Looking at proteins: representations, folding, packing, and design

Biophysical Society National Lecture, 1992

Jane S. Richardson,* David C. Richardson,* Neil B. Tweedy,* Kimberly M. Gernert,* Thomas P. Quinn, Michael H. Hecht,‡ Bruce W. Erickson,§ Yibing Yan,§ Robert D. McClain, Mary E. Donlan, and Mark C. Surles

*Department of Biochemistry, Duke University, Durham, North Carolina 27710; ‡Departments of Chemistry and Molecular Biology, Princeton University, Princeton, New Jersey 08544-1009; and §Department of Chemistry, University of North Carolina, Chapel Hill, North Carolina 27599-3290 USA

ABSTRACT Looking at proteins is an active process of interpretation and selection, emphasizing some features and deleting others. Multiple representations are needed, for such purposes as showing motions or conveying both the chain connectivity and the three-dimensional shape simultaneously. In studying and comparing protein structures, ideas are suggested about the determinants of tertiary structure and of folding (e.g., that Greek key β barrels may fold up two strands at a time). The design and synthesis of new proteins "from scratch" provides a route toward the experimental testing of such ideas. It has also been a fruitful new perspective from which to look at structures, requiring such things as statistics on very narrowly defined structural categories and explicit attention to "negative design" criteria that actively block unwanted alternatives (e.g., reverse topology of a helix bundle, or edge-to-edge aggregation of β sheets).

Recently, the field of protein design has produced a rather unexpected general result: apparently we do indeed know enough to successfully design proteins that fold into approximately correct structures, but not enough to design unique, native-like structures. The degree of order varies considerably, but even the best designed material shows multiple conformations by NMR, more similar to a "molten globule" folding intermediate than to a well ordered native tertiary structure. In response to this conclusion, we are now working on systems that test useful questions with approximate structures (such as determining which factors most influence the choice of helix-bundle topology) and also analyzing how natural proteins achieve unique core conformations (e.g., for side chains on the interior side of a β sheet, illustrated in the kinemages).

INTRODUCTION

Protein crystallography is rather like a dog chasing cars. Both are probably foolish and certainly heroic endeavors. But in both cases, the real problem is: suppose you actually manage to catch one, what in the world do you do with it (Victor Bloomfield, 1979, personal communication).

Our answer to that, other than chewing on the tires, has always been that you look at them. Looking, and especially seeing, is an active process of interpreting, of finding and emphasizing patterns, of adding things and deleting things. This is particularly clear in the case of proteins, for two reasons:

First, proteins are so complex that showing every atom is almost immediately rejected as hopelessly confusing. Color plate 1 *a* shows an all-atom brass model of a small protein, which is only a little more illuminating than a list of all the x, y, z's. But even in the days before computer graphics, we found ways to show meaningful subsets of the information, as in plates *1, b* and *c,* where the backbone fold is emphasized by adding UV-fluorescent dye in tygon tubing and then the atoms are deleted by turning off the room lights. Kinemage 1 shows a similar progression from confusion to simplification, for a computer graphics example.

Second, for something smaller than the wavelength of visible light, there is no such thing as showing how it really looks on the molecular level. This is healthy because it encourages experimentation with multiple representations. Each representation emphasizes different features of the structure and can help you see those features in the first place (while helping you ignore other things, for better and for worse).

Representations of proteins in one, three, and four dimensions

The two most traditional representations for molecules are either stick figures for the bonds or spheres for the atoms, paralleling the two major types of physical models: brass-rod "Kendrew" models (e.g., plate 1 *a*) and plastic-ball "CPK" models (e.g., plate 2). CPK models are very informative during the process of putting them together, but the completed models all look alike. Computer versions of CPK models have successfully imitated their appearance and most of their disadvantages (the fact that the inside is completely hidden, and the difficulty of identifying an atom or group), without, so far, imitating the real virtue of CPKs, which is the physical "feel" for the bumps, constraints, and degrees of freedom one obtains by manipulating them.

T. P. Quinn's present address is Department of Biochemistry, University of Missouri, Columbia, MO 65211.

R. D. McClain's present address is Nalco Chemical Co., Sugar Land, TX 77487.

M. E. Donlan's present address is Molecular Simulations Inc., Waltham, MA 02154.

M. C. Surles' present address is San Diego Supercomputer Center, San Diego, CA 92186.

Address correspondence to Jane Richardson.

Abbreviations used: PDB, Brookhaven Protein Data Bank (Bernstein et al., 1977).

0006-3495/92/11/1186/24 $2.00

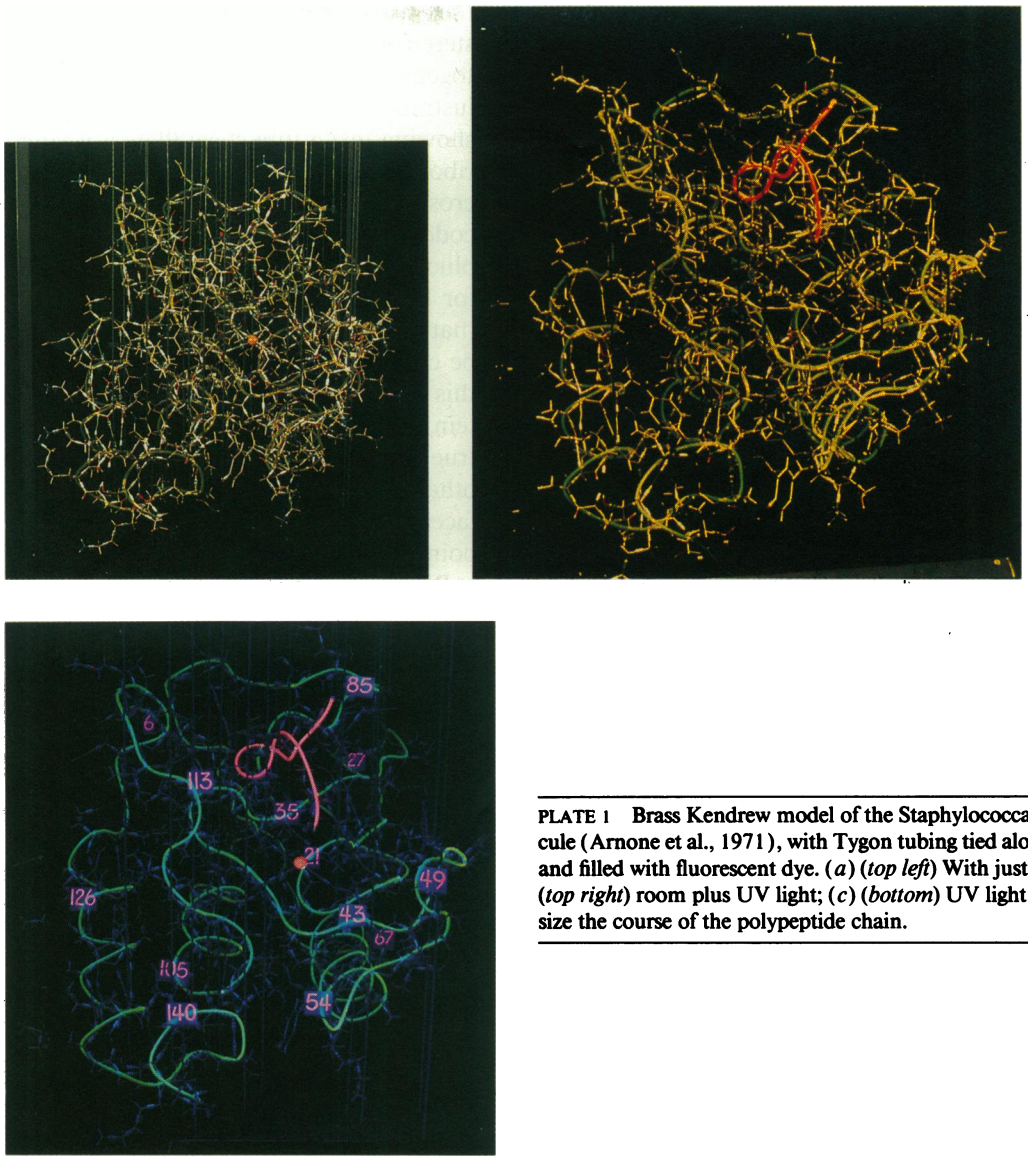


PLATE 1 Brass Kendrew model of the Staphylococcal nuclease molecule (Arnone et al., 1971), with Tygon tubing tied along the backbone and filled with fluorescent dye. (*a*) (*top left*) With just room lights; (*b*) (*top right*) room plus UV light; (*c*) (*bottom*) UV light only, to emphasize the course of the polypeptide chain.

