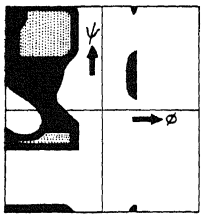

Special issue:
The Ramachandran map and
protein and peptide structures

Proteins: The hard sphere, structure and energetics

Raghavan Varadarajan, Frederic M. Richards* and Patrick R. Connelly†
Department of Molecular Biophysics and Biochemistry and
†Department of Chemistry, Yale University, 260, Whitney Avenue, P. O. Box 6666, New Haven, Connecticut 06511, USA

The Ramachandran map was derived by treating atoms as hard spheres and restricting allowed conformations so as to avoid steric overlap between non-bonded atoms. The high packing density characteristic of most solids and reflected in individual protein molecules can be anticipated in terms of hard sphere atoms. The constraints of high packing density and avoidance of steric overlap have been used to develop algorithms which attempt to predict sets of sequences consistent with a given protein main-chain architecture. However hard sphere models cannot predict the energetics of specific interactions within the protein–aqueous solvent system. Accurate and detailed measurements of such interactions coupled with high resolution structural information are essential in understanding the relationship between protein sequence and structure. We describe some recent results of such studies in the ribonuclease-S system.

IN the proper solvent a polypeptide chain with an appropriate amino acid sequence will fold up into a compact molecule with a unique three-dimensional structure and biological function. A complete understanding of this process, leading to predictions of a protein's three dimensional structure from its sequence, is far from being attained. Even if the folded protein structure represents the global minimum in free energy for the protein–aqueous solvent system, no feasible procedure for calculation of this minimum presently exists. In part this is because proteins are only marginally stable, having free energies of folding of the order of 10 kcal/mol. This small number is the difference between two large numbers, the free energies of the folded and unfolded states. For the fully unfolded chain there are an enormous number of possible conformations, in each of which every part of the polypeptide is in contact with solvent. The area of the solvent–protein interface is dramatically reduced in the compact folded conformation, and there are only a few solvent molecules in the inside of most proteins. This interior space is well filled with atoms of the peptide chain.

*For correspondence.

Although it is obvious that folding leads to a large decrease in the number of possible conformations, it was the work of G. N. Ramachandran and co-workers that showed that one could represent all of the broad features of the conformation of a polypeptide chain simply by specifying the restrictions on the two torsional angles involving the main chain single bonds of each amino acid residue[1]. The earlier work of Pauling and Corey provided the covalent geometry of the bonded atoms[2, 3]. Each atom was then assigned a radius as a hard sphere. A single restriction, the avoidance of steric overlap of non-bonded atoms, produced the map which defines all permitted conformations of the peptide chain. The repetitive structures, the helices and the extended beta strands, and all non-repetitive conformations, the various turns and loops no matter how convoluted, must fit this map.

In one sense the success of such an approach is remarkable. Atoms, of course, are not hard spheres. The electron clouds of neighboring atoms can interpenetrate a significant distance. The electrons in molecules are not distributed with spherical symmetry around the atom centers, and the chemical behavior of the atom is not isotropic. Nonetheless, the hard sphere approximation has been extremely useful in theoretical chemistry and has led to models which can mimic molecular properties remarkably well in many instances. Further work, following the lead of the Ramachandran group, has refined the map on the basis of detailed quantum mechanical calculations[4], but the basic conclusions represented by the shape of the allowed conformational space have not been altered. The importance of this map cannot be overestimated. It has found widespread use, even today, in essentially its original form. The avoidance of steric overlap between non-bonded atoms was and is the message.

The tendency of non-polar amino acid side chains to pack together in the protein interior where contact with water is avoided (the hydrophobic effect) is generally regarded as a major force leading to compact protein

molecules[5, 6]. The similar tendency of hydrocarbons and water to form separate phases when mixed has led many to regard this phenomenon as a model for the behavior of nonpolar residues in the protein folding process. The physical origins of these phenomena are the subject of intense debate at this time[7, 8]. However, even without knowing these origins, the above observations can be described semi-quantitatively in terms of the minimization of the interfacial area between the molecules of the two components[9].

