoms as "hydrophobic." We find large regions (of the order of 1000 Å^3) of carboxypeptidase that contain *ca.* 97% hydrophobic atoms. These regions are roughly channel shaped and hold two-thirds of the hydrophobic atoms of the molecule. The channels are located between the helices and the β sheet, sometimes extending to the molecular surface. The density of the hydrophobic regions is surprisingly low being close to that of liquid benzene.

The importance of intramolecular hydrogen bonding and hydrophobic bonding in determining the tertiary structure of globular proteins is still a matter of some controversy. The difficulty with a facile assessment of the role of hydrogen bonding is the small free energy of formation of amide hydrogen bonds in aqueous solutions.[1] Cooperative secondary features such as α helices or β pleated sheet structures clearly involve some substantial contribution of hydrogen bonding to the overall potential energy. What is not clear is the importance of hydrogen bonding in connecting these secondary elements together. Kauzmann's discussion of hydrophobic interactions[2] has led to a very simple model for the structure of globular proteins which does not require direct reference to hydrogen bonding. In this model the protein is represented as a hydrophilic shell enclosing a hydrophobic core. Qualitative assessments of the known protein structures have supported this picture in that relatively few polar residues are found in the protein interior.[3] Recently investigations have raised some interesting difficulties. Klotz[3] and Lee and Richards[4] have pointed out that a number of hydrophobic residues are on the protein surface. The latter authors, particularly, question whether hydrophobic interactions have much importance in determining protein structure. Klapper,[5] in a different approach, has made use of scaled particle theory to estimate the atom density, free volume, and compressibility of the protein interior. He finds that the average density is so high as to make a "wax-like" rather than an "oil-like" description more useful. In sum, both of the widely discussed approaches to simple models for protein structure have yet to gain complete acceptance. In this paper we will make use of the availability of atomic coordinates for globular proteins to investigate further intramolecular hydrogen bonding and hydrophobic interactions. Our initial questions are: (1) do hydrogen bonds provide direct links between secondary structural elements; (2) can hydrophobic regions be quantitatively identified; and (3) if so, what are the

(1) I. M. Klotz and J. S. Franzen, *J. Amer. Chem. Soc.*, **84**, 3461 (1962).
(2) W. Kauzmann, *Advan. Protein Chem.*, **14**, 1 (1959).
(3) I. M. Klotz, *Arch. Biochem. Biophys.*, **138**, 704 (1970).
(4) B. Lee and F. M. Richards, *J. Mol. Biol.*, **55**, 379 (1971).
(5) M. H. Klapper, *Biochim. Biophys. Acta*, **229**, 557 (1971).

properties and functions of such regions? We have chosen carboxypeptidase for this initial study because of the high quality X-ray data, the extensive secondary structure, and the availability of a general review.[6]

Methods and Results

Hydrogen Bonding. The X-ray data do not yield the hydrogen positions directly. Given this uncertainty, we have chosen to estimate the *maximum* number of intramolecular hydrogen bonds. Thus, we will count any donor–acceptor pair with a heteroatom separation of less than or equal to 3.5 Å as a *potential* hydrogen bond.[7] This procedure clearly overestimates the number of hydrogen bonds because some pairs will have unsuitable angular alignments. Backbone donors and acceptors less than three residues apart could only form highly strained hydrogen bonds and are discarded. Hydrogen bond formation with other protein molecules in the unit cell is not considered. In general, we treated all noncarbon atoms as possible hydrogen bonding sites. Specifically, we took as proton donors peptide NH, side-chain NH's of Lys, Arg, His, and Trp, and side-chain OH's of Ser, Thr, and Tyr; we took as proton acceptors peptide carbonyl O, carboxyls of Glu and Asp, OH's of Ser and Thr, but not Tyr, and S of Met and Cys. The N and O of the amide side chains cannot be distinguished from the X-ray data. We checked each position as both a proton donor and proton acceptor site.

