CMLS, Cell. mol. life sci. 53 (1997) 790–802
1420-682X/97/100790-13 $ 1.50 + 0.20/0

CMLS Cellular and Molecular Life Sciences

Protein stability: still an unsolved problem[1]

F. M. Richards

Department of Molecular Biophysics and Biochemistry, Yale University, New Haven (Connecticut 06520, USA)

Abstract. This brief review suggests that molecular packing, the efficient filling of space, may be the most generally applicable factor that leads to the unique structures of most globular proteins. While simple in concept, the details of packing can lead to very subtle effects. The mechanical properties of a protein, dynamics and deformations under stress, tend to be asymmetric. In terms of structural alterations and thermostability, responses to genetic mutations are context dependent and remain difficult to predict with any confidence. Through small shifts proteins can frequently accommodate major changes in composition of the core region without substantial alteration in the basic chain conformation. Extending a jigsaw puzzle analogy, all of the pieces (side chains) are convex, varying flexible, and cannot be packed together without leaving cavities. Although large cavities do occasionally occur, a relatively even distribution of empty space is more common, and the overall packing does seem to specify the unique native structure. While it might appear that the translation machinery of the cell could have been designed with any set of α amino acids, the packing requirements, while strong, must be flexible enough to permit nondestructive single site mutations. This flexibility, combined with the need to produce a unique structure, may limit the average number of allowed side chain rotamers per residue. This in turn will reduce the allowable asymmetry of the side chains in order to maintain the largest number of structural motifs. It may be hard to improve on current set of amino acids.

Key words. Protein structure; packing; cavities and stability.

Introduction

At the time of Kaj Linderstrøm-Lang's birth in November 1896 the existence of substances called proteins had been known for some time, but the definitions used to describe them were qualitative at best. Proteins tended to coagulate with heat; they would move in an electric field in a direction which depended on whether the solution was acidic or basic (Sorensen[2] had not yet invented the term 'pH'); they contained nitrogen (by the Kjeldahl[2] procedure); some 10 different α amino acids had been isolated from proteins, but compositional data was extremely crude, and there was no information at all on how many types of amino acids there were; both the proposed macromolecular nature of these substances and the significance of the peptide bond were subjects of major arguments. The defining experiments based largely on physical chemistry had not yet begun. The early history of protein chemistry is provided in the fascinating book by J. S. Fruton [1].

The six decades of Lang's life carried him through the major developments in protein chemistry, developments in which he played a leading role. One of the saddest aspects is that he could only have just known about the structure of myoglobin, produced by the first successful interpretation of the single crystal X-ray diffraction pattern of a protein, which was reported in 1959, the year of his untimely death. The validation of his own work and ideas was just about to begin.

During the past century the amount of detailed information about proteins has been increasing and today

[1] This manuscript was produced from memory some months after the symposium was held. It corresponds roughly, but undoubtedly not accurately, to what was said in the lecture. It is not intended as a definitive review.

[2] Kjeldahl was Head of the Chemical Department of the Carlsberg Laboratory from 1876 to 1900, followed by S. P. L. Sorensen from 1901 to 1938 who, in turn, was succeeded by K. U. Linderstrøm-Lang from 1939 to 1959.

can only be described as a torrent. Our ability to assimilate this vast amount of data has not kept pace. Each group of closely related proteins is represented not only by a large number of individual papers but by its own set of specialist journals. From only the most cursory survey of this vast literature it would appear that there are very few general statements about the properties of proteins that are always true.

In principle, protein structures and properties, as with all other forms of matter, can be derived by the proper application of quantum mechanics. However, the number of atoms in the smallest protein renders such calculations, ab initio, impossible at present. At best we can deal at the quantum level with the properties of a small number of atoms, immersed in a larger group of atoms treated classically, further surrounded by an outer shell treated as a continuum. It is very difficult to use such a procedure to predict the properties of an entire macromolecule. The fallback position is to treat the entire system classically making as much use as possible of the entire field of small molecule chemistry, and the statistical mechanics of polymer systems. Even here the calculational aspects can be formidable, and more qualitative considerations have proved useful.

For biologically relevant solvent conditions, the free energy difference between the thermodynamic state of the native protein (active) and non-native state(s) (inactive[3]) is of the order of 10 to 15 kcal mol^{-1}, more or less independent of the size of the protein, with the native form being energetically favoured. The enthalpic and entropic energy differences between the two states are in the range of hundreds of kcal mol^{-1}. With the free energy as the small difference between these large numbers, a standard problem in physics, protein stability always appears to be marginal. The prediction of this small difference from first principles is central to the eventual solution of 'the protein folding problem' and its corollary protein stability.

Input parameters

The general results from small molecule chemistry, outlined in table 1, can be applied to macromolecules essentially unchanged. The parameters in the potential functions related to bond lengths, bond angles and torsional angles are derived from small molecule data because the quality of the structural data from the macromolecules themselves is usually not good enough to merit use for any further refinement of these values. On the other hand the nonbonding parameters may be improved by such refinements since there is no unambiguous small molecule reference state for the protein interior. Almost invariably the potential functions contain only pairwise terms. The relative importance for current qualitative considerations of higher order interactions, reflecting such properties as polarizability, has not yet been definitively established.

Added to these data are estimates of effects which are not necessarily new in kind but which are expected to be amplified in importance due to the large number of atoms in even a small protein. The introduction of solvent is very much a research area with currently no general agreement on the best approach to the analysis of the solvent-protein boundary, arguably the region of greatest relevance to biological function.

Proteins as macromolecules

As the molecules increase in size and have larger and larger numbers of formal charges, special attention must be given to the electrostatic effects of the long-range coulombic interactions between the ionizable groups. There are instances where very large structural changes are produced by modest shifts in pH. For the hemagglutinin from the influenza virus, for example, relative positional changes within the molecule of as much as 100 Å are observed and the secondary structure is altered (fig. 1A). The small pH change inducing these shifts reflects alterations in the state of just a few ionizable groups. The biological behaviour of the molecule in causing membrane fusion is dramatically altered by this change in structure. In this case the change is irreversible and suggests a metastable state for the starting structure.

Table 1. General parameters and forces.

Taken over from data on small molecules	Factors of increasing importance for larger molecules
Covalent geometry	macromolecule-solvent interface
Bond lengths	(solvent accessible surface)
Bond angles	
Torsion angles	hydration layers outside of interface
Hydrogen bond geometry	electrostatic interactions (single charges/dipoles)
Atom size-van der Waals' radii	packing density (excluded volume/cavities)
Group charge distributions	

[3] In earlier days the 'inactive' form was called the 'denatured' state. Today, to avoid the somewhat pejorative aspect of this term, the same state is more commonly described as 'unfolded'. This appears to be intended to make the studies of such systems more acceptable as both interesting and fundable science.