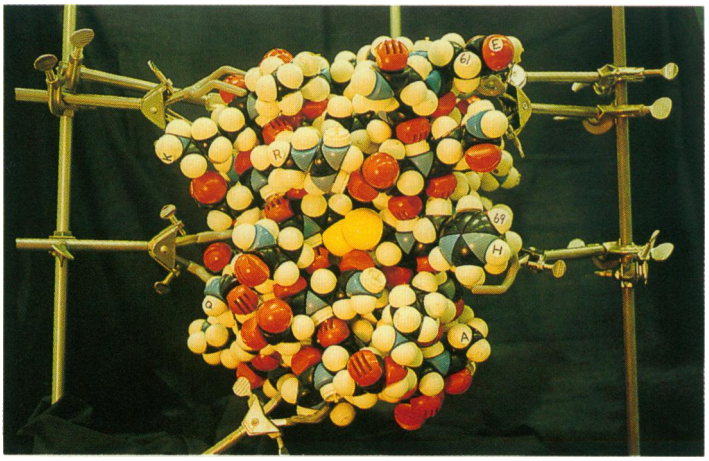


PLATE 2 Plastic CPK space-filling model of the Felix molecule (Hecht et al., 1990), with sulfurs of the SS bridge showing in yellow.

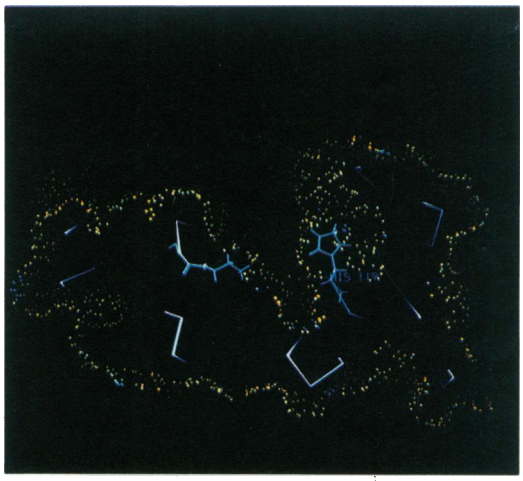


PLATE 3 A thin slice through ribonuclease A (PDB file 5RSA; Wlodawer and Sjolin, 1983), with vectors for the Cα's (*white*) and active site His (*blue*). Dots outline the exposed surface (Connolly, 1983) color-coded with green (C), red (O), blue (N), and yellow (H). Most dots are yellow.

Brass models or stick figures concentrate on the bond connectivity of the molecule, while CPK models or atom-sphere figures concentrate on the outside shape in three dimensions. In fact, one of the most basic and fascinating things about protein structures is that they are inherently both one- and three-dimensional at the same time. The interplay between those two aspects is crucial to protein folding, evolution, mobility, and much of function; therefore, it is desirable to show both together, and several types of representation at least attempt to do so.

Dot surfaces (Connolly, 1983) are primarily a way of showing outside surface, but they have the advantage that you can see through them, so that a stick model can be included inside. When working interactively, one can in fact be aware of both the identity and connectivity of the amino acids and of the three-dimensional shapes they are creating. As for all protein representations, stereo helps, color helps, real-time rotation helps, and all together are even better. In static two-dimensional illustrations, comprehensibility unfortunately requires showing just a thin slice. Plate 3 is such a slice through ribonuclease, outlining with dots the surface shape of a cross-section through the active site. The dots are color-coded by atom type; there are a few patches of red or blue, but primarily the dots are yellow, which is the color for hydrogens. This slide was made before we learned that to make sense of the dot surfaces, hydrogens should be colored to match the atom they bond to. However, this version makes the very important point that the protein surface consists almost entirely of hydrogens. That is true both for the outside, which contacts solvent and other molecules, and also for the internal contact surfaces between parts of the protein. We will return to this point later.

Ribbon schematics, either hand-drawn (plate 4; Richardson, 1985) or computer-generated (Fig. 1; Carson and Bugg, 1986), also attempt to combine one and three dimensions. Obviously, they show the path of the polypeptide chain, with cues that try to make the connectivity and relative positions unambiguous. However, they also reflect important parts of the interactions in three dimensions. For instance, the plane of the β-strand arrows or of the helical ribbons follows the hydrogen-bonding direction, so that when the planes of neighboring arrows turn in synchrony you perceive them as belonging together in a unified sheet.

Plate 5 shows another such attempt, for pancreatic trypsin inhibitor, made from wooden closet pole with a mitre/bevel saw. It uses distance, angle, and dihedral values for successive α-carbon positions, in the same way as a wire backbone model (Rubin and Richardson, 1972), but the wood is thick enough so that the parts of it almost touch at hydrogen-bonding distance. The connectivity is clear, but there is also the strong feeling of a compact but nonintersecting chain in three dimensions.

It is also possible to show explicitly the interactions

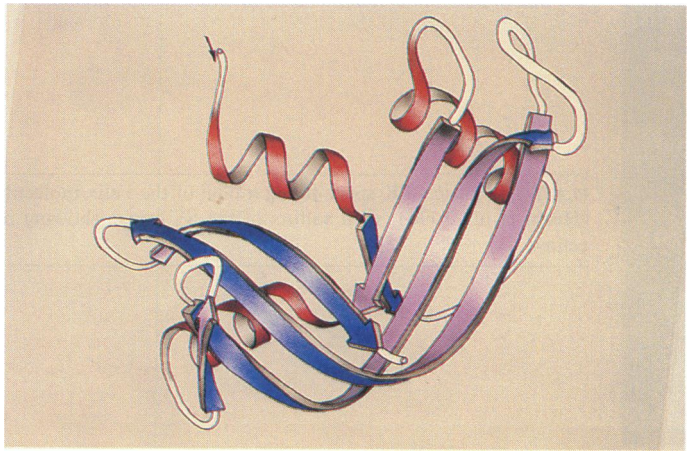


PLATE 4 Ribbon schematic drawing of the ribonuclease A backbone. Helices are shown as red spirals and β strands as arrows, with the outside of the sheet blue and the inside lilac.