Some years after the appearance of the Ramachandran map two algorithms were proposed for calculating the surface area of macromolecules of known structure by numerical integration[10, 11]. As in the Ramachandran map, the hard sphere approximation represented by the van der Waals envelope of the molecule was the starting point. (Subsequently, analytical solutions for calculating of the area of intersecting spheres have been provided by Richmond[12] and by Connolly[13].) At the level of atomic resolution, the meaning of 'surface' and 'area' is not always immediately obvious. The accessible surface and the molecular surface of a macromolecule are two slightly different geometrical constructs (see definitions provided by Richards[14]). Either one may be used to represent the interface between the protein and the solvent in which it is dissolved. In most cases, the solvent is water whose molecules are taken to be spheres with a radius of 1.4 or 1.5 Å. Although bulk water is normally considered to be a fluid whose structure is not controlled by packing considerations due to the overwhelming importance of the tetrahedral hydrogen bonding, the van der Waals envelope of the water molecule is just as real as that of any other molecule and will play a role in interactions with dissolved solutes[15]. The concept of surface area is useful as a single parameter representation of the hydrocarbon–water interface, and it requires no detailed knowledge of the water structure. Detailed structural effects are smoothed out in the averaging process. The surface of a protein, however, is not that of a hydrocarbon. It is a patchwork of small polar and non-polar regions all in contact with water. The polar and charged atoms interact with water in very specific ways not easily describable by a 'surface' formalism. The interaction energetics will depend on the specific geometry of the interface in each case[16]. The free energy change for the transfer of non-polar molecules from a hydrocarbon to aqueous solvent is frequently taken as 20–25 cal/mol/Å^2 of accessible surface area[14]. Due to its nature as an average, this value can be expected to become less and less certain as the total size and shape of the relevant 'area' decreases. Its use in the binding of small ligands or in interpreting the effects of small surface changes must be considered very uncertain.

In the solid state, molecules will tend to pack into as dense an arrangement as possible in order to maximize the non-polar attractive forces. Even when special factors such as charge interactions or hydrogen bonds play a role, close packing is usually observed in molecular crystals. Occasionally this is assisted by the incorporation of solvent molecules which can be either water or organic species. The packing arrangement will depend on the molecular shape, but large empty cavities are assiduously avoided[17, 18].

These same principles seem to apply to individual globular protein molecules, macromolecules which have a low axial ratio and are large enough to have a defineable 'inside'. It has been known for many years that the unfolding of proteins into extended, solvent-surrounded conformations is accompanied by very small percentage changes in volume (see for example, Holcomb and Van Holde[19]). This was taken as initial evidence that there were no large empty cavities in the folded structures. The overall size and shape were also known to be compatible only with a dense relatively solvent free structure. The precise amount of internal solvent can only be determined in individual cases where high resolution X-ray structures are available. To date the amount of internal solvent found is generally small but structurally important where it occurs. Chothia[9], analysing some of the then known protein structures, showed that the average volume occupied by interior residue side chains was closely equivalent to that found for the same groups in small molecule crystals, confirming the apparent overall packing density.

The high packing density observed in proteins appears to be the result of the interplay of two competing criteria: the efficient filling of space and the avoidance of steric overlap. Ponder and Richards[20] have suggested that the attainment of effective packing in the protein interior may be the controlling factor in deciding which folded conformation a peptide chain of defined sequence will actually assume.

This latter study started with a re-examination of the structures of some 17 proteins which had been refined to high resolution. In agreement with earlier studies[21], the conformations of the residue side chains were found to fit quite accurately the staggered positions expected for the single bond dihedral angles. The standard deviations were substantially smaller than those found earlier in the less well refined structures. A corrollary to this observation is that a protein does not 'store' energy in the form of distorted side chains to any significant extent. Hence it appeared reasonable to model amino-acid side chain conformations in proteins using a set of rotamers derived from examination of a set of high resolution protein crystal structures.

Using this rotamer approximation, interior packing units were examined. A packing unit is taken to be a roughly spherical interior region of protein which is filled with the side chains of half a dozen or so residues.