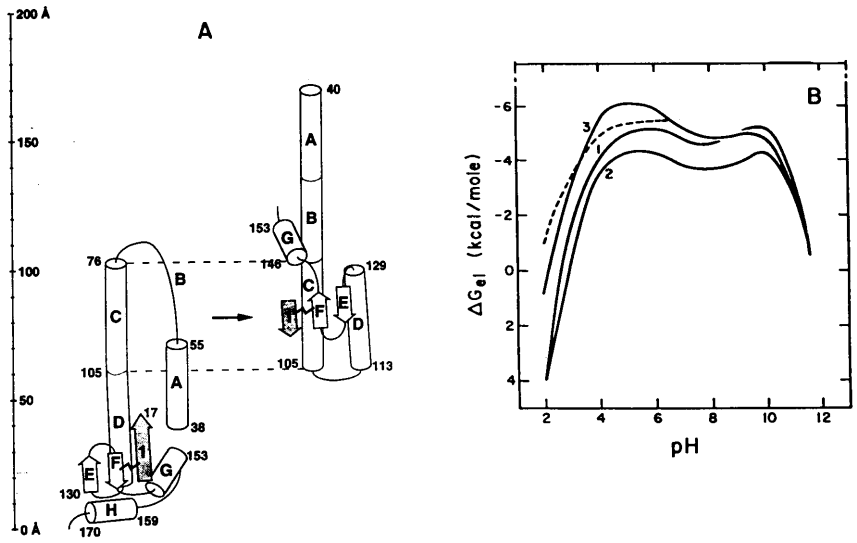

Figure 1. Electrostatic effects on protein structure. (*A*) The pH-induced change in the structure of a portion of the hemagglutinin from influenza virus, HA. Bromelain cleavage at neutral pH produces a soluble trimer one monomer of which is the reference structure shown in the left of panel A. Acidification and treatment successively with trypsin and thermolysin yields a soluble acid form trimer, a monomer of which is shown on the right of panel A. The dramatic changes in both secondary and tertiary structure caused by the pH change are clear in the comparison. Parts of the molecule are shifted in relative position by as much as 100 Å. (Adapted and reprinted with permission from: Bullough P. A., Hughson F. M., Skehel J. J. and Wiley D. C. (1994) Structure of influenza hemagglutinin at the pH of membrane fusion. Nature **371:** 37–43, © 1997 Macmillan Magazines Ltd.) (*B*) The effect of pH on the free energy of stabilization: calculated values showing the pH dependence of the electrostatic contribution to the free energy of stabilization of ribonuclease-S are given by the solid lines: 1, ionic strength = I = 0.01, no specific ion binding; 2, I = 0.15, no anion binding; 3, I = 0.15 with anion binding. The net charge changes from 0 at the isoelectric point near pH 10 to about +14 at pH 4 with no significant change in the stability of the enzyme. The dashed line is derived from experimental data on ribonuclease A collected between pH 7 and 2 and an ionic strength of 0.16. (Adapted and reprinted with permission from: Matthew J. B. and Richards F. M. (1982) Anion binding and pH dependent electrostatic effects in ribonuclease. Biochemistry **21:** 4989–4999, © 1997 American Chemical Society).

Among proteins whose 3D structure is known such changes are relatively rare. Many, perhaps most, of this group are stable over a wide pH range where the net charge varies substantially (fig. 1B). While pH usually affects biological activity, the charge changes may have little or no effect on the overall geometry of the molecule. It would thus appear that the coulombic interaction of ionizable groups is not universally important in specifying the basic structure of a protein.

One universal law does apply to proteins: no two atoms can occupy the same space at the same time. In the absence of special interactions, the van der Waals attraction of identical molecules in the crystalline state strongly favours packing with a coordination number of 12, the densest packing that is available for any molecular shape (see Kitaigorodski, [refs 5, 6]). We may picture a protein molecule as a small solid sample which is subject to this same constraint. However, the residues in the protein are neither identical nor independent due to the covalent connectivity of the polypeptide chain. Thus the symmetry implied by the coordination number of 12 found in molecular crystals of small molecules is not found in the interior of a protein. However, the efficient filling of space might still be expected in order to maximize the van der Waals stabilization energy. There is considerable evidence that this is, in fact, the case. There are relatively few water molecules in the interior of the protein (i.e. the space is well filled with protein atoms). There are very few atom-sized interior voids. The inside of a protein has a unique structure and thus resembles a solid rather than a dense liquid [4].

Molecular packing

For ease of visualization consider the example of two-dimensional packing shown in figure 2. It is clear from this simple case that the number of possible 'structures' is a very sensitive function of the packing fraction (fig. 3A). With dense packing only one structure is possible. With only a modest loosening of the packing all structures are possible. In the first case we could say that the structure is controlled by the packing requirement (filling space without overlap). In the second situation, whatever unique structure might be present, it could not be ascribed to the effects of packing at all. Where do real proteins fit into such a picture?

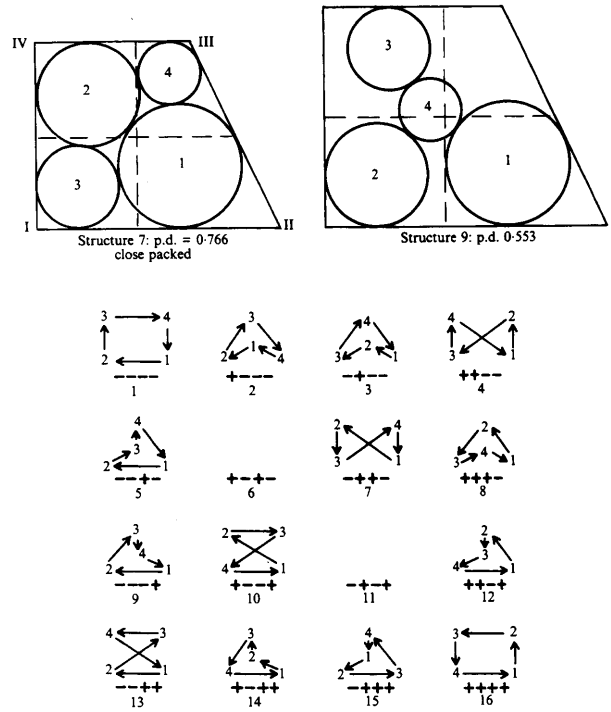

Figure 2. A two-dimensional example of packing. In the top left panel a trapezoid is drawn just to enclose four circles of differing size arranged as shown. There is no other way to place these particular circles in this defined space. The 'structure' of this assembly of circles is unique. A 'structure' is defined as the list of the cross products (+ or −) of the centre to centre vectors taken serially, $(1 \rightarrow 2 \times 2 \rightarrow 3)$, $(2 \rightarrow 3 \times 3 \rightarrow 4)$, $(3 \rightarrow 4 \times 4 \rightarrow 1)$, $(4 \rightarrow 1 \times 1 \rightarrow 2)$. There are 14 possible structures. Rotations of any group to allow the circles to occupy different quadrants of the trapezoid, labeled I–IV, are also considered different 'structures'. There are thus a total of 56 possible 'structures'. In the top right panel the trapezoid has been expanded while maintaining its shape. The increase in empty space allows the circles to move, a little or a lot depending on how much extra space is allowed. The bottom panel shows examples of the circle positions for the 14 allowed structures. (Reprinted with permission from: Richards F. M. and Lim W. A. (1994) An analysis of packing in the protein folding problem. Quart. Rev. Biophys. **26:** 423–498).

Jay Ponder analysed this situation for a real structure in three dimensions with his program PROPAK [8]. The results of applying this procedure to the interior of the small protein crambin are shown in figure 3b. In this case the 'inverse protein folding problem' has been examined. For a particular set of positions in the polypeptide chain, we can ask how many sequences of amino acid residues can satisfy a given packing constraint? Only five residues which are in contact with each other in an interior 'packing unit' were allowed to change in this particular example, the rest of the protein, both main chain and side chains, being fixed. The shape of this curve is very similar to the one shown in figure 3a, and has the same interpretation. The packing is sufficiently good in the native protein that we might conclude that packing is important in 'specifying' the structure. But evolution has not provided the highest possible packing density. The dynamic behaviour of proteins is almost certainly affected by packing density. The adjustment of this property to the optimal level for function may require less than maximally efficient geometrical packing. Nonetheless minimization of empty space within a protein, while avoiding steric overlaps, seems to be a strong guiding principle for stable structures. The observed level of packing can be satisfied with a small number of sequences, not restricted just to one, and, for a given sequence, various arrangements of rotamer shapes can often be accommodated. An appropriate level of underpacking would appear to be important if evolution is to occur through random mutagenesis as is commonly thought.