Carboxypeptidase contains eight helical regions and eight strands of β structure (Table I). In addition, about 45 turns of various types have been described.[8] Taken together, helix, β, and turn regions account for ~90% of the residues in the protein. Exact criteria for assignment of residues to secondary features have not been clearly established. Generally, one can use various internal coordinates (*e.g.*, Ramachandran angles) or particular hydrogen bonding patterns. These two approaches coincide quite well for helical segments and there is little ambiguity about helical assignments in carboxypeptidase. β sheets present a more difficult problem because one can find regions of extended polypeptide chain that do not make appropriate interstrand hydrogen bonds. We have chosen to assign as β structure all extended chain which, on the average, makes at least one hydrogen bond per residue to another section of β chain. Quiocho and Lipscomb[6] have required two hydrogen bonds per residue in their assignments. Dickerson and Geis[9] have chosen an intermediate value (Table I). As will become apparent, the different β assignments will not alter the major conclusions of this paper.

Our first concern was the hydrogen bonding *within* each type of secondary structure. For the backbone atoms we find that the conventional models[10] for helical and β elements are quite satisfactory representations for carboxypeptidase. The helical regions have better than 75% of the available backbone sites engaged in the expected 1–3 and 1–4 hydrogen bonding patterns. The unoccupied sites are found at the ends of a helix; typically, the first three NH's and the last three C=O's of each helix have no suitable hydrogen bonding patterns within the protein. Less than 5% of the helical backbone sites are so located as to hydrogen bond to other parts of the protein (Table II).

Table I. Structural Elements for Carboxypeptidase

Helices	This work	Dickerson and Geis[a]	Quiocho and Lipscomb[b]
	14–28	14–29	14–28
	72–88	72–88	72–88
	94–100	94–103	94–103
	112–120	115–123	112–122
	172–186	174–184	173–187
	215–232	215–233	215–231
	253–262	254–262	254–262
	285–306	288–305	285–306
β sheet	32–39	32–37	32–36
	44–54	45–54	49–53
	59–66	61–67	60–67
	103–109	103–111	104–109
	190–197	190–197	190–196
	200–205	200–205	200–204
	235–242	238–243	239–241
	265–271	265–271	265–271
Hairpin turns[c]	30–33		
	41–44		
	56–59		
	89–92		
	148–151		
	169–173		
	213–216		
	232–235		
	244–247		
	275–278		
	277–280		

[a] Reference 9. [b] Reference 6. [c] Reference 8.

The precise details of hydrogen bond formation for β structure can be expected to depend on whether parallel or antiparallel strand alignment occurs. The maximum fraction of filled sites is similar for both arrangements and is $(N - 1)/N$ where N is the number of strands. Here we assume that one-half of the sites on each of the two edge strands cannot bond *within* the sheet. For carboxypeptidase this model predicts a maximum occupancy of 88% of the β backbone sites. Our β assignments (Table I) show 76/123 or 62% of the sites have suitable separations to be counted as hydrogen bonded. Using Quiocho and Lipscomb's assignments the corresponding occupancy is 77%. In either case the unoccupied sites occur, as expected, at the edges of the sheet. About 10% of the β backbone sites are found to be hydrogen bonded to other regions of the protein.

Possible hydrogen bonding conformations for backbone atoms involved in high-angle "hairpin" turns have been described.[10] No turns of these precise geometries are actually found in carboxypeptidase, but if the turn parameters are relaxed somewhat,[11] 11 hairpin turns involving 42 residues are present.[8] The 1,4(C=O)—NH bond is the only backbone hydrogen bond predicted for this structure. Curiously, only two of the 11 turns

(6) Coordinates courtesy of W. Lipscomb: F. A. Quiocho and W. N. Lipscomb, *Advan. Protein Chem.*, **25**, 1 (1971).
(7) G. C. Pimentel and A. L. McClellan, "The Hydrogen Bond," W. A. Freeman, San Francisco, Calif., 1960.
(8) I. D. Kuntz, *J. Amer. Chem. Soc.*, **94**, 4009 (1972).
(9) R. E. Dickerson and I. Geis, "Stereo Supplement to the Structure and Action of Proteins," Harper & Row, New York, N. Y., 1969.
(10) G. N. Ramachandran and V. Sasisekharan, *Advan. Protein Chem.*, **23**, 284 (1968).
(11) Hairpin turns were taken as those nonhelical residues for which the C^{α}_i to C^{α}_{i+3} distance is less than 5.5 Å and the vectors linking C^{α}_i to C^{α}_{i+1} and C^{α}_{i+2} to C^{α}_{i+3} make an angle greater than or equal to 135°.