In general there is no obvious positional relationship of these residues to each other in the linear sequence of the peptide chain. One can then ask the question: How many sequences are there for the residues in the packing unit which will pack equally well within specified limits? (This is a version of the 'inverse folding problem' first suggested by Drexler[22].) In an attempt to answer this question an algorithm was developed. The procedure required that the conformations of the main chain and all other residues be left in their native positions. The packing unit residues are removed back to the beta carbon atoms and the rotamer library is then surveyed to refill the space with all possible sets of rotamers. In general there is no unique answer to this packing puzzle. There are a relatively small number of possible sequences, but this number is very sensitive to the choice of the minimum acceptable packing density which is specified as an input parameter.

In principle one should try to repack the entire protein at the same time since only the overall packing density can be specified with any degree of confidence. Unfortunately, because of the combinatorial nature of the problem, the answer for the whole protein is computationally inaccessible. Additionally, the packing density can vary quite a lot from one part of a protein to another; the main chain can alter its conformation to some extent and still be considered to represent 'the same structure'; and the possibility of internal solvent must be considered. All of these factors should be considered in a general calculation. They will tend to increase the number of sequences which would be considered acceptable. However, even this number, large as it would be, would be a tiny fraction of all of the combinatorially possible sequences for a peptide of a given length. The answer to the relation between sequence and three dimensional structure could be contained in this packing proposal, but it has not been possible to prove it by computation up to this point.

Experiments directed at testing the packing hypothesis are essential, and they are underway in several laboratories. The most extensive data currently available is from Sauer's laboratory at MIT in experiments on the lambda repressor system. The structure of the N-terminal fragment of this protein, which contains the operator binding region, has been determined by Pabo and Lewis[23]. A large number of mutants have been made involving the residues of some internal packing units. A report on some of the early phases of this work has appeared[24]. Random mutagenesis of a three residue packing unit led to about 110 sequences that provided biologically active molecules. This is about 1.4% of the 8000 possible sequences. Further studies have isolated a restricted set of both active and inactive sequences. An initial comparison of these mutants with the calculated list of 'acceptable' sequences for the same set has indicated an encouraging level of agreement (Lim and Ponder, unpublished results). However there are discrepancies, and these should lead to refinement of the packing criteria. One should note that the current algorithm allows for no main chain flexibility and thus the derived list of permitted sequences is too restrictive. The extent of main chain movement that should be allowed is not known and computational procedures to efficiently account for such motion have not yet been developed.

With regard to permitted sets of interior residues, two major questions are: (i) 'how large a cavity can be tolerated', and (ii) 'to what extent a cavity can be accommodated by rearrangements of the surrounding protein or incorporation of solvent to fill the space?' In general the answers to these questions will differ from protein to protein and may change for different interior sites within a given protein. In particular the range of allowed substitutions will depend strongly on the free energy of folding of the protein under the assay conditions used to assess success. It is therefore important to obtain quantitative measures of thermodynamic changes and structural responses associated with cavity formation in proteins.

In recent years there have been a number of detailed studies on the effects of amino acid substitutions on protein structure and stability[25–28]. A number of other systems are poised to produce similar data, including *E. coli* thioredoxin in the authors' laboratory. In such studies, one proceeds by disrupting mutant and wild-type protein structures in solution by means of increasing temperature or concentration of denaturant, while following the reaction process optically or calorimetrically. Subsequently, one extracts the various thermodynamic parameters characterizing the disruption process and extrapolates the extracted parameters to a particular set of reference conditions to reveal the energetic differences in stability. However, it has been difficult to obtain quantitative values for specific interactions from such measurements since one cannot ascribe the relative energetic contributions of the folded and unfolded states to the net stability differences. In addition, interpretation of such studies is often complicated by factors such as irreversible denaturation, possible presence of folding intermediates, and the need for making large extrapolations (to zero denaturant concentration, neutral pH, or a reference temperature). We have avoided these complications in our approach by investigating the effects which specific substitutions of buried hydrophobic residues have on the simple binding reactions between S-protein and various S-peptide analogues.