To what extent does the assembly of the protein chain resemble a three-dimensional jigsaw puzzle, where the size and shape of each individual residue are both of prime importance? To examine this question in some detail there is now a large body of experimental data on the effect of mutations on protein stability. We might expect a variety of effects just from consideration of steric overlaps (fig. 4). Addition of atoms at a given position or a change in shape might easily produce steric problems and require a conformational change of the surrounding atoms to reduce the energetic consequences. Simple deletion of atoms would not cause any steric difficulties, but would lead to cavity formation unless a possible structural shift could arrange to fill the new space. The only mutation that would automatically be expected to produce an increase in stability would be the filling of a pre-existing cavity large enough to accommodate the new atom(s) without any accompanying conformational change. This appears to be an unusual situation. Almost all mutations are accompanied by some conformational change, making prediction of the effects on stability difficult. In most cases mutations lead to lowering of the stability.

The concept of hydrophobic collapse appears prominently in current discussions of the folding problem. Removal of nonpolar material from contact with water is certainly a large driving force, but there is often an implication that the process is controlled solely by the total amount of surface area buried, and that the size and shape of the individual residues are of secondary importance. If this were true, consideration of packing details would not be important in structure prediction.

Some examples of mutations at specific sites

1) The studies by Terwilliger on the gene V protein demonstrate the effects of residue shape [9]. The residues at positions 35 and 47 in the protein interior

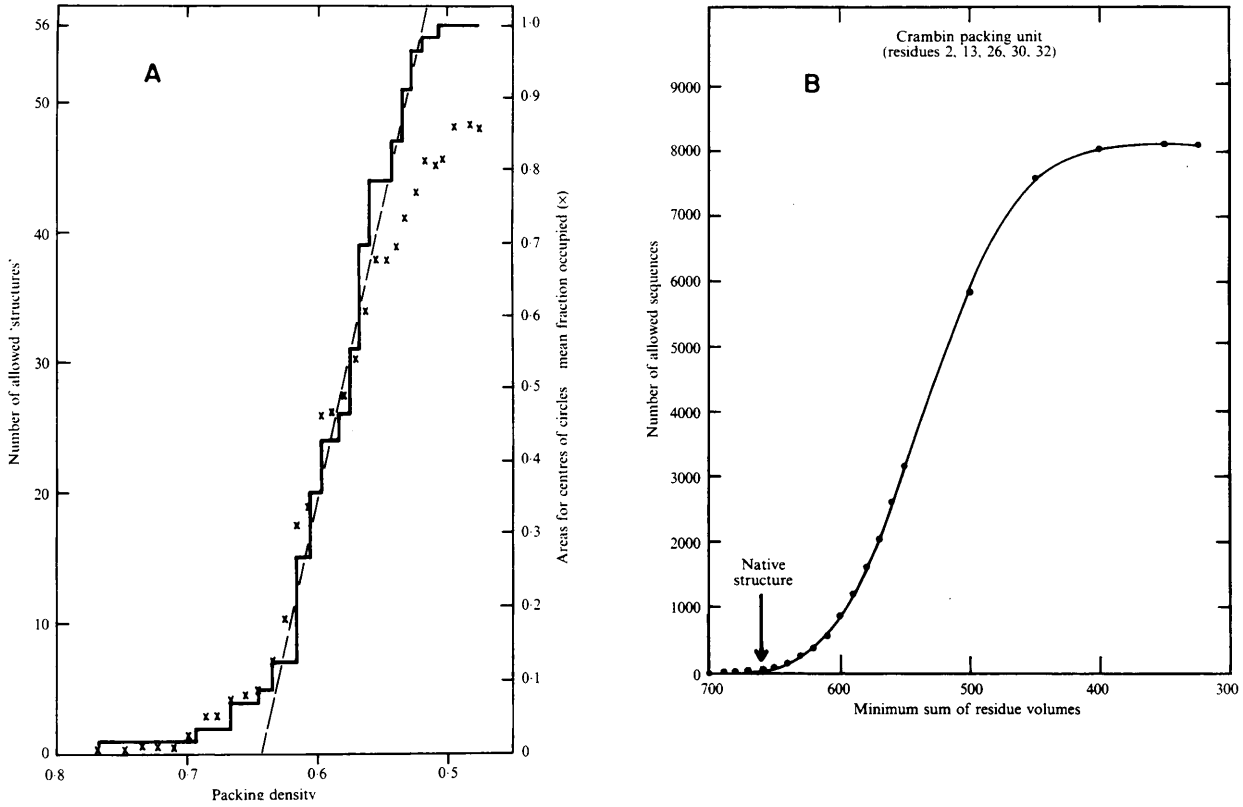

Figure 3. Number of 'structures' as function of packing fraction. (*A*) The number of possible structures, in the 2D model described in figure 2, as a function of the packing fraction, i.e. the fraction of the trapezoidal area occupied by the circles. (*B*) A 3D example from the native conformation of the main chain of the protein crambin. The number of compatible sequences of five interior residues are shown as a function of the specified minimum sum of the standard volumes of the residues in the sequence. This latter number is directly related to the packing fraction. The volume of the cavity containing the five residues is defined by the rest of the protein and remains constant. The arrow shows the value actually found in the native protein. These calculations were made with the program PROPAK of J. Ponder [8]. The similarity in the general shape of the two curves is clear by inspection. (Reprinted with permission from: Richards F. M. and Lim W. A. (1994) An analysis of packing in the protein folding problem. Quart. Rev. Biophys. **26:** 423–498).

are in contact with each other and can be thought of as occupying a single cavity of clearly defined shape. He prepared a number of double mutants with various residues in these two positions. For each pair the complementary mutant was also made where the residue positions were switched (fig. 5). Only nonpolar residues are involved so no change in the electrostatic component of the stability would be expected. The residues at positions 35 and 47 are not part of the active site of the enzyme. The residues in each switched pair were identical except for position. They had the same intrinsic volume and represented an identical loss in surface area upon burial. The effects on the overall stability of the protein were remarkably variable. Thus steric effects which alter the relative positions of other parts of the structure seem to be the only remaining possibility. The general conclusion is that both the size and the shape of the cavity containing residues in the protein interior can be very important in defining the response of the protein to mutation.

2) Brian Matthews and colleagues have made a vast number of mutants of the lysozyme of bacteriophage T4 and determined their detailed structures. In one series of experiments [11, 12], they carried out single site Leu → Ala mutations in various different positions in the peptide chain. The difference in buried nonpolar area is identical for each of these mutations and is thus a constant factor in the energetic effects of all these mutations. However, the structural consequences varied considerably as did the accompanying changes in overall stability (fig. 6, open circles). At one limit the four methylene equivalents removed from the leucine by the mutation simply leave an intact cavity with no other change. This leads to the largest change in stability. At the other limit, the protein structure shifted to fill the cavity almost completely with other atoms, leaving only the constant contribution from the difference in buried area. The linear relation found between the decrease in free energy of stabilization and cavity volume would

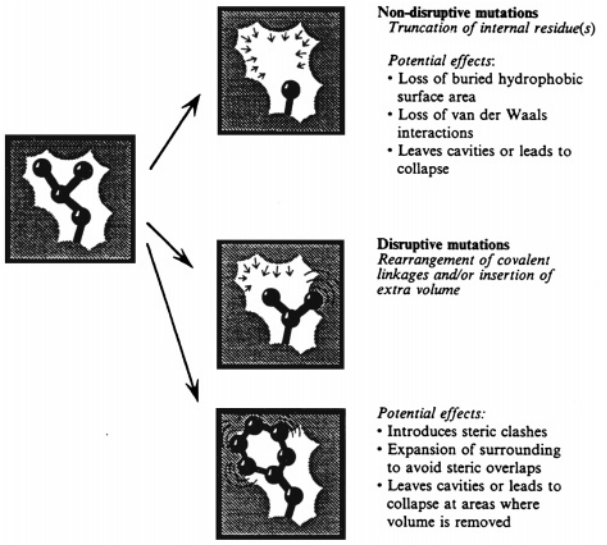

Figure 4. A cartoon illustrating classification of two types of single site mutants, disruptive and nondisruptive, and their potential effects on protein stability. (Reprinted with permission from: Richards F. M. and Lim W. A. (1994) An analysis of packing in the protein folding problem. Quart. Rev. Biophys. **26:** 423–498).