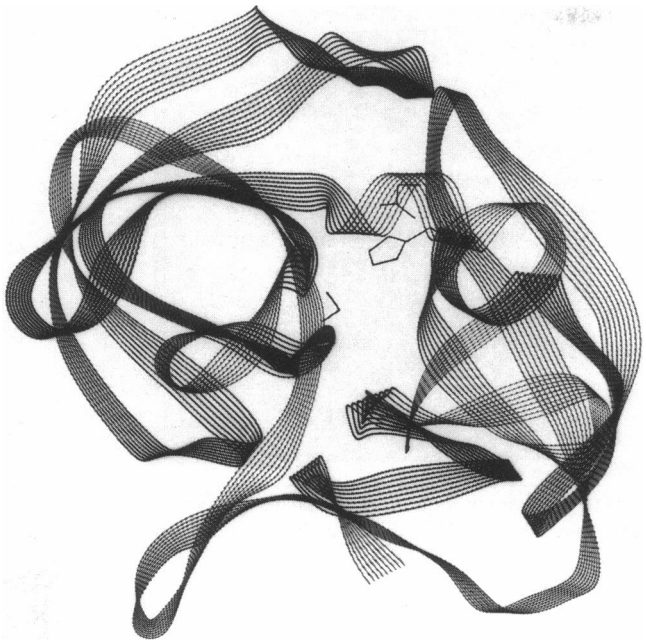


FIGURE 1 Computer-drawn ribbon schematic of a serine protease molecule (PDB file 2SGA; James et al., 1980). It has two domains of antiparallel β structure, with the active-site side chains in a cleft between them.

between parts of a protein chain. One simple method is adding H-bond lines to a stick figure (as in Kinemage 1), using dotted or thinner or less brightly colored lines than for the covalent bonds. Van der Waals contacts can be represented directly with a modified form of dot surface, as shown in Fig. 2 and later in Kinemage 7. These "small-probe contact dots" (used in Richardson and Richardson, 1987, 1989*b*) are calculated by bumping a small radius probe around the surface of each atom, as in the standard Connolly algorithm. However, instead of leaving dots only on the open areas where the probe touches nothing else, you do the opposite and leave dots only on the contact portions, where the probe does touch another atom. Wherever two atoms come within half an Ångstrom of actually touching, there will be a patch of dots from each side, color coded by atom type so that hydrophobic contacts are green and hydrogen bonds are slightly intersecting pairs of blue and red patches. This lets you evaluate in fine detail the goodness of fit between two pieces of molecule. The example is a slice through a helix–helix contact, showing that close to half of each sidechain's contact is with backbone atoms.

Another important part of looking at proteins is examining electron density maps for x-ray crystal structures, or the experimental constraints and preferably even the spectra for NMR structures. This lets one evaluate levels of accuracy, both for the structure as a whole and for particular parts of it, and shows subtleties that are not very well reflected in the coordinate data, such as the relationships of multiple, partially occupied waters, the contacts between molecules, or the nature of disordered regions.

An additional complication is that protein structures are really four-dimensional, because conformational changes are often very important. Movements are extremely difficult to study or show with physical models, but once you know the end states they are easy to show on the computer, either as both structures superimposed or by switching back and forth between them. For example, some mutants of T4 lysozyme have the two domains in different relative positions in the crystal structures (e.g., Jacobson et al., 1992). Plate 6 is a superposition of three of those structures, showing the "hinge bending" motion of the domains. Kinemage 2 "animates" those same changes and also a relative twisting motion of the domains that occurs between other crystal structures. The animation allows the eye to perceive much more subtle and/or more complex changes. For an explanation of kinemages, see Richardson and Richardson, 1992.

Much of what we have discussed about types of representations is directed toward the researcher trying to find new, significant relationships in a protein structure. But a second, especially crucial role of models, drawings, and computer graphics is to make explicit a relationship that you have found, enabling other people to see it as well. This often can be done just by making the relevant part a heavier line or a brighter color, or by deleting most of everything else, but it always requires explicit effort. The total process of looking scientifically at proteins involves communication as well as perception.

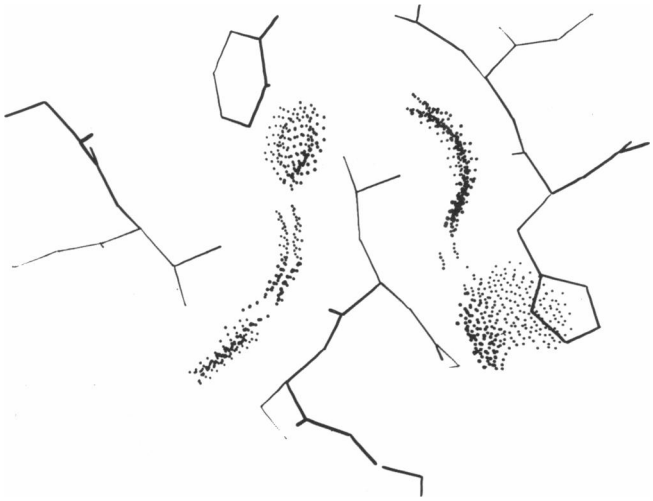


FIGURE 2 "Small probe" dots to show the van der Waals contacts between the A helix and its neighbors in the four-helix bundle of myohenerythrin (PDB file 2MHR; Sheriff et al., 1987). The central leucine, for instance, touches both sidechains (e.g., the Phe ring) and also backbone of other helices.

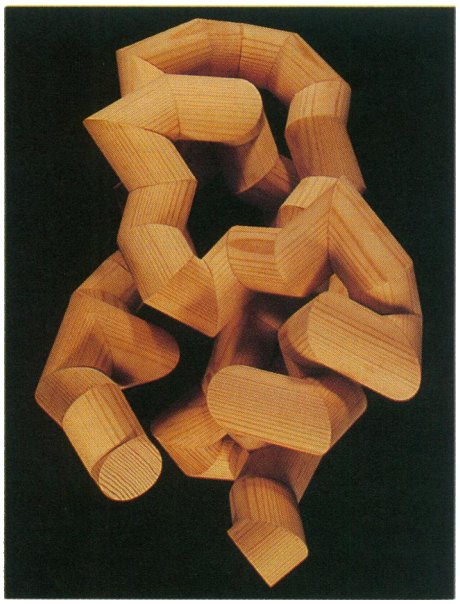


PLATE 5 Wooden model of the Cα backbone of basic pancreatic trypsin inhibitor (PDB file 5PTI; Wlodawer et al., 1984), sized to nearly touch at H-bonding distance.

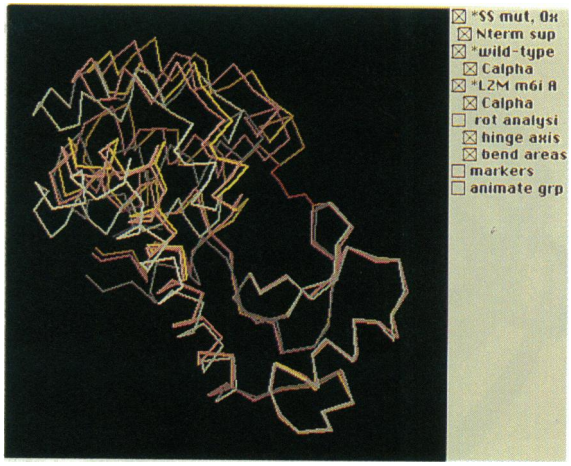


PLATE 6 Domain motions in T4 phage lysozyme (PDB file 3LZM; Jacobson et al., 1992). Cα's of the NH_2 terminal domain are superimposed, for wild-type (*pinktint*), disulfide mutant (*yellowtint*), and M6I mutant (*white*).

Looking at protein "folds" for clues to the folding process

One consideration that always seems present for us when looking at a protein structure is the question of how it got that way. That is probably because it was Chris Anfinsen who first got us interested in solving and studying protein structures, and like him we have always considered the major aim to be understanding protein folding. To represent that problem in a general way, plate 7 shows the one-dimensional sequence of pancreatic trypsin inhibitor and the three-dimensional structure into which it folds, with a ribbon backbone and full sidechains. As a metaphor for that process, plate 8 shows a flat sheet of paper and the object into which it can be folded by origami. In both cases, the starting material is simple and has few interesting properties, but the final product has complex structure and biological function: BPTI inhibits trypsin, and the canary might fly or sing. Also, in both cases, it is not obvious how the pattern of creases in the paper or the order of amino acids in the sequence produces the final structure. In the absence of an instruction book, one way of learning some of the folding rules is to study, for instance, origami pieces of many kinds of birds, to see how the paper layers in the wings are arranged relative to the tails, and then correlate that with

PLATE 7 3-D structure of trypsin inhibitor (ribbon backbone, stick figure side chains) versus the one-dimensional amino-acid sequence from which it folds.