Bovine pancreatic ribonuclease-A (RNase-A) may be cleaved with subtilisin at the peptide bond between residues 20 and 21 to give the S-protein and the N terminal S-peptide[29]. These fragments may be separated,

and can be reconstituted to give rise to the product ribonuclease-S (RNase-S) that has a structure very similar to that of RNase-A[30]. Residues 3–13 of the peptide portion of RNase-S form an α-helix, just as these residues do in uncleaved RNase-A. Residues 16–20 are evidently not important for binding, and are not clearly defined in the crystal structure of RNase-S[30]. The hydrophobic residues in S-peptide thought to be particularly important for binding are Met-13 and Phe-8 (refs. 31, 32). In order to study interactions among buried hydrophobic groups, peptides have been synthesized in which the methionine at position 13 has been replaced by seven other hydrophobic amino acids (glycine, alanine, α-amino-*n*-butyric acid, valine, leucine, isoleucine, phenylalanine). For this series the parent peptide, S15, corresponds to the first 15 residues of S-peptide and has an amidated C-terminus. Figure 1 shows the location of residue 13 in a ribbon diagram of the S15:S-protein complex. The thermodynamic properties of the reactions of the various peptide analogues with S-protein have been determined by titration microcalorimetry[33].

Probing the nature of stabilizing interactions in this manner can be done isothermally, without the use of denaturants, and in a simple 1:1, well defined reaction. Changes in the structure of the unbound peptide (the equivalent of the denatured state) can be monitored by circular dichroism. Additionally, seven of the peptide protein complexes have been crystallized and the structures are currently being determined by X-ray crystallographic analysis.

Table 1 gives the values of the different thermodynamic parameters calculated relative to the S15·peptide so that for $J=G$, H, or S, $\Delta\Delta J_x=\Delta J_{M13x}-\Delta J_{S15}$ where X is the single letter code for the substituted amino acid. ANB is used to denote α-amino-*n*-butyric acid. The differences in enthalpies, free energies and entropies do not change appreciably with pH, although the absolute quantities do, as has been reported previously[34]. From the data for M13G it appears that removal of the set of hydrophobic contacts due to a single side chain can result in a destabilization of about 5 kcal/mol. This is comparable to the free energy of folding of many globular proteins and emphasizes just how marginally stable most proteins are.

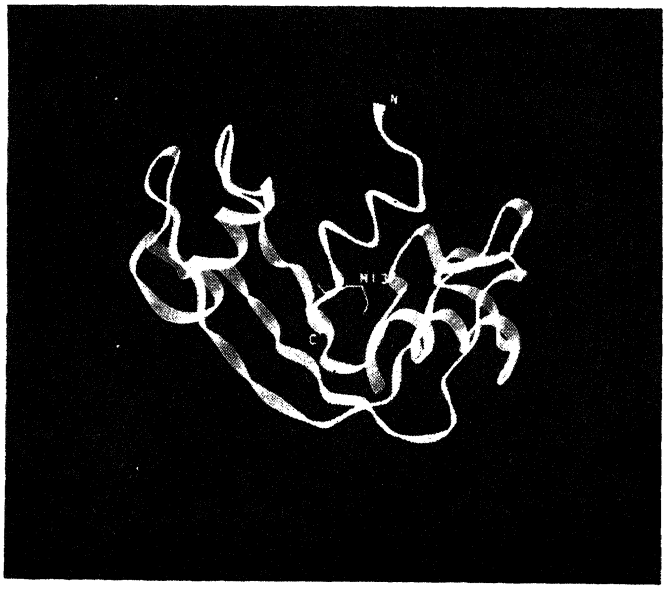

Figure 1. Location of residue 13 in a ribbon diagram of the S15:S-protein complex.

Table 1. Difference thermodynamic parameters at 25°C for interaction of S-peptide analogues with S-protein relative to S-peptide/S-protein interaction.*

	$\Delta\Delta G^\circ$ kcal/mol		$\Delta\Delta H$ kcal/mol		$T\Delta\Delta S^\circ$ kcal/mol	
Peptide	pH6	pH5	pH6	pH5	pH6	pH5
M13G	5.0	—	(−1.3)	—	(−6.3)	—
M13A	4.1	—	4.7	—	0.6	—
M13ANB	1.6	1.6	7.2	7.1	5.5	5.5
M13V	0.3	−0.1	3.2	2.8	2.8	2.9
M13I	0.2	—	4.8	—	4.6	—
M13L	0.6	0.2	5.2	6.0	4.6	5.8
M13F	2.7	2.5	3.5	4.5	0.8	2.0

*Average errors are ±0.2 for $\Delta\Delta G$; ±0.8 for $\Delta\Delta H$; and ±0.9 for $T\Delta\Delta S^\circ$ (Connelly *et al.*[33]).