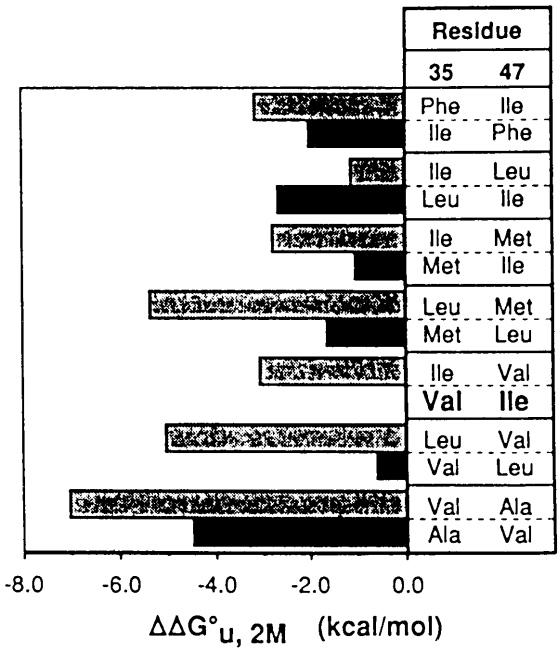

Figure 5. For the gene V protein: comparison of stabilities, expressed as the differences in free energy of unfolding under a standard set of conditions relative to the wild type gene V protein, of pairs of mutant proteins with identical core composition but with residues at positions 35 and 47 reversed. The wildtype protein is indicated by bold face type. (Reprinted with permission from: Terwilliger T. C. (1995) Engineering the stability and function of gene V protein. Adv. Protein Chem. **46:** 177–216, © 1997 Academic Press, Florida).

indicate that the structural shifts responsible for the volume changes contributed minimally to the observed change in stability. Since the mutations are identical and involve only atom deletions, the side chain shape factors do not play any role in this series of experiments. However, various disruptive mutations at two of these same sites show no such relation between residual cavity volume and energy of stabilization as expected (fig. 6, filled circles).

3) *E. coli* thioredoxin is a small, spherical, very stable, multifunctional protein consisting of a central beta sheet surrounded on each side with a pair of alpha helices (fig. 7). As one of its functions it serves as an essential component in the assembly of filamentous phage such as M13. In the protein interior residues L42 and L78 are in contact. Most of the amino acids were substituted in each position, especially all of those capable of providing a formal charge (table 2). With the exception of proline and tyrosine, all of the mutants were biologically active but with varying effects on stability. The melting temperature might drop 10 to 20 °C indicating a loss of thermal stability. However, since T_m for the wild type is about 85 °C, these 'less stable' mutants still appeared to be very stable at room temperature. The CD spectra were all very similar at 25 °C, but the stabilities towards guanidine hydrochloride denaturation were changed, reflecting the T_m shifts. From the known wildtype structure it seems remarkable that all the charged residues could be accommodated with minimal effects. The changes that did occur were not large enough to destroy the biological activity. Unfortunately it has not been possible so far to crystallize these mutants, and no high resolution structural data are yet available except for the recent NMR studies of the Lys 78 mutant reported by DeLorimier et al. [14]. These authors found that chemical shift changes were localized around the site of mutation, and most of the structure was largely unchanged. However, the dynamics of the protein were altered as reflected in both conformational and hydrogen exchange behaviour, and these changes were not restricted to the vicinity of the mutation. In the absence of any structural information at all, the biological activity data might easily be totally misinterpreted in cases such as this.

4) In ribonuclease-S we have a protein peptide complex which can be dissociated and recombined reversibly (fig. 8). Such a bimolecular reaction can be characterized in detail in terms of the thermodynamic parameters. That is, both the enthalpy and free energy can be measured as a function of temperature, yielding the derived values for the entropy and heat capacity changes in this association reaction in addition to this primary data. Through chemical synthesis rather than genetics, a series of variants at position 13 in a truncated version of S-peptide,

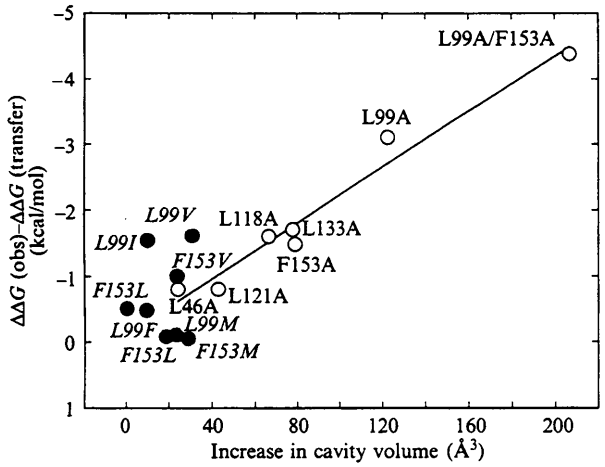

Figure 6. Thermodynamic and structural data from mutants of phage T4 lysozyme. The change in cavity volume derived from the X-ray crystal structures is shown on the abscissa. The observed free energy of stabilization has been corrected for the expected contribution from the change in hydrophobicity attributed to the mutation as estimated from the change in buried nonpolar accessible surface area between the wildtype and the mutant. The residual free energy plotted on the ordinate is the penalty attributed to cavity formation. There are six single and one double mutant representing L→A and F→A nondisruptive mutations, shown as open circles. These fall quite nicely on the straight line shown. The other mutants, defined as disruptive, show no significant correlation. (Reprinted with permission from: Richards F. M. and Lim W. A. (1994) An analysis of packing in the protein folding problem. Quart. Rev. Biophys. **26:** 423–498).

S-15, were prepared. Sg-15 comprises the first 15 residues of S-peptide, and the terminal carboxyl group is amidated. This shortened version has a binding constant equivalent to that of the full length peptide. The thermodynamics of association of the various mutants were established and the crystal structures determined.

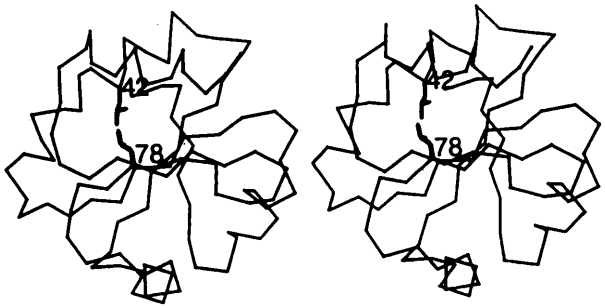

Figure 7. Stereodiagram of the alpha carbon chain of *E. coli* thioredoxin. The side chains shown are of residue 78, coming off the beta sheet, and residue 42, pointing in from an alpha helix packed against the sheet. These two residues, in contact with each other, are part of the internal hydrophobic core of this protein. (Reprinted with permission from: Hellinga H. W., Wynn R. and Richards F. M. (1992) The hydrophobic core of *Escherichia coli* thioredoxin shows a high tolerance to nonconservative single amino acid substitutions. Biochemistry **31:** 11203–11209, © 1997 American Chemical Society).

Table 2. Activity of mutant thioredoxins in phage M13 maturation.