Table II. Fraction of Sites Potentially Hydrogen Bonded

	Helix	β	Hairpins	All residues
Backbone sites				
Donor	93/111, 84%	45/60, 75%	16/40, 40%	182/297, 61%
Acceptor	93/114, 82%	45/63, 72%	14/42, 33%	190/307, 62%
Side-chain sites				
Donor	33/49, 67%	14/26, 54%	8/18, 44%	80/147, 54%
Acceptor	24/47, 51%	8/17, 47%	4/13, 31%	50/117, 43%
Intraelement hydrogen bonds	100 for 114 residues	42 for 63 residues	3 for 42 residues	

show this bonding pattern. There is no distinct hydrogen bonding pattern for hairpin turns. An average of 37% of the backbone sites are occupied.

Taking the protein as a whole, a substantial number of backbone hydrogen bonds are not made. The (maximum) occupancy for all backbone sites leaves 232 of 605 sites unoccupied. It is unlikely that many of these sites are involved in intermolecular bonding between adjacent protein molecules. Most (>85%) are located at or very near the molecular surface. It seems likely that water molecules serve as appropriate donors or acceptors at these positions.

There are few guidelines in the literature for the conformation of the side-chain hydrogen bonds. For carboxypeptidase we find: (1) for any structural element, the fraction of side-chain sites involved in hydrogen bonding is usually less than that for the backbone atoms; (2) in helical regions an appreciable number (28/96) of the side-chain hydrogen bonds are to backbone sites in a manner supporting the helix; (3) a similar fraction (8/43) of β side-chain hydrogen bonds act to strengthen the β sheet; (4) a detailed examination of the side-chain results shows the 4–1 bonding of a serine or threonine to the C=O of the backbone is a very common occurrence in helical regions.[10] Ser and Thr OH sites serve as proton donors much more frequently (2:1) than as proton acceptors. Curiously, at most only six "salt links" are present out of a possible 25. Three of the six involve His residues which may not be ionized in the protein near physiological pH.

To summarize: the backbone and side-chain hydrogen bonds support the classical models for helices and β sheets; hairpin turns are very poorly bonded within each turn; the (minimum) total number of unoccupied sites for the protein is high (366/869 or 42%), with most of these "empty" sites located at or near the molecular surface and presumably interacting with the solvent.

We turn to the role of hydrogen bonds in connecting secondary structural elements, particularly the connections between the helical and β regions. The possible helix–helix and helix–β bonds are given in Table III. Using our liberal β assignments there are at most 12 helix–β and five helix–helix hydrogen bonds for the entire molecule. Quiocho and Lipscomb's assignments reduce the possible helix–β bonds to four. As expected, most of the interelement hydrogen bonds involve side-chain donors and acceptors, although there is a tendency for the terminal carbonyl groups of the helices to be used. Curiously, the helices act predominately as proton acceptors to the β donors. The interelement hydrogen bonds are located primarily at the start and end of the structural regions and are near the molecular surface rather than inside. Although such placement

Table III. Interelement Hydrogen Bonds

Donor	Acceptor	Length, Å
Helix–Helix Hydrogen Bonds		
Trp-73, ε NH	Thr-119, OH	3.37
(Lys-84, NH_3^+	Gln-28, NH_2?	3.47)[a]
Ser-95, OH	Glu-302, COO^-	2.10
Tyr-259, ε NH	Asp-215, COO^-	2.79
(His-303, NH	Glu-218, COO^-	3.44)[a]
Helix–β Hydrogen Bonds		
H Tyr-19, OH	Pro-46, C=O	2.43[b]
β Lys-35, NH_3^+	Asp-23, COO^-	2.98
β Arg-45, NH^+	Asn-114, C=O	3.37[b]
β Thr-54, OH	Leu-100, C=O	2.64[b]
β Met-103, NH	Ile-99, C=O	2.93[b]
H (Ser-178, OH	Ile-38, C=O	3.40)[b,c]
β Tyr-204, OH	Asn-220, δ C=O	2.65
β Gly-235, NH	Lys-231, C=O	3.45[b]
β Thr-236, OH	Glu-292, COO^-	3.20[b]
β Tyr-265, OH	Glu-218, COO^-	2.51
β Ser-266, OH	Ser-258, OH	2.72
H (Val-305, NH	Met-103, S	3.48)[c]