Although there is some qualitative correlation between residue size and binding affinity, none of the thermodynamic parameters listed in Table 1 correlate quantitatively with any simple property of the substituent amino acid. The energetics of packing interactions of residue 13 with its surroundings are highly distance dependent. Such interactions would not generally be expected to scale simply with the surface area or hydrophobicity of the buried residue. They will depend on the detailed structural changes in both protein and aqueous solvent caused by the amino acid substitution. In several instances there appear to be compensating changes in the enthalpy and entropy of binding. Note particularly the cases of S15, M13L and M13I which differ solely in the relative placement of a methyl group along the side chain. The $\Delta\Delta G$'s are all very similar but the $\Delta\Delta H$'s are not. This underscores the danger of interpretations based solely on measurements of free energy changes and emphasizes the need for obtaining data on as many thermodynamic parameters as possible.

We have shown[33] that the relative stabilities are governed by the difference in the energetics of taking the residue in the isolated peptide helix from the aqueous solution and placing it in contact with the hydrophobic core of the protein. The observed thermodynamic differences may thus be broken into three components: (a) changes in the hydration of residue 13 in the free peptides relative to S15, (b) changes in the hydration of residue 13 in the complexed peptides relative to the S15 complex, and (c) changes in the

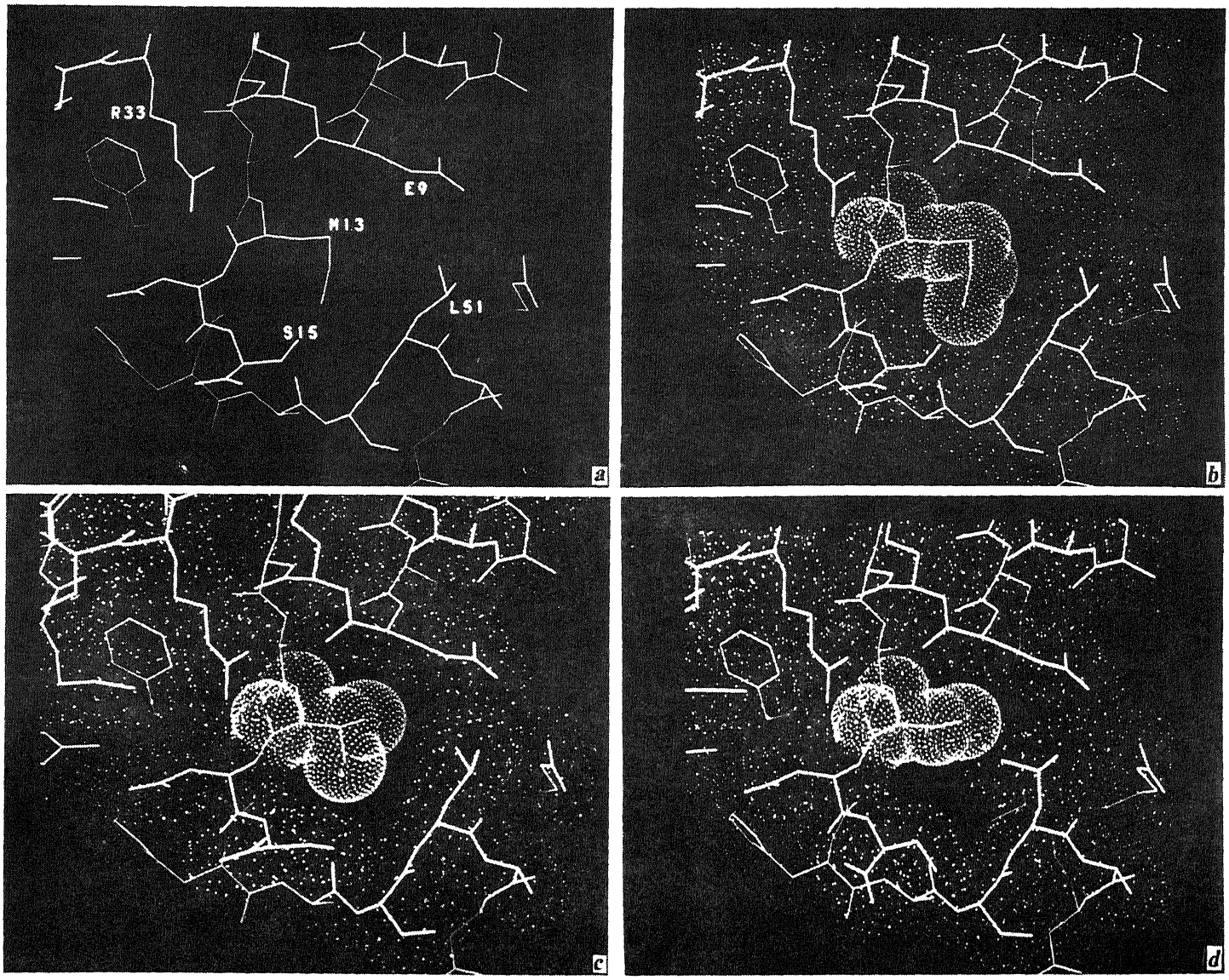

Figure 2. *a*, A slice through the region surrounding residue 13 in the complex of S15 with S-protein. All labels are positioned near the gamma position of the corresponding side chain. ***b–d,*** van der Waals surfaces of residues in the complexes of (***b***) S15, (*c*) M13V, and (*d*) M13ANB with S-protein shown in the same orientation as in part *a*. In each case residue 13 is highlighted by using a dot density ten times that used for the rest of the protein. Note the movement of L51 in the latter two cases. The density for residue 15 is not clearly defined past the alpha carbon atom at the present level of refinement so the precise orientations of this side chain and the C-terminal amide group of the peptide are currently not known.