Amino acid	Position 42	Position 78	Amino acid	Position 42	Position 98
Ala	+	+	Leu(wt)	+	+
Arg	+	+	Lys	+	+
Asn			Met		+
Asp	+	+	Phe	+	+
Cys		+	Pro	−	−
Gln	+	+	Ser	+	+
Glu	+	+	Thr	+	+
Gly		+	Tyr		−
His			Val	+	+
Ile	+				

Blank = mutant was not isolated; + = supported phage assembly; − = did not support phage growth. Note that Trp was not present in the mutagenic oligonucleotide pools. Residues 42 and 78 are interior and fully buried in the structure of wildtype thioredoxin. (Adapted with permission from: Hellinga H. W., Wynn R. and Richards F. M. (1992) The hydrophobic core of *Escherichia coli* thioredoxin shows a high tolerance to nonconservative single amino acid substitutions. Biochemistry **31:** 11203–11209, © 1997 American Chemical Society).

The wild type methionine was changed to other nonpolar residues involving a change in both size and shape (table 3). Very little is left of any simple relationship between the thermodynamic parameters and the structures. Some pictures of the structures of the extremes in the series, Gly and Phe, are shown in figure 9. The protein attempts unsuccessfuly to fill the cavity left by the M13G mutation. L51 changes its rotamer conformation to cover the cavity and partial filling is obtained with one water molecule which manages to find enough main chain H-bonding partners to allow it to occupy space in this otherwise hostile nonpolar environment. At the other end of the size range the Phe replacement is too large to fit in the original binding pocket and slight main chain movements allow steric access, but the new fit is not good and extra cavity space appears which simply cannot be filled by any compensating movement within the protein. Both the underpacked M13G and the overpacked M13F are bound more weakly than the wildtype peptide (table 3).

This type of behaviour is not uncommon. In single site mutants, the structural changes are generally greatest near the site of mutation, and moving away, decrease radially in all directions. Even the small changes are so complex that the linkage relations do not allow assignments of the energetic changes to unique parts of the altered residue and its immediate contacts. Such problems have been analysed in detail in the work of DiCera on thrombin [18]. The RNase-S example also provides another fairly common observation, a specific structural change at some distance from the site of mutation. The 1.5 Å movement of the 66-69 loop which is 20 Å from residue 13, is the largest motion seen in the comparison of the X-ray structures. There is no convincing explana-

Figure 8. A ribbon diagram of ribonuclease-S. The helix of the dissociable S-peptide (residues 1–20) is shown in blue in the centre background with the side chain of methionine coming off the alpha carbon atom of residue 13 in red. S-protein (residues 21–124) and S-peptide can be separated and recombined reversibly with loss and regain of activity.

tion yet of how the changes in binding can produce a major movement over such a distance.

5) Staphylococcal nuclease has been studied extensively as a model system for protein folding [19]. The ease of crystallization increases its attraction. R. Wynn has carried out a set of studies on residue 23 which is located inside the protein and is not part of the groove providing the active site (fig. 10). Modifications at this position can have only a remote site effect on biological activity. Conversion of the wildtype valine to a cysteine residue at position 23 provides a fully active variant which serves as the reference molecule for further modifications produced chemically on the free SH group. A series of n-alkane thiols have been attached through a disulfide bond to C23 in the unfolded protein to produce a series of unnnatural side chains which differ consecutively by one CH_2 group in length. The protein is then put under refolding conditions. Surprisingly, all of the variants produced so far have not only refolded but crystallized quite well. The CD spectra are virtually identical, and nucleolytic activity is retained (fig. 11).

Wynn et al. have said: 'The strictly aliphatic side chains continuously increase in volume, hydrophobicity and flexibility. Volume and hydrophobicity would be expected to increase approximately linearly with side chain length, but the number of possible conformations taken up by a side chain should increase geometrically with the number of rotatable bonds. The methyl cysteine disulfide is almost isosteric with methionine, the most flexible nonpolar natural amino acid with 3 rotatable bonds or 27 rotamers. The pentyl derivative would have well over 1000 potential rotamers. This would be expected to have a dramatic effect on the side-chain conformational entropy....' [20]. The butyl-containing residue already has a volume slightly larger than tryptophan. The heptyl form would be 3 methylene equivalents bigger still, an additional 78 $Å^3$, with more than 10 000 rotamers. The cyclopentyl and cyclohexyl forms differ from their straight chain analogues only by the loss of two hydrogen atoms and of some additional volume due to ring closure. The main change is in the marked reduction in the flexibility of the chain. The crystal structures of all of these modified forms have been established at resolutions of 2.0 to 2.3 A with R factors in the range of 16% to 18% [21] (fig. 12).

The overall structures of these variants are very similar to the starting V23C reference structure. The six N terminal and 8 C terminal residues are not seen in the electron density maps as in the original wildtype structure. Apart from the side chain at position 23 the structure of the protein is remarkably constant. All the

Table 3. Difference thermodynamic parameters relative to S-15, the truncated S-peptide, for peptide binding to S-protein[a,b]

Peptide	$\Delta\Delta H$ (kcal mol^{-1})	$\Delta\Delta C_p$ (kcal mol^{-1} K^{-1})	$\Delta\Delta G°$ (kcal mol^{-1})	$T\Delta\Delta S°$ (kcal mol^{-1})
M13V	3.5	0.17	−0.1	3.6
M13I	5.4	0.21	0.1	5.3
M13L	5.7	0.10	0.3	5.2
M13NLe	7.9	0.19	0.8	7.1
M13ANB	9.2	0.17	1.5	7.7
M13F	4.7	0.03	2.6	2.1
M13A	2.8	−0.20	4.3	−1.5
M13G	(−1.3)[c]		(5.0)[c]	(−6.3)

[a]The reference temperature is 25 °C, the solvent aqueous acetate buffer pH 6.0, total ionic strength 0.15 M. Under these conditions the free energy, enthalpy and entropy ($T\Delta S$) of binding of the reference peptide, S-15, are: −9.5, −39.3, −29.8 kcal mol^{-1} respectively. [b]The data are taken from [16] except for M13G, see [17]. (Reprinted with permission from: Thomson J., Ratnaparkhi G. S., Varadarajan R., Sturtevant J. M. and Richards F. M. (1994) Thermodynamic and structural consequences of changing a sulfur atom to a methylene group in the M13N1e mutation in ribonuclease-S. Biochemistry **33:** 8587–8593, © 1997 American Chemical Society). [c]The binding is sufficiently weak that these numbers are only approximate.

side chains, including those providing the cavity for residue 23, are clearly seen and occur in rotameric forms, as in the native enzyme.

There are significant cavities adjacent to residue 23 in the V23C reference stucture. These are unusual in size, but comparable cavities are known in other native proteins. The methyl, ethyl and propyl homologues proceed to fill these cavities and the movement of any other atoms in the protein, as analysed by difference distance matrices, is less than 0.5 Å. There are no detectable cavities in the propyl structure. The larger side chains cannot be accommodated without motion in the rest of the protein. The major response is the rigid body movement of helix 1 away from the mouth of the beta barrel. Cavity space again appears as the fit of the units is not perfect. This cavity space grows and decreases cyclically as the length of the chain at 23 increases. The complete filling of space reflects the subtle interplay between side chain volume, shape and flexibility. Any simple correlation between cavity volume and thermodynamic parameters thus disappears.

One of the most interesting aspects of this study is an examination of what is not seen. In the homologous series the omit maps show the side chain electron density unambiguously for the methyl and ethyl groups, but with the propyl and larger groups many of the outer atoms are not seen well or at all (figs 11 and 12). For the butyl chain reduction of the contour level the map fails to show any outer atoms in the chain above the noise. This is reminiscent of the behaviour of the fatty acid chains in a lipid bilayer. The longer chains either occupy a very large number of alternative sets of rotamers, or rapidly interconvert between different rotamer forms. X-ray analysis will not distinguish between the time-dependent and fixed but randomly distributed structures. This should be an interesting case to examine by NMR.