[a] Both acceptors have closer alternative donors. [b] Would not be counted using β assignments of Quiocho and Lipscomb.[6] [c] Both donors have closer alternative acceptors.

probably maximizes the structural impact of these bonds, pairs of helices are connected by much less than one hydrogen bond and no helix is held to the β sheet by more than two hydrogen bonds. Remembering that some of these bonds may have unfavorable angles, their number seems insufficient to generate a high degree of internal rigidity. We note in passing that a recent analysis of the hydrogen bonding in myoglobin shows only five hydrogen bonds linking the eight helices together.[12]

We have, then, a simple picture of the role of hydrogen bonding in carboxypeptidase. Rigid structural features such as helices and β sheet presumably owe their rigidity to the cooperative hydrogen bonding present. Turns appear to be "weak links" with little internal hydrogen bonding. The tertiary structure requires the interconnection of the structural elements. The small numbers of interelement hydrogen bonds prompt us to look elsewhere for the major source of tertiary structure.

Hydrophobic Bonding. To assess the role of hydrophobic interactions in influencing protein structure, we must begin with a method which can locate and characterize those regions of carboxypeptidase with high hydrophobic character. Two approaches have been generally used to define hydrophobicity. The more common method uses the concept of a hydrophobic *residue* (*e.g.*, Leu, Ile, Phe, etc.). We will use a definition somewhat similar to that of Lee and Richards.[4] We take all side-chain carbon *atoms* (C_β and beyond)

(12) H. C. Watson, *Progr. Stereochem.*, 4, 299 (1970).

Table IV. Properties of Hydrophobic Regions in CPA

	Av per region	Total hydrophobic regions	Rest of protein	Total protein
Hydrophobic atoms	37	630	246	876
Other atoms	1	15	1546	1561
Total atoms	38	645	1792	2437
Cross section	7 Å × 10 Å			
Volume, Å^3	950	16,100	26,900	43,000[a]
Volume/atom		25.0	15.0	17.6
Density, g/cm^3 [b]		0.93	1.55	1.39

[a] Protein volume calculated from specific volume (0.720) is 41,000 Å^3. [b] See text.

as hydrophobic. Both approaches have disadvantages. The residue-based definition of necessity includes the highly polar backbone within any hydrophobic region. The atom-based definition is in difficulty with carbons α to polar terminal groups (*e.g.*, the carbonyl carbon in Gln and Asn). The latter problem can be removed by more restrictive definitions, but, for our present purposes, will not actually create any ambiguity.

To locate hydrophobic portions of carboxypeptidase, we prepared section maps (each 5-Å thick) showing the positions of all nonhydrogen atoms. The protein required only eight sections, and there is little overlapping of atoms. Examinations of the sections (*e.g.*, Figure 1) showed regions of surprising hydrophobicity. Boundaries can be drawn without severe convolutions that include no more than two polar atoms out of 100 side-chain carbons. We found 17 such regions containing 10–120 atoms in volumes of 200–3000 Å^3. The averaged properties of all the regions are given in Table IV. The larger regions are channel-like rather than circular in cross section.[13] They are located primarily in the molecular interior although several of the larger regions extend to the protein surface. There are no large hydrophobic "patches" on the surface that do not connect into the molecular interior.

We are observing the segregation of hydrophobic side chains in a manner similar to that pictured by Kauzmann.[2] The maximum dimensions occupied by any cluster of side chains is limited by the "boundaries" fixed by the polar backbone. Thus, one expects a channel of undetermined length but of cross-sectional dimension twice the length of an extended side chain (*ca.* 10 Å). In carboxypeptidase several of the channels are aligned to give a hydrophobic sheet approximately 20 × 40 × 8 Å.