packing interactions of residue 13 with the hydrophobic core of the protein in the various complexes.

At the present time there is no generally accepted method for evaluating term (a) although a recent analysis[7] suggests that it is relatively small. However, we hope that detailed structural studies coupled with theoretical calculations will provide estimates of the latter two terms in the various protein–peptide complexes.

X-ray crystallographic data have been collected for the S15, M13ANB, M13V, M13I, M13L and M13F complexes to a resolution of 1.7–1.8 Å. The *R*-factors at current levels of refinement are between 18 and 19%. Shown in Figure 2 are residues in the environment of residue 13 in the S15, M13V and M13ANB complexes. There do not appear to be large changes in hydration in the immediate environment of residue 13 in these complexes. Furthermore, although there are small rearrangements of the rest of the protein around the site of substitution, these do not result in filling of the cavities created by substitution of Met by V and ANB. The side chain of M13ANB appears to be restricted to a single well defined conformation having a similar position to the gamma 2 methyl group in the M13V complex. Another interesting observation is that cavity formation in M13V is not accompanied by a noticeable change in the free energy of binding. While the largest changes in position relative to the S15 complex occur at L51, it is possible that the observed changes in binding thermodynamics have an appreciable contribution from small movements of several residues that occur

throughout the protein. A more quantitative analysis will be carried out once refinement of the structures of all the complexes is complete.

It is thus clear that a quantitative understanding of the forces that stabilize folded proteins, even in relatively simple systems, is a formidable task. In the past decade there have been substantial technological advances that have greatly enhanced the rate and accuracy with which protein structural and thermodynamic data can be acquired. Large increases in available computational power have also facilitated theoretical treatment of such data. All these advances notwithstanding, it is important to note that the underlying principles on which the Ramachandran map was based are still widely used and without serious challenge.