Normal behaviour of interior residues is regained when the C5 and C6 chains are converted to their cyclic forms (fig. 12). The structures show well-developed density encompassing all the atoms of the rings. The cyclohexyl form refines to the chair conformation, the cyclopentyl ring to a conformation favoured for five-membered saturated rings. The entropic penalty appears to be lowered by the ring formation to the point where the single rotamer conformations are energetically acceptable as is usually the case for internal residues.

The linear side chains are very flexible and appear to be liquid-like in the interior of a structure which is otherwise very clearly a unique solid, with all interior atoms in defined locations with comparable B factors. If all the residues were nonpolar chains of varying length the hydrophobic driving force would be there and the collapse would occur on folding, but the micellar structure formed would not have a unique time-independent conformation, and the specific ligand binding activity would be much harder, if not impossible, to attain.

These disulphide-containing derivatives may also be used to provide a thermodynamic cycle where the equilibrium constants for all four sides can be individually measured. The unfolding of the SH and SS forms in urea can be studied individually, for example while the energetics of the chemical mutations can be obtained through the redox potentials of the folded and unfolded forms of the protein. Such experimental measurements are not possible with genetic mutations although they can be simulated in theoretical calculations. Thermodynamic evidence for the existence of structure in the 'unfolded' state can thus be obtained. This part of the story is not considered further in this paper. (See Wynn et al. [22, 23].)

Conclusions

The chemical structure of the nucleic acid bases seems eminently suitable for their function (i.e. the encoding of genetic information, replication, and transcription with the dominant interaction of base pairing and the one to one coding). The translation apparatus is complex but will produce a peptide chain of α amino acids

Figure 9. A thick section through the middle of a stick model of RNase-S centred on residue 13. (*A*) Stick model with the side chains of selected residues identified. Some of the following panels have dot surfaces at the van der Waals radius surrounding selected atoms or side chains. (*B*) Variant M13F. Residue 13 is shown with a high dot density surface; the surrounding residues are shown at low density. (*C*) Variant M13F (blue) and wildtype (red) overlaid, showing approximate atomic shifts which correspond to a small increase in cavity volume to provide room for Phe 13. Residues not forming part of the cavity wall are shown in white. (*D*) same as (*C*) with dot surfaces shown for both structures. The increase in cavity volume due to mismatch in the side chain shapes can be seen. (*E*) Hypothetical Gly mutant. The wildtype structure is shown with the side chain of Met 13 removed back to the α-carbon atom to show the appearance of the cavity if there were no change in structure on mutation. (*F*) Actual M13G structure shown with dot surfaces. The change in the L51 position and the water molecule at the + position are the partially successful attempts to fill the cavity left by the mutation. (*G*) Overlay of M13G (yellow) and wildtype (red). The red structure is identical to that shown in panel (*C*). The shifts are small, but generally in the opposite direction to M13F and tend to collapse the cavity. The rotamer change in L51 is clearly visible. (*H*) Same as (*G*) with dot surfaces shown. M13 can be seen under L51 and shows some of the cavity not filled by either L51 or the water molecule. (Reprinted with permission from: Varadarajan R. and Richards F. M. (1992) Crystallographic structures of ribonuclease-S variants with non polar substitution of position 13: packing and cavities. Biochemistry **31:** 12315–12327, © 1997 American Chemical Society).

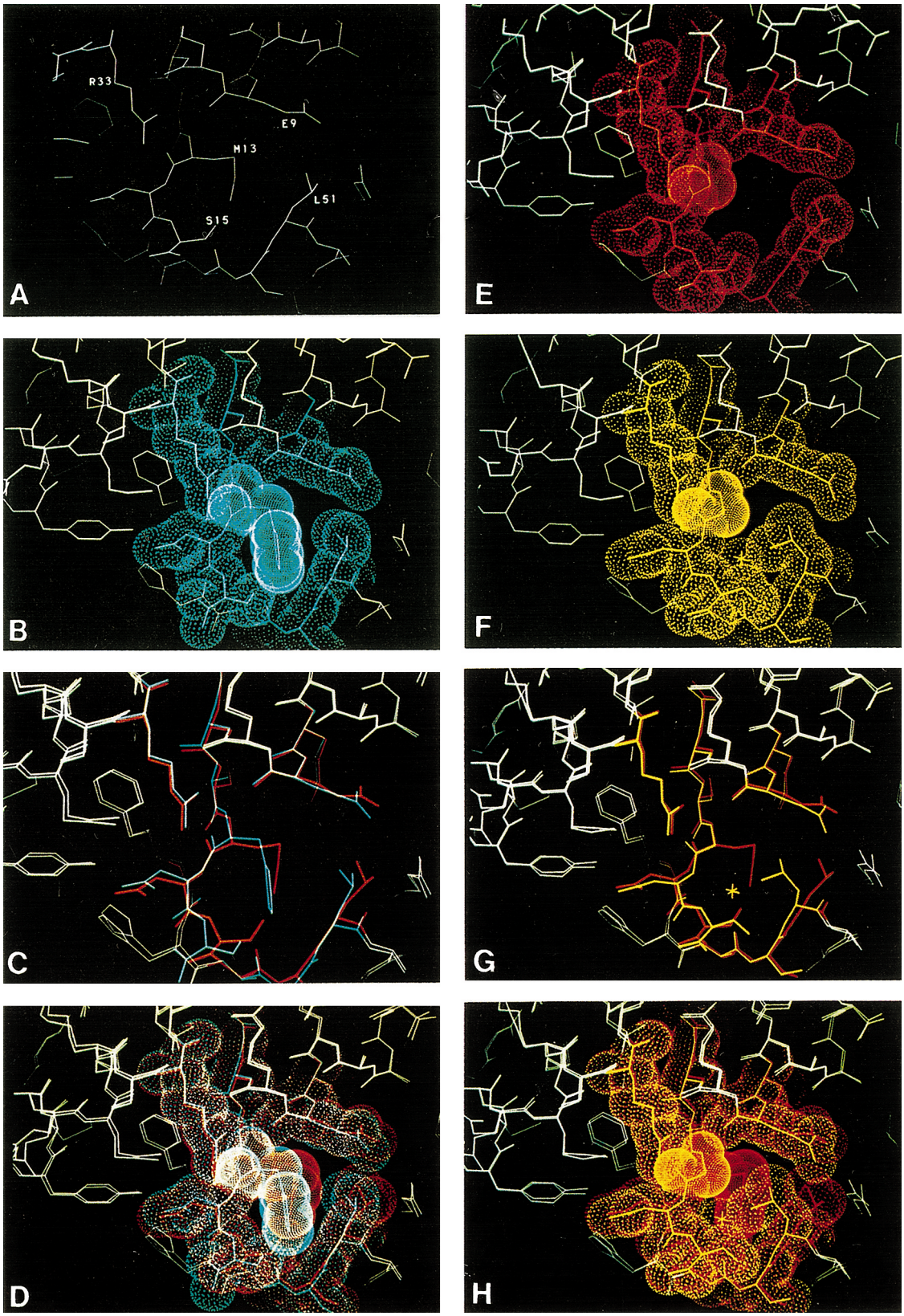

Figure 9.