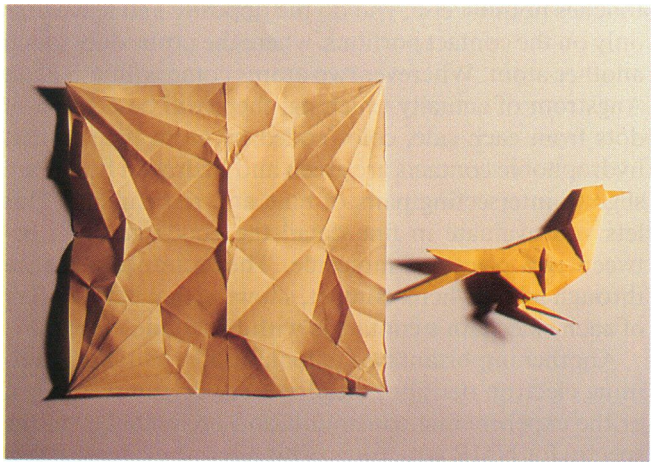


PLATE 8 An origami canary versus the creased piece of paper from which it was folded.

the crease patterns on the pieces of paper. We will now look at several different groups of protein structures to see if there are illuminating regularities in their arrangement that suggest what can or cannot be true about their folding.

The first such structure/folding issue concerns the organization of parallel α/β structures. Since each strand goes in the same direction in a parallel β sheet, the connections need to get back around to the starting end of the sheet somehow; they do this in what is called a crossover connection, which lies across one surface of the sheet and often includes an α-helix. Fig. 3 shows that the crossover has a choice of two possible directions to go; one choice gives a right-handed spiral and the other a lefthanded one. You can test the handedness by moving your thumb from the first strand to the second strand and letting the fingers curl around to follow the path of the connection; if that works with the right hand, the crossover is righthanded. The interesting thing is that in real proteins, with only one well documented counterexample, crossover connections are always right-handed (Richardson, 1976). That fact puts very strong constraints on all of the structures in the parallel α/β category.

The simplest parallel α/β structures are the singly wound barrels, of which triose phosphate isomerase is the classic example (plate 9 shows a side view, and Fig. 4 an end view). They contain a central cylinder of eight parallel β strands, surrounded by an outer cylinder of eight helices from the right-handed crossover connections. The chain goes up in a strand, around in a helix loop, and back into the next strand over; it winds around the barrel, always in the same direction and moving over by one strand each time. As an analogy to this structure type, plate 10 shows the Castel del Monte in Apulia, Italy. The central courtyard is the hydrophobic core, with the main part of the castle as the β barrel, the eight towers as the helices, and the entrance as the chain termini. This is obviously a sound structural design, although the castle is not as graceful as the protein because it is not twisted and it does not have the interplay of the sequence winding through the structure. The protein

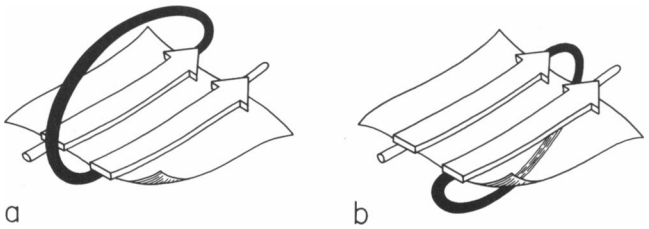


FIGURE 3 Connections between parallel strands in β sheet: (*a*) right-handed vs. (*b*) left-handed crossover connections. Right-handed ones predominate overwhelmingly in the known protein structures.

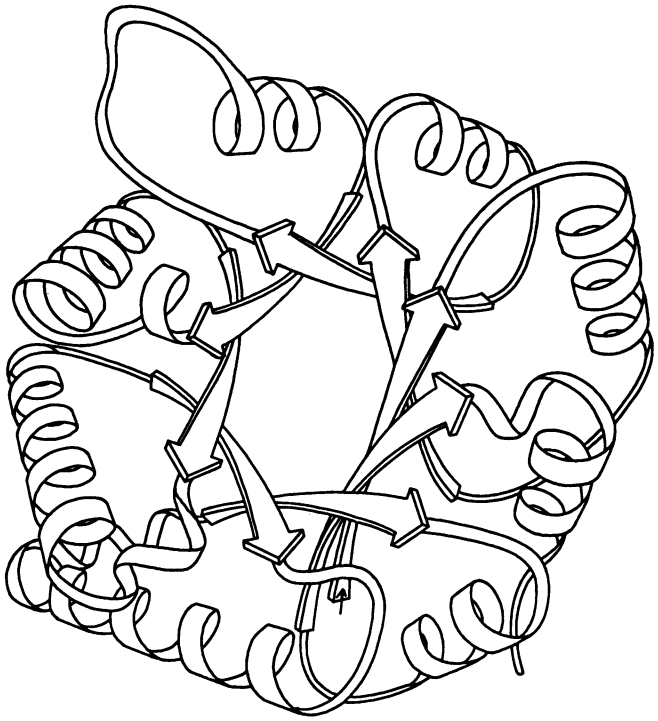


FIGURE 4 Hand-drawn ribbon schematic of the classic "TIM barrel" fold of triose phosphate isomerase (PDB file 1TIM; Banner et al., 1975), in end view. Arrows show the central ring of eight parallel β strands, surrounded by an outer ring of eight helices in the crossover connections.

version would not work as well under the demands of gravity, however.

The commonest α/β structure is the doubly wound β sheet, many of which are nucleotide-binding domains, as first discovered in a comparison of lactate dehydrogenase with flavodoxin (Fig. 5; Rao and Rossmann, 1973). When initially discovered, such pairs of structures were considered so similar as surely to be related, but later it was realized that they represented especially favorable arrangements that could also arise by convergent evolution. Kinemage 3 demonstrates why this structure is called "doubly wound", by building up the chain sequentially from NH_2 to COOH terminus. It starts in the middle of the sheet and winds to the left, depositing helical crossovers on the back side of the sheet, and then switches back to the middle and winds to the right, putting helical loops on the front side. This tertiary structure is shaped by the demands of using right-handed crossovers, protecting both sides of the parallel β sheet, and preferring to move over by only one strand at a time. The reverse doubly wound structure, which would wind from the edges in to the center and have the COOH-terminus in the middle, seems equally plausible a priori. However, it does not ever occur, either

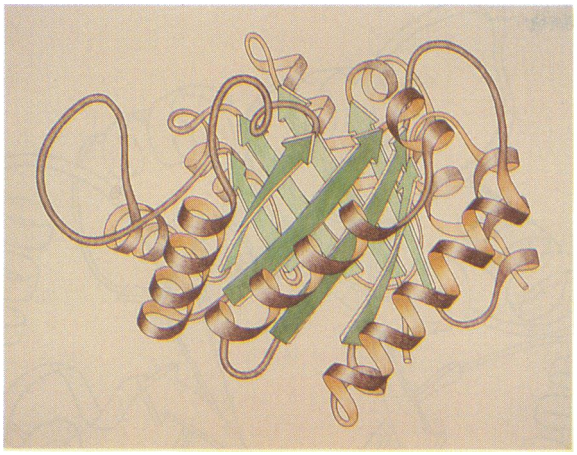


PLATE 9 Ribbon schematic drawing of triose phosphate isomerase (PDB file 1TIM; Banner et al., 1975), showing the central eight-stranded parallel β barrel as green arrows and the outer ring of α-helices as brown spirals.

because of a preference to start folding at the NH_2-terminus, or because perhaps all of these doubly wound proteins are actually related to one another after all. It still remains a puzzle, in fact, why all crossovers are right-handed (the resulting structure should be somewhat more stable, but not enough to account for such an overwhelming preference); perhaps an early folding step involves a concerted, handed, loop formation, but there is no firm backing for such a process in what is understood of the energies involved.