We have made an estimate of the density of the hydrophobic and nonhydrophobic regions of carboxypeptidase. Using the section maps, we calculated the volume per nonhydrogen atom, and from that the atom density. To account for the hydrogens and (in the polar regions) the atomic weights of oxygen and nitrogen, we calculated apparent "atomic" weights for an average nonhydrogen atom in the polar and nonpolar side chains and backbone. These are all very close to 14.0. Using this value and the atom densities yields the following densities (g/cc): hydrophobic regions, 0.93; remainder of protein, 1.55; total protein, 1.39.

Thus the protein interior displays substantial density inhomogeneities over volumes of the order 10^3 Å^3.

(13) Note that the exact size, shape, and number of regions will depend on the manner of sectioning. However, the total volume and number of atoms involved is independent of sectioning details for "thick" sections.

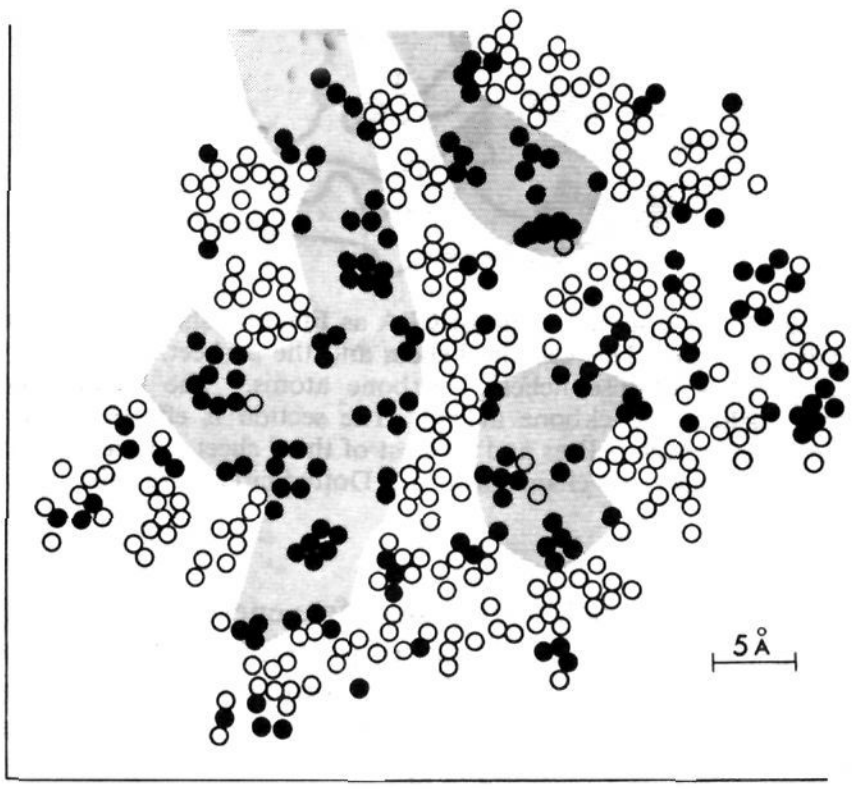

Figure 1. Section of carboxypeptidase, perpendicular to the Z axis of Quiocho and Lipscomb,[5] 5.1 Å thick, showing the X,Y locations of all nonhydrogen atoms located between Z of −8.00 and −12.00. Filled circles represent side-chain carbon atoms. The shaded areas contain a high proportion of side-chain carbons. Atom diameters are smaller than scale.

Carboxypeptidase (and presumably other globular proteins) can then be described as a relatively dense polar framework in the spaces of which are low density hydrophobic clusters. The general impression is of an irregular honeycomb.

As mentioned above, the main hydrophobic portions are found to be located in the protein interior. There is, at least in carboxypeptidase, a distinct pattern to their arrangement. Notice that the interstices between the helices and the β sheet are largely hydrophobic atoms. In the plane shown in Figure 1, for example, only 8 out of 89 atoms located between secondary structural elements are polar. Even these eight atoms tend to be at the molecular periphery and rarely break up the hydrophobic regions (Figure 2).