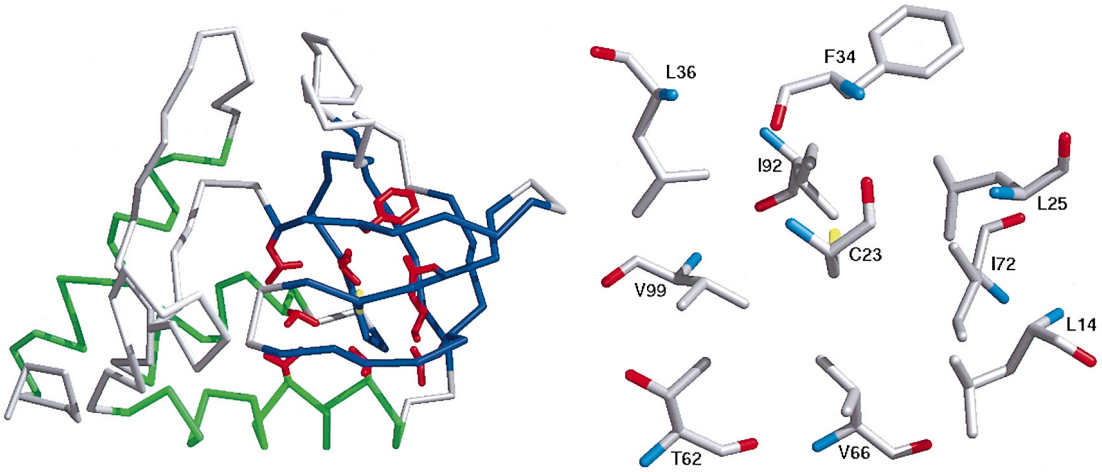

Figure 10. Staphylococcal nuclease. Alpha carbon chain of the mutant V23C, with residue 23 and the residues providing the wall around 23 are shown in red. The beta strands are blue and the alpha helical residues green. An expanded view of residue 23 with its surroundings identified are shown in the right panel. (Reprinted with permission from: Wynn R., Harkins P., Richards F. M. and Fox R. O. (1996) Mobile unnatural amino acid sidechains in the core of staphylococcal nuclease. Protein Sci. **5:** 1026–1031, © 1997 Cambridge University Press, New York).

of any type that happen to be attached to the tRNAs complementary to the message. The three-letter genetic code restricts the maximum number of protein amino acids, but it is not obvious why that number should be 20. Stipulating this number, any set of 20 could be chosen as far as the coding is concerned. The loading of the tRNAs depends on the specificity of the synthetases. These enzymes, in turn, could be designed to recognize any set of amino acids. Each amino acid already has its own biosynthetic route. With this level of complexity,

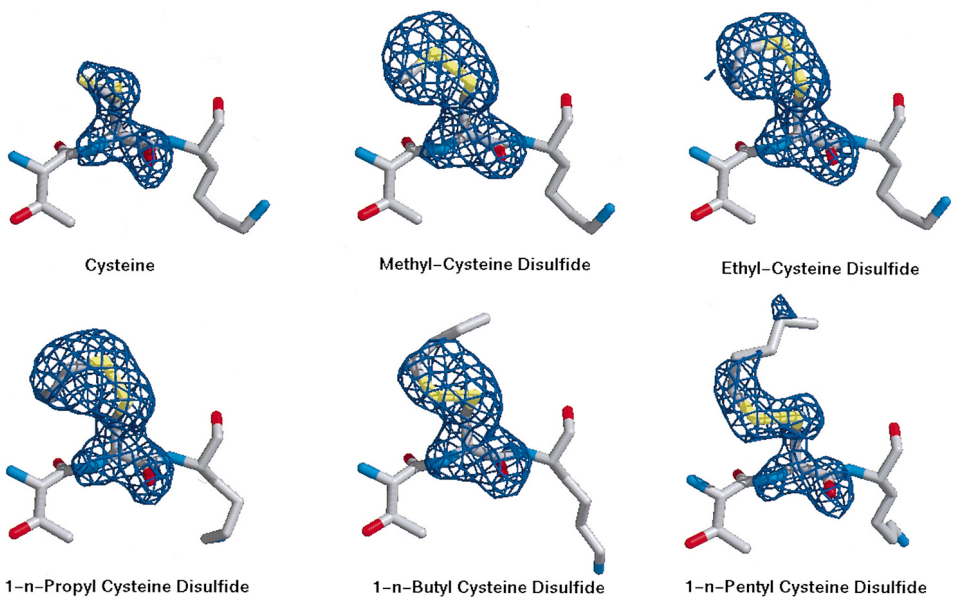

Figure 11. Omit maps showing the electron density in birdcage form for the disulfide chemical modifications at position 23. These mutants were prepared with the n-alkyl methyl through pentyl mercaptans. (Reprinted with permission from: Wynn R., Harkins P., Richards F. M. and Fox R. O. (1996) Mobile unnatural amino acid side chains in the core of staphylococcal nuclease. Protein Sci. **5:** 1026–1031, © 1997 Cambridge University Press, New York).

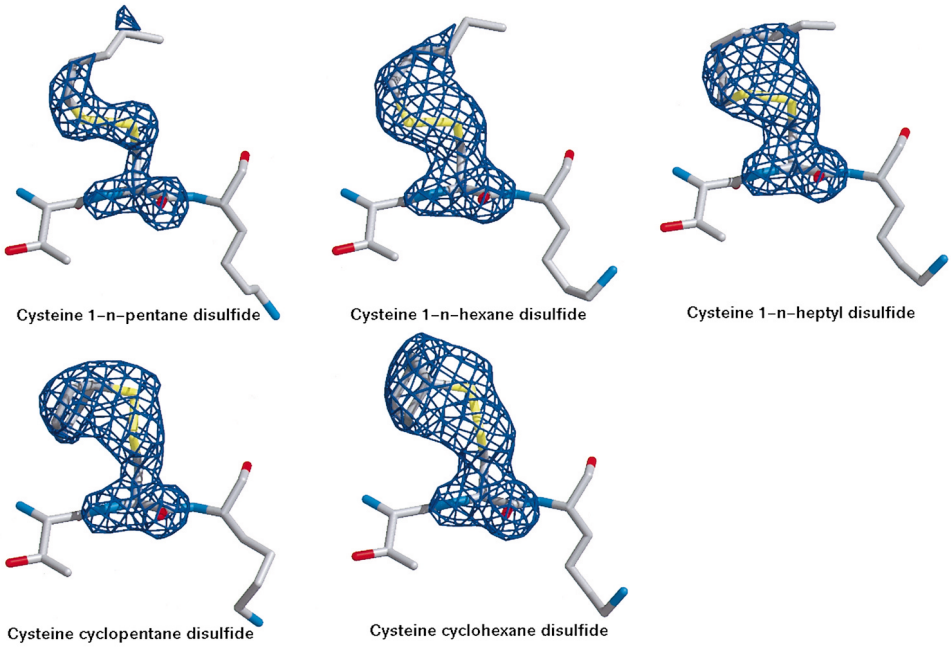

Figure 12. Same as figure 11 but showing the C5, C6 and C7 n-alkyl derivatives and the cyclopentyl and cyclohexyl forms. (Reprinted with permission from: Wynn R., Harkins P., Richards F. M. and Fox R. O. (1997) Comparison of straight chain and cyclic unnatural amino acids embedded in the core of staphylococcal nuclease. Protein Sci. **6:** 1621–1626, © 1997 Cambridge University Press, New York).

we might imagine that a set of pathways which would provide any specified set of amino acid side chains could readily be laid out. Why were the current 20 selected?

I am not aware of any detailed answer to the above question, but there are some limits on the possible candidates. The packing in native proteins must be good enough to exclude essentially all solvent molecules, but the fit must not be perfect as in an ordinary jigsaw puzzle. If it were, interior single site mutations would always be disruptive, either through under- or overpacking. At a minimum, two simultaneous mutations would be required to maintain the native structure and its biological function. Evolution as we understand it today could not have occurred under such stringent requirements.