The second structure/folding issue concerns the formation of antiparallel β sheets. As an introduction to the relevant features, Kinemage 4 and Fig. 6 both show a β hairpin out of trypsin inhibitor. It is a two-stranded β ribbon with the usual β-sheet twist (Chothia, 1973), which is right-handed as you look along the strand direction; the twist is an indirect consequence of the handedness of the amino acids. For antiparallel sheet, the backbone hydrogen-bonds come in pairs, alternating a narrow pair and a wide pair. The sidechain direction also alternates from one side of the sheet to the other, with adjacent sidechain pairs on the two strands in register. Therefore, on one side of the ribbon (the *top* side here) sidechains stick out from between narrow pairs of H-bonds, and on the other (*back*) side they stick out from between wide pairs of H-bonds. Salemme (1983) has shown that those two sides of a β ribbon are not quite equivalent, and that conformational preferences cause the ribbon to curl somewhat toward the narrow-pair side. That can be seen in this example, which has a rather pronounced curl toward the side facing us. Another obvious property of β hairpins is that the two strands are neighbors in the sequence, since they are joined by a turn at one end. For protein structure in general, entropy favors the interaction of two adjacent pieces of the chain rather than two that are distant in the sequence.

Indeed, in β sheet proteins, such as the immunoglobulin domain in plate 11, most of the connections are hairpins between adjacent strands (shown in purple). Some β proteins have all their connections adjacent, but the great majority include connections that skip over several strands, like the blue ones in plate 11 which skip two strands and cross over an end of the barrel. We call this arrangement a Greek key structure (Richardson, 1977), because if you visualize the β sheet as a cylinder, open it out flat and make a "topology" diagram of how the strands are connected, you end up with a pattern (Fig. 7) that looks like the classic design used to decorate Greek vases (see plate 12). This is one of the natural patterns that lets a line do something interesting without crossing over itself, but the question remains as to why this should be a favored structure for proteins, especially since it violates the adjacency rule. This puzzled us at first, but we think the only plausible explanation lies in a favored folding pathway. If the protein initially formed a long, two-stranded β ribbon with the normal twist, then that pair of strands could curl up into a structure that

PLATE 10 Castel del Monte (Apulia, Italy), which has an eight-fold structure around a central opening and eight outer towers, each containing a right-handed spiral stair. Photo courtesy of Kasper Kirschner.

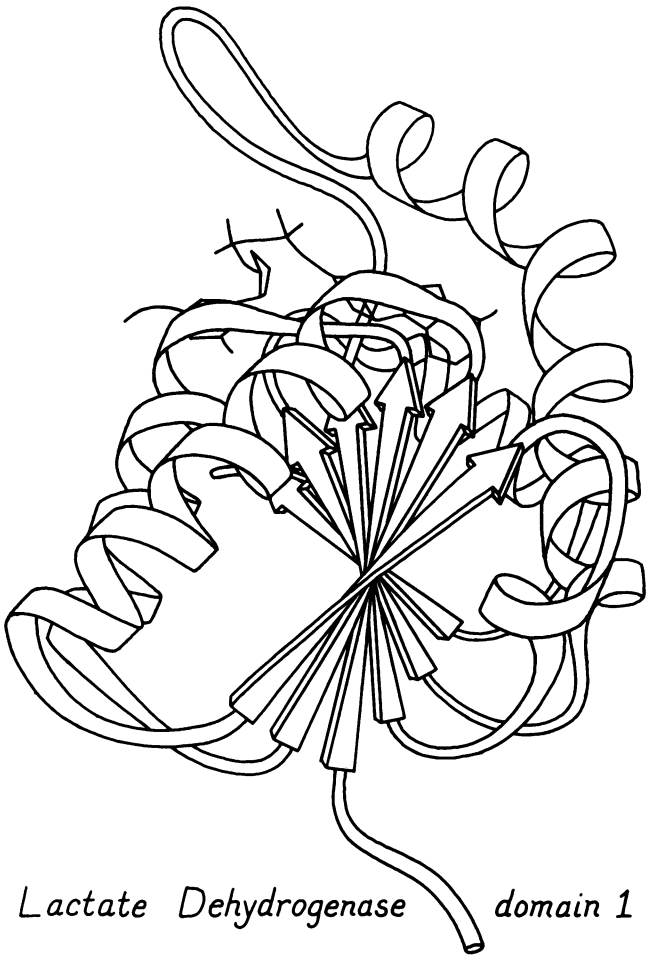


FIGURE 5 Schematic ribbon drawing of the classic "doubly-wound" backbone fold of the nucleotide-binding domain from lactate dehydrogenase (PDB file 6LDH; Abad-Zapatero et al., 1987). The chain starts in the middle of the β sheet and winds to the front with loops at the left, then skips back to the middle and winds toward the rear with loops at the right (animated in Kinemage 3).

produces the Greek key topology automatically, as shown in Fig. 8. The Greek key connections make sense if the protein is going together two strands at a time rather than one strand at a time.

To appreciate the organization of a Greek key protein, see Kinemage 5, which explores the eight-stranded Greek key β barrel of satellite tobacco necrosis virus protein (STNV). First the eight strands are built up sequentially from N H_2 to COOH terminus, and then instead they are built up by the strand pairs of the proposed Greek key hairpin. Fig. 9 shows this strand pairing on a schematic drawing of STNV. For all of the known Greek key barrel structures, the optimal strand pairing when judged by number of strands and number of H-bonds is also the one with its hydrophobic sidechains on the side of preferred curl (see above), which should promote its curling together and then self-associating during folding. Recent findings, both theoretical ideas like the preferred direction of curl and experimental results such as the organization seen in newly determined β barrel proteins, all give circumstantial support to the concept of Greek keys folding as two-stranded ribbons; however, there is still absolutely no direct experimental evidence on this issue.

Plate 13 *a* shows γ crystallin, with its six β strands (out of eight) of Greek key topology highlighted; the paired strands have the same color, shading from yellow at the ends to dark orange near the central hairpin turn. This representation emphasizes our ideas about how these domains might fold. Of course that is not the only useful thing to show in this structure. Plate 13 *b* color codes the sheet structure of each domain, with one four-stranded sheet in yellow and the other in green; this emphasizes the packing of the β sheets in the final structure. Plate 13 *c* shows γ crystallin from an evolutionary perspective, where the color coding shows four-strand units that are homologous in both sequence and structure. Each unit includes three strands from one sheet and one from the opposite sheet. There was almost certainly one duplication of the four-strand unit to make the domain structure, and another duplication to create two similar domains (all three representations show the twofold relationship between the domains). These are three very different descriptions, but all of them are reasonable and informative. One advantage, in fact, of these schematic drawings is that people realize they involve subjective

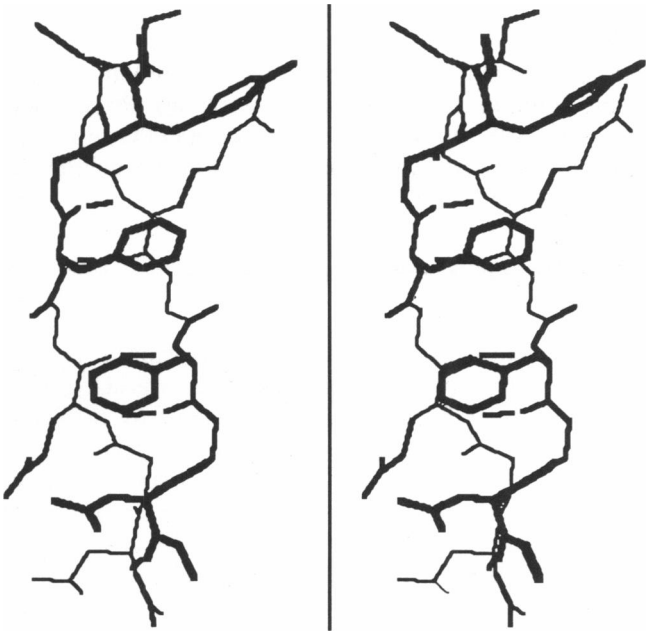


FIGURE 6 Stick figure, in stereo, of the two-stranded β hairpin from basic pancreatic trypsin inhibitor (BPTI; PDB file 5PTI; Wlodawer et al., 1984), taken from Kinemage 4. Main chain atoms, H-bonds, and side chains on the inside curve of the hairpin are shown; those side chains are between narrow pairs of H-bonds.