Discussion

We return to the question of the role of hydrogen bonding and hydrophobic bonding in maintaining protein structure. Particularly, we ask what forces contribute to the relative positioning of helical and β sheet elements in carboxypeptidase. It seems that the small number of interelement hydrogen bonds would not be able, of itself, to organize the eight helices with respect to the β sheet. Using the most optimistic estimates of

Figure 2. The same section of CPA as Figure 1 showing all nonhydrogen atoms between the helices and the β sheet. The heavy dotted lines enclose helical backbone atoms. The heavy solid lines enclose β backbone atoms. The section is effectively perpendicular to the helices and to most of the β sheet. Filled circles again represent side-chain carbons. Dotted circles are other nonhydrogen atoms.

the number of such bonds and a free energy of −2 kcal/bond yields ∼−30 kcal/mol of proteins; −10 kcal is a more likely figure. Both helices and strands of β structure have a residual dipole moment. Although the exact magnitude and direction of moment are not known, antiparallel arrangement of the structural elements would be expected from simple electrostatic considerations. The β sheet shows three antiparallel and four parallel pairings. Similarly, among the case helix–helix contacts there are four parallel and three antiparallel interactions. Assuming that the residual moment has the same sense for both helix–helix and β strands, helix–β contacts show six antiparallel and two parallel arrangements. It is not feasible to make an accurate estimate of the overall energetics without a detailed calculation. However, as a very crude model, assume that the residual moment directed along the chain was as large as 2 D.[14] Two point dipoles of this magnitude, located 7 Å apart (the closest polar–polar contacts between helices or helix and β sheet), have a potential energy of interaction of about −0.2 kcal/mol. If each interacting structure has ten such dipoles, the net interaction is more than additive, giving on the order of −8 kcal/mol per paired structure. Inasmuch as we are overestimating the interacting moments and underestimating their separation, the overall dipole contribution to the stabilization of the protein from the eight helices and the β sheet is probably not more than a few tens of kilocalories. The large number of hydrophobic contacts (of the order of several hundred) and the moderate lowering of free energy that occurs when they are placed in contact in an aqueous medium suggests that hydrophobic bonding may contribute on the order of 100 kcal/mol of protein and hence may dominate the interelement interactions.

(14) The dipole moment of the peptide group is approximately 3.7 D. It points roughly along the C=O bond so that the residual moment will depend upon the chain propagation direction for a particular secondary structure.[10]

The lack of strongly directional forces such as hydrogen bonding and the low density of the hydrophobic regions poses the problem of how to generate internal rigidity in the globular proteins. Regardless of the aptness of "oily" or "waxy" descriptions, *per se*, the existence of greatly restricted mobility of all types of (internal) side chains is clear from both X-ray and nmr data. Klapper[5] suggests that the rigidity arises from a very high and unusual packing density. We find that the hydrophobic regions are somewhat denser than liquid hexane but about as dense as liquid benzene. Density alone cannot be used to infer molecular mobility or molecular order, as the case of water and ice dramatically illustrates. The source of the internal rigidity probably lies in the constraint of constant (and minimum) volume imposed by the solvent. Thus, any internal motion must be highly cooperative and hence possess a high activation barrier even in the absence of highly directional specific bonding. An every-day example is the difficulty in leaving a crowded elevator unless someone in front steps out first. Perhaps the most striking aspect of the lack of a heavily interconnected tertiary structure is the possibility of moderately large internal slippage of the secondary elements with respect to each other. Such rearrangements are expected for multisubunit systems, but somewhat similar changes may be possible in molecules the size of carboxypeptidase.

In sum, it appears that hydrophobic bonding models[2,15] provide a satisfactory rationalization of the relative positions of the helices and the β sheet in carboxypeptidase. The helices turn to direct hydrophilic side chains to the molecular surface.[15] The hydrophobic side chains come together in long channels which join with a hydrophobic plane from one side of the β sheet. The other side of the sheet contains primarily polar side chains which interact with a region containing many hairpin turns. Those hydrophobic groups on the protein surface are normally connected with the deeper lying channels and may serve a role in protein–protein interaction.

Acknowledgment. Partial support by the Academic Senate of the University of California (Grant No. 2-509942-35488) and the National Institutes of Health (GM 19269) is gratefully acknowledged.

(15) M. Schiffer and A. B. Edmundson, *Biophys. J.*, 7, 121 (1967).