1. Ramachandran, G. N. and Sasisekharan, V., *Adv. Protein Chem.*, 1968, **23**, 283.
2. Pauling, L. and Corey, R. B., *Proc. Natl. Acad. Sci. USA*, 1951, **37**, 235.
3. Marsh, R. E. and Donohue, J., *Adv. Protein Chem.*, 1967, **22**, 235.
4. Pullman, B. and Pullman, A., *Adv. Protein Chem.*, 1974, **28**, 347.
5. Kauzmann, W., *Adv. Protein Chem.*, 1959, **14**, 763.
6. Tanford, C., *The Hydrophobic Effect*, Wiley, New York, 1980.
7. Privalov, P. L. and Gill, S. J., *Adv. Protein Chem.*, 1988, **39**, 191.
8. Muller, N., *Acc. Chem. Res.*, 1990, **23**, 23.
9. Chothia, C., *Nature*, 1975, **254**, 304.
10. Lee, B. and Richards, F. M., *J. Mol. Biol.*, 1971, **55**, 379.
11. Shrake, A. and Rupley, J. A., *J. Mol. Biol.*, 1973, **79**, 351.
12. Richmond, T. J., *J. Mol. Biol.*, 1984, **178**, 63.
13. Connolly, M., *J. Appl. Crystallogr.*, 1983, **16**, 548.
14. Richards, F. M., *Annu. Rev. Biophys. Bioeng.*, 1977, **6**, 177.
15. Savage, H. F. J. and Finney, J. L., *Nature*, 1986, **322**, 717.
16. Finney, J., *Philos. Trans. R. Soc. London*, 1977, **B278**, 3.
17. Kitaigorodsky, A. I., *Order and Disorder in the World of Atoms*, Springer-Verlag, New York, 1967.
18. Pertsin, A. J. and Kitaigorodsky, A. I., *The Atom-Atom Potential Method, Applications to Organic Molecular Solids*, Springer-Verlag, New York, 1987.
19. Holcomb, D. N. and Van Holde, K. E., *J. Phys. Chem.*, 1962, **66**, 1999.
20. Ponder, J. W. and Richards, F. M., *J. Mol. Biol.*, 1987, **193**, 775.
21. Janin, J., Wodak, S., Levitt, M. and Maigret, B., *J. Mol. Biol.*, 1978, **125**, 357.
22. Drexler, K. E., *Proc. Natl. Acad. Sci. USA*, 1981, **78**, 5275.
23. Pabo, C. O. and Lewis, M., *Nature*, 1982, **298**, 443.
24. Lim, W. A. and Sauer, R. T., *Nature*, 1989, **339**, 31.
25. Shortle, D. and Meeker, A. K., *Proteins*, 1986, **1**, 81.
26. Kellis, J. T., Nyberg, K., Sali, D. and Fresht, A., *Nature*, 1988, **333**, 784.
27. Matsumura, M., Becktel, W. J. and Matthews, B. W., *Nature*, 1988, **334**, 406.
28. Yutani, K., Ogasahara, K., Tsujita, T. and Sugino, Y., *Proc. Natl. Acad. Sci. USA*, 1987, **84**, 4441.
29. Richards, F. M. and Vithyathil, P. J., *J. Biol. Chem.*, 1959, **234**, 1459.
30. Wyckoff, H. W., Tsernoglu, D., Hanson, A. W., Knox, J. R., Lee, B. and Richards, F. M., *J. Biol. Chem.*, 1970, **245**, 305.
31. Hearn, R. P., Richards, F., Sturtevant, J. M. and Watt, G. D., *Biochemistry*, 1971, **10**, 806.
32. Scoffone, E., Rocchi, R., Marchiori, F., Moroder, L., Marzotto, A. and Tamburro, A., *J. Am. Chem. Soc.*, 1967, **89**, 5450.
33. Connelly, P. R., Varadarajan, R., Sturtevant, J. M. and Richards, F. M., *Biochemistry*, 1990, (in press).
34. Schreier, A. A. and Baldwin, R. L., *J. Mol. Biol.*, 1976, **105**, 409.

ACKNOWLEDGEMENTS. We thank Drs J. W. Ponder and W. A. Lim for discussions of their unpublished results and Dr J. M. Sturtevant for his comments and support. We acknowledge the support of the National Institute of General Medical Sciences, Grant Number GM-22778 (F. M. Richards) and National Science Foundation, Grant Number PCM-8417341 (J. M. Sturtevant). R. Varadarajan is supported by a postdoctoral fellowship from the Damon Runyon-Walter Winchell Cancer Fund.