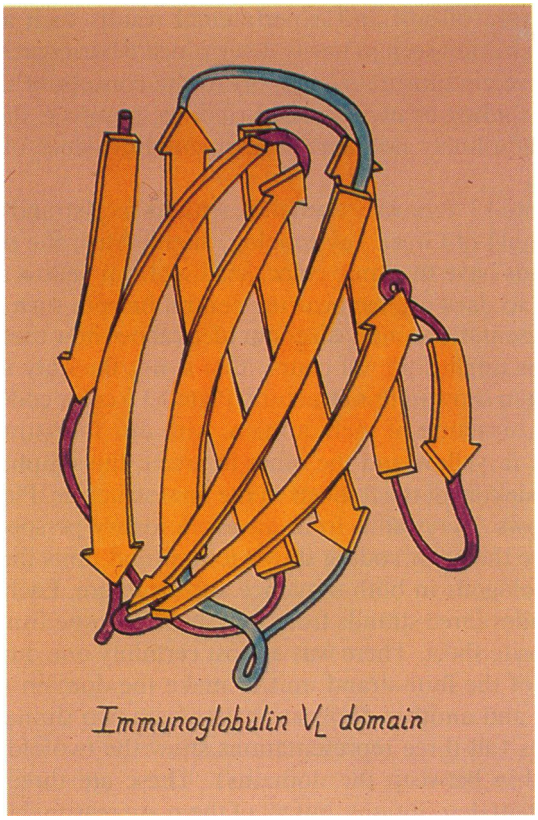


PLATE 11 Ribbon schematic drawing of an immunoglobulin V_L domain, with near-neighbor connections between β strands in purple and "Greek key" connections in blue.

interpretation. It is worth remembering, though, that the same thing is true of computer drawings; the message that gets communicated is strongly dependent on decisions about viewpoint, coloring, and what to include and not to include.

Testing the ideas from looking: protein design

As is evident from the Greek key discussion, it is often frustrating to have the protein structures suggest hypotheses about their folding, but to have no way of testing those ideas experimentally. More than ten years ago, we decided that one possible way to answer such questions would be by trying to design new proteins from scratch. De novo design is the inverse of the prediction problem: instead of starting from a sequence and predicting the three-dimensional structure it will fold into, you start with a proposed structure, choose a sequence that would fold into that structure, and then make it and see whether indeed it does. The design project ties in closely with looking at proteins, from two different directions: first, it allows us to test some of the ideas that came from looking at known structures; and second, it has provided an entirely different perspective which has suggested new and very productive questions to ask and things to look for in those structures. The following discussion will try to give a brief idea of what the design process is like, describe some of those new perspectives on what to look for, and summarize where we think the field stands at present.

The first step in a protein design is the choice of a tertiary structure type and of the specific backbone framework that is to be the design goal. A "de novo" design is one that is not based on, or homologous to, any specific natural protein, although it usually is meant to embody the simplest common core of some set of similar proteins. The two designs we have worked on the most, Betabellin and Felix, are based on proteins like the two shown in plates 14 and 15: Cu, Zn superoxide dismutase to represent antiparallel β barrels and cytochrome b562 to represent four-helix bundles.

Plate 16 shows a ribbon drawing of Betabellin, which has a four-stranded, simply connected up-and-down β sheet in the front and another identical sheet in the back (Richardson and Richardson, 1987, 1989*a*). Betabellin is made by peptide synthesis, and in the original form incorporated a two-armed cross-linker which attached to

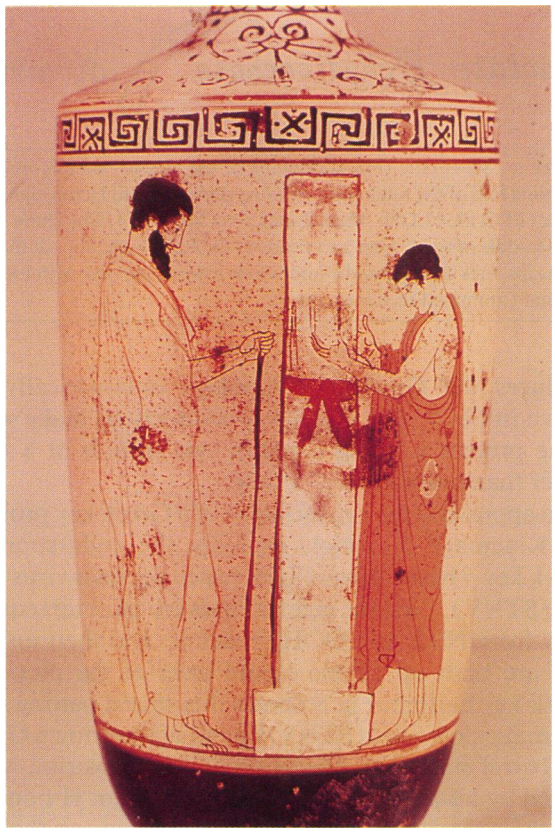


PLATE 12 The traditional "Greek key" border pattern, on an Attic white lekythos (5th century B.C.; National Museum, Athens).

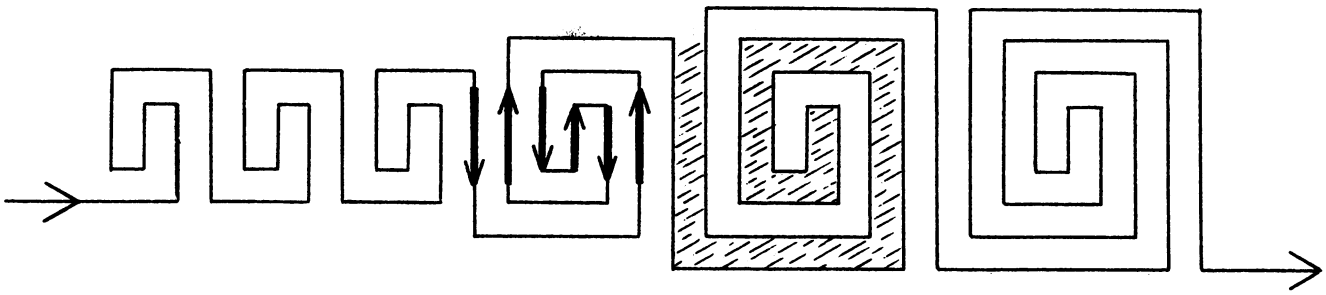


FIGURE 7 Stylized "Greek key" patterns of increasing complexity, all with the counterclockwise swirl direction of protein β barrels viewed from the outside. One of the motifs has arrows to show how the β strands fit in, and one has shading between the strands that curl around together as a pair.

the resin at the top of the drawing, and from which the two identical half-chains could be synthesized at the same time (Unson et al., 1984). There is a disulfide bridge across the inside of the β barrel, like the one which connects the two sheets of an immunoglobulin domain. The hairpin turns are meant to be short, separated by 6-residue antiparallel β strands, so that each sheet includes 32 residues.

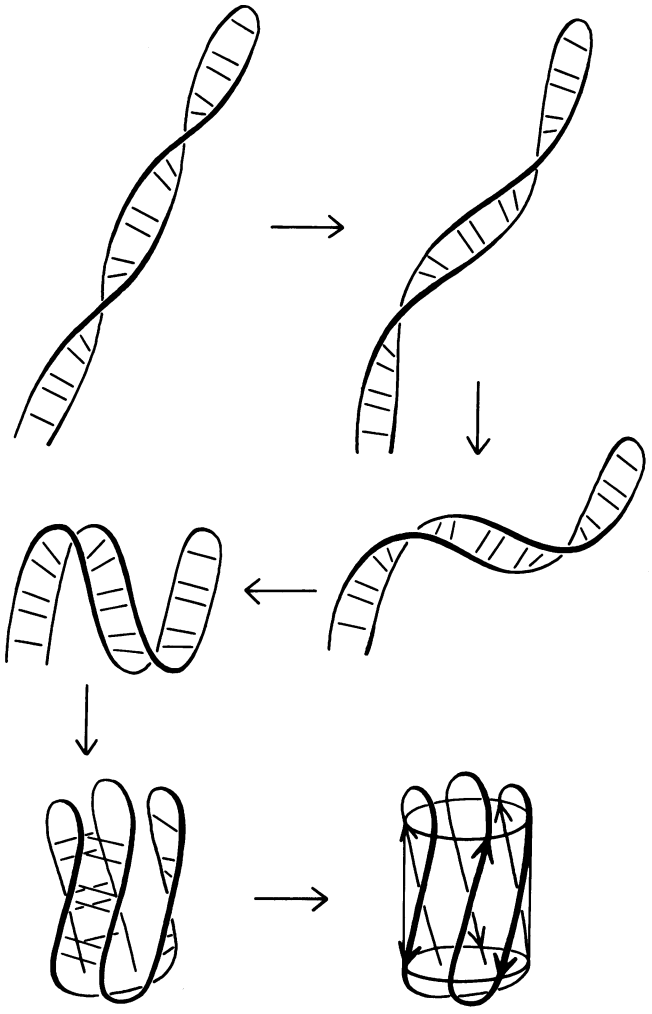


FIGURE 8 Hypothetical scheme for how a two-stranded β ribbon could fold up into a β barrel, with the original preference of twist and curl directions automatically producing the Greek key topology, the swirl handedness, and the choice of strand pairing that are seen in the known structures of this type.

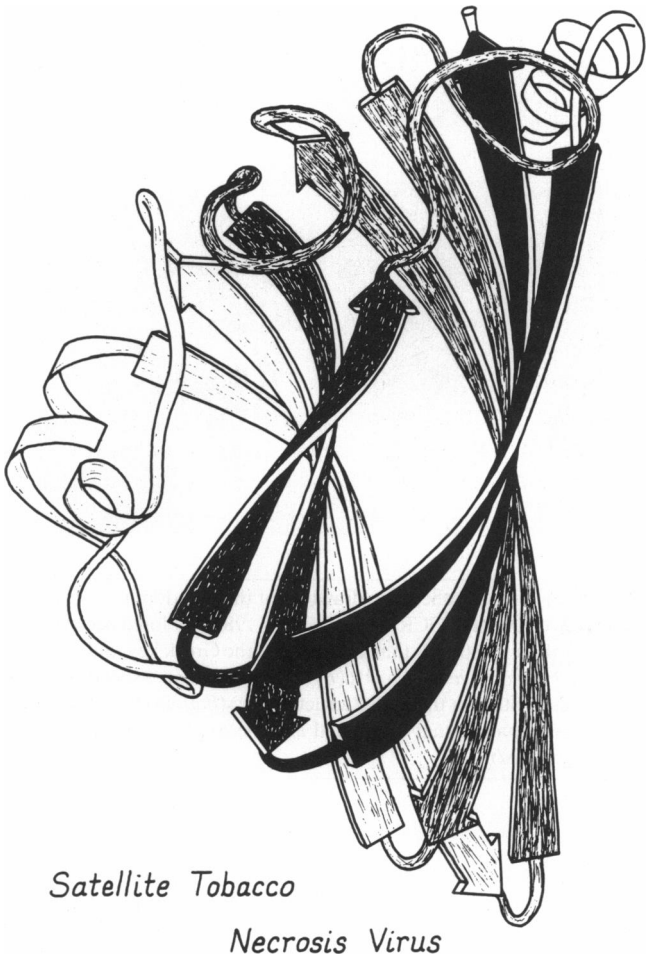


FIGURE 9 The Greek key β barrel of satellite tobacco necrosis virus (STNV; PDB file 2STV; Jones and Liljas, 1984), with matched shading on the preferred pair of strands that wind around side-by-side to form the structure. Alternative animations are shown in Kinemage 5.

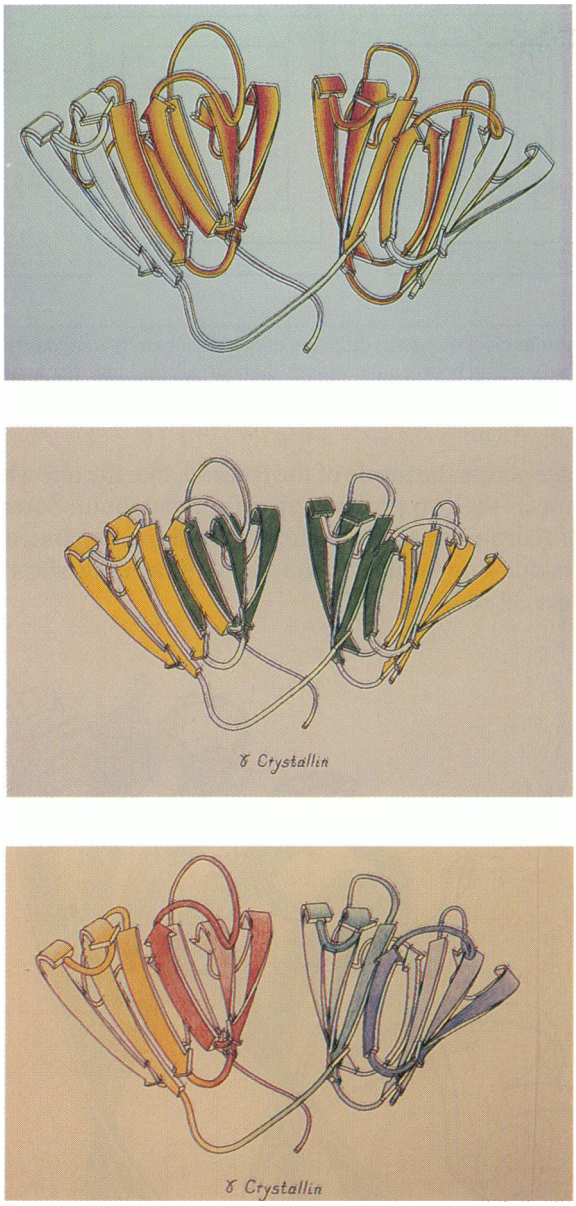


PLATE 13 Ribbon schematic drawings of the two-domain γ-crystallin molecule (PDB file 1GCR; Wistow et al., 1983), colored according to: (*a*) (*top*) the six β strands that are paired in the Greek key folding model (*shaded from orange to yellow*); (*b*) (*middle*) the two β sheets that pack against each other in the folded structure; (*c*) (*bottom*) the four-regions along the sequence that originated as gene duplications (*blue, cyan, yellow,* and *red*).

Felix is a four-helix bundle of 79 residues made by gene synthesis and expression methods (Hecht et al., 1990); its structure and sequence are shown schematically in Fig. 10. Felix uses 19 different amino acids, in a nonrepeating, native-like sequence. As in most helix-bundle proteins, each α-helix in Felix is connected to its nearest neighbor, going up and down around the bundle. Cytochrome b562 is an example of the commoner of two versions, in which the chain turns to the right at the top of the first helix (looking from outside); Felix was designed to turn to the left at the top of its first helix, like the less common examples, so that success would be somewhat stronger evidence that we could actually control the topology.

We have also worked on several other de novo designs, mostly summarized in Richardson and Richardson, 1989*a*. They include one singly wound and one doubly wound parallel α/β structure, and also Betadoublet, a β sandwich similar to Betabellin but with a quite different internal arrangement of side chains and made by gene synthesis and expression (Quinn et al., 1991).

The design process involves many considerations, some quite straightforward and others less obvious. On the straightforward side, the designed sequence should for instance predict the correct secondary structures by various algorithms, and the model should put hydrophobic sidechains on the inside and hydrophilics on the outside (e.g., the helix wheel in plate 17). We have tried, in fact, to take into account everything people know, or think they know, about the determinants of protein structure, and to make the best compromises between those criteria when they conflict. A given design is also influenced by how it will be made, such as including an NH_2-terminal Met and convenient restriction sites in the Felix sequence. Then there are two interesting general categories of more subtle considerations that had not occurred to us before starting. One is that we spend more and more of our design effort on avoiding alternative structures, and the other is that we often need to collect statistics on very much more narrowly defined categories than is usually done.

To illustrate this latter point, for choosing sequence to form helices it was not good enough to consider overall helix propensity or even to use occurrence in centers versus first and last turns, so we needed to tabulate occur-

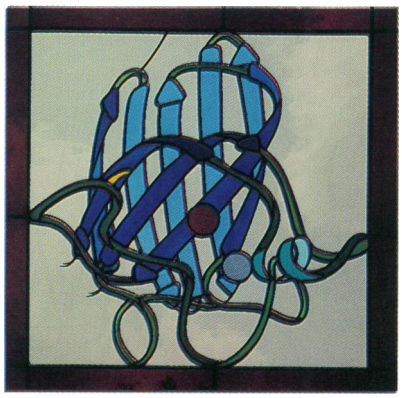


PLATE 14 Ribbon schematic (in stained glass) of the Cu, Zn superoxide dismutase subunit (PDB file 2 SOD; Tainer et al., 1982), with the antiparallel β barrel in shades of blue.

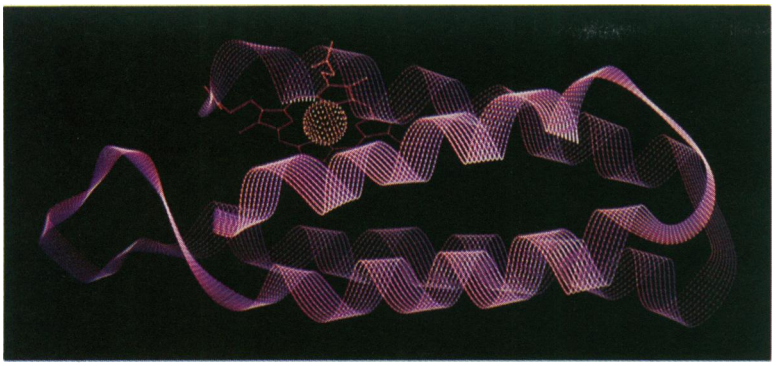


PLATE 15 A natural four-helix bundle protein: backbone ribbon of cytochrome b562 (PDB file 256B; Lederer et al., 1981) in purple, with red heme.

rence frequencies in individual specific positions relative to the helix ends. This meant defining the exact endpoints in a robust way. Most helices have a residue at each end with the peptide on one side definitely part of the helix and the peptide on the other side definitely doing something else (such as the Cα marked as "C-cap" in plate 18); some definitions include that residue as helical but most do not. The key to this difficulty is to give the ambiguous position a new name and use it as the reference endpoint: we call it the N-cap residue when at the beginning of the helix and the C-cap when at the end. That choice gives much stronger correlations with amino acid occurrence frequencies than definitions based either on ϕ, ψ angles or on hydrogen bonding (Richardson and Richardson, 1988). For instance, in a plot of the normalized occurrence frequencies of Asn in each position relative to the N-cap (the *vertical dotted line* in plate 19 *a*), some adjacent positions vary by factors of 4. For Gly the C-cap position is a factor of 12 preferred over C − 1 or C + 1, and fully one-third of all helices end with Gly. In more generally defined statistics, few preference factors are as great as 2 or 3. A strong, specific preference is clearly useful when you are trying to control exactly where a helix should end.

Of course, in order to understand such statistics, we need to look at the protein structures to see what is actually going on. Plate 19 *b* is a superposition of 17 helices that start with an Asn in the N-cap position. The asparagine side-chains are shown in green, because in black and white they look exactly like peptides. There is an H-bond from the Asn sidechain Oδ to one of the backbone NH

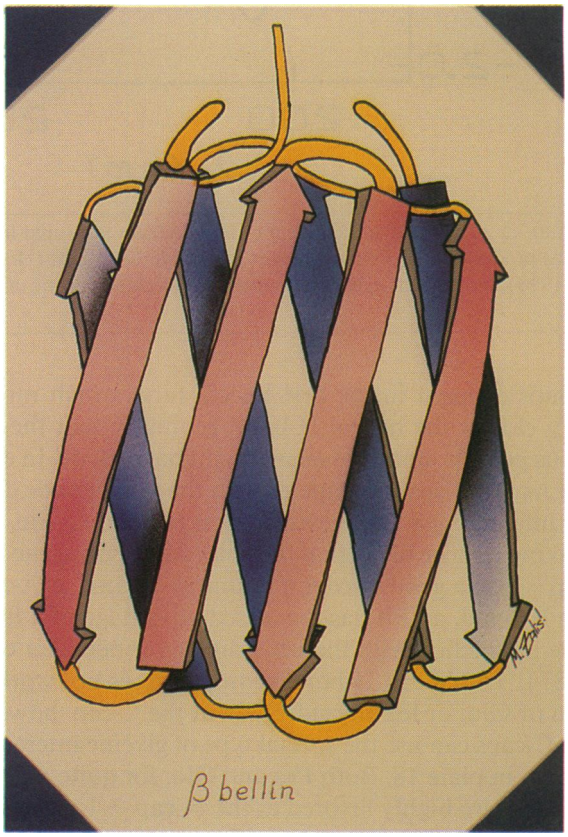


PLATE 16 Ribbon schematic drawing of the designed structure of Betabellin (Richardson and Richardson, 1987), with two identical four-stranded β sheets (*pink* and *blue*) joined by a two-armed crosslinker (*yellow*).

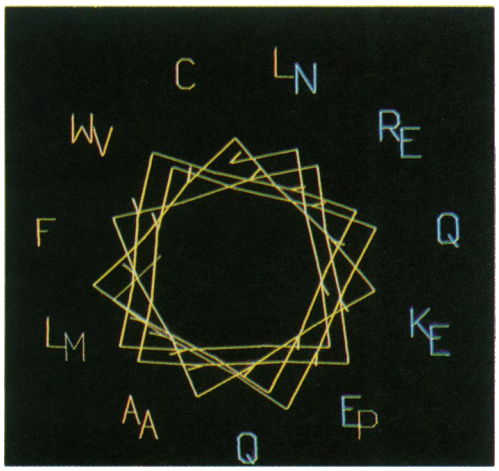


PLATE 17 End view of the first α-helix of Felix, laid out as a "helix wheel" with residue labels at the Cβ positions. Hydrophobics are in orange and hydrophilics in blue, showing the buried and exposed sides of the helix.