For physiologically relevant conditions it appears that the folded structure usually has to be unique to arrive at the necessary specificity in the interactions of the protein with other molecules, large or small. If a homologous series of straight chain aliphatic amino acids had been selected to represent the nonpolar amino acids, similar to those reported in this study, a unique structure would never be formed. The protein interior would be truly a micelle, stable in the sense of a compact molecule because of the strong hydrophobic component, but not a unique structure because of the enormous entropic penalty of fixing these flexible chains and the many ways in which the internal volume of varying shape could be filled. Fifteen of the 20 current amino acids have only two or fewer rotatable side chain bonds. Three (Glu, Lys, Arg) of the other five have three or four rotatable bonds, are almost invariably on the protein surface, and use their flexibility to help ensure that their charged groups can indeed reach the solvent. Met and Gln, each with three bonds, are the most flexible of the nonpolar or mildly polar residues. The chunky aromatic rings, attached through only two bonds to the main chain, severely limit the tertiary structure options for filling space efficiently, thus leading towards a unique structure for any particular sequence. This could easily be overdone. Suppose that the heme group occurred as a regular side chain. The shape and special requirements of this large flat plate would severly limit the number of basic scaffolds that might be constructed.

The general rules appear to be:

1) Have a range of side chain sizes and polarities, but do not have very large axial ratios. (For example, we can densely pack a set of long rigid rods, simply by bundling them with their long axes strictly parallel to each other);
2) have a few awkward shapes, but not too many (they are useful in leading to unique structures);
3) set the flexibility in a rather narrow range;
4) have a few small ones to fill in empty spaces as required.

These rules do not restrict our options to the standard set of 20, but they make it a reasonable choice and

suggest that, given the range of activities required from proteins, there may not be as many other options as one might have thought. Biologically there may be other reasons for the original choice. Either directly or as precursors, the amino acids serve many other functions in an organism in addition to their role as subunits in proteins. A set strictly reserved for making proteins might be unduly 'expensive'.

The rules must be very carefully applied. The marginal stability of almost all proteins seems to be a biological necessity. This is not usually associated with the specific biological function, but perhaps with the necessity to maintain the option of rapid disposal when required. As the structures of the proteins from thermophilic bacteria become available, it is clear that proteins that are stable at over 100 °C are common in certain environments. The mystery is what is the origin of this thermostability. In the structural sense it is clearly very subtle. The maximum energy of stabilization may be roughly constant with the midpoints for the heat and cold denaturation simply proceeding up and down the temperature scale in tandem. Slightly denser packing may be observed in the thermophilic structures, but it is not yet clear how general this is [24]. It is possible that the peak of high temperature enzyme stability for industrial processes may be tempered by instability caused by cold denaturation during shelf storage at room temperature. On the other hand the dramatic increase in stability that can be produced by chemical crosslinking could overcome this problem in practical applications.

A general solution to the protein stability problem, good enough for even moderately accurate predictions, still eludes us, but progress is remarkably rapid. The current level of excitement seems certain to continue for a least a few decades. Linderstrom-Lang would be particularly tickled by what is happening if only he were here to see it.

1 Fruton J. S. (1972) Molecules and life, Wiley-Interscience, New York
2 Bullough P. A., Hughson F. M., Skehel J. J. and Wiley D. C. (1994) Structure of influenza hemagglutinin at the pH of membrane fusion. Nature **371:** 37–43
3 Matthew J. B. and Richards F. M. (1982) Anion binding and pH dependent electrostatic effects in ribonuclease. Biochemistry **21:** 4989–4999
4 Hermans J. and Scheraga H. A. (1961) Structural studies of ribonuclease. V. Reversible change of configuration. J. Am. Chem. Soc. **83:** 3283–3292
5 Kitaigordosky A. I. (1961) Organic Chemical Crystallography. Consultants Bureau, New York (authorized English translation 1961, original Russian text published 1955)
6 Kitaigordosky A. I. (1973) Molecular crystals and molecules. Academic Press, New York
7 Richards F. M. and Lim W. A. (1994) An analysis of packing in the protein folding problem. Quart. Rev. Biophys. **26:** 423–498
8 Ponder J. W. and Richards F. M. (1987) Tertiary template for proteins – use of packing criteria in the enumeration of allowed sequences for different structural classes J. Molec. Biol. **193:** 775–791
9 Terwilliger T. C. (1995) Engineering the stability and function of gene V protein. Adv. Protein Chem. **46:** 177–216
10 Sandberg N. S. and Terwilliger T. C. (1991) Energetics of repacking a protein. Proc. Natl Acad. Sci. USA **88:** 1706–1710
11 Eriksson A. E., Baase W. A., Zhang X.-J., Heinz D. W., Blaber M., Baldwin E. P. et al. (1992) Response of a protein structure to cavity-creating mutations and its relationship to the hydrophobic effect. Science **255:** 178–183
12 Eriksson A. E., Baase W. A. and Matthews B. W. (1993) Similar hydrophobic replacements of Leu 99 and Phe 153 within the core of T_4 lysozyme have different structural and thermodynamic consequences. J. Molec. Biol. **229:** 747–769
13 Hellinga H. W., Wynn R. and Richards F. M. (1992) The hydrophobic core of *Escherichia coli* thioredoxin shows a high tolerance to nonconservative single amino acid substitutions. Biochemistry **31:** 11203–11209
14 DeLorimier R., Hellinga H. W. and Spicer L. D. (1996) NMR studies of structure, hydrogen exchange, and main-chain dynamics in a disrupted-core mutant of thioredoxin. Protein Sci. **5:** 2552–2565
15 Varadarajan R. and Richards F. M. (1992) Crystallographic structures of ribonuclease-S variants with non polar substitution of position 13: packing and cavities. Biochemistry **31:** 12315–12327
16 Thomson J., Ratnaparkhi G. S., Varadarajan R., Sturtevant J. M. and Richards F. M. (1994) Thermodynamic and structural consequences of changing a sulfur atom to a methylene group in the M13Nle mutation in ribonuclease-S. Biochemistry **33:** 8587–8593
17 Connelly P. R., Varadarajan R., Sturtevant J. M. and Richards F. M. (1990) Thermodynamics of protein-peptide interactions in their ribonuclease-S system studies by titration calorimetry. Biochemistry **29:** 6108–6114
18 DiCera E. (1997) Site-specific linkage thermodynamics. Adv. Protein Chem. (in press)
19 Shortle D., Wang Y., Gillespie J. R. and Wrabl J. O. (1996) Protein folding for realists: a timeless phenomenon. Protein Sci. **5:** 991–1000
20 Wynn R., Harkins P. C., Richards F. M. and Fox R. O. (1996) Mobile unnatural amino acid side chains in the core of staphylococcal nuclease. Protein Sci. **5:** 1026–1031
21 Wynn R., Harkins P. C., Richards F. M. and Fox R. O. (1997) Comparison of straight chain and cyclic unnatural amino acids embedded in the core of staphylococcal nuclease. Protein Sci. **6:** 1621–1626
22 Wynn R. and Richards F. M. (1993) Partitioning the effects of changes in a protein to the folded or unfolded forms by using a thermodynamic cycle: a change in *E. coli* thioredoxin does not affect the unfolded state. Biochemistry **32:** 12922–12927
23 Wynn R., Anderson C. L., Richards F. M. and Fox R. O. (1995) Interactions in nonnative and truncated forms of staphylococcal nuclease as indicated by mutational free energy changes. Protein Sci. **4:** 1815–1823
24 DeDecker B. S., O'Brien R., Fleming P. J., Geiger J. H., Jackson S. P. and Sigler P. B. (1996) The crystal structure of a hyperthermophilic archeal TATA-box binding protein. J. Molec. Biol. **264:** 1072–1084
25 St. Clair N. L. and Navia M. A. (1992) Cross-linked enzyme crystals as robust catalysts. J. Am. Chem. Soc. **114:** 7314–7316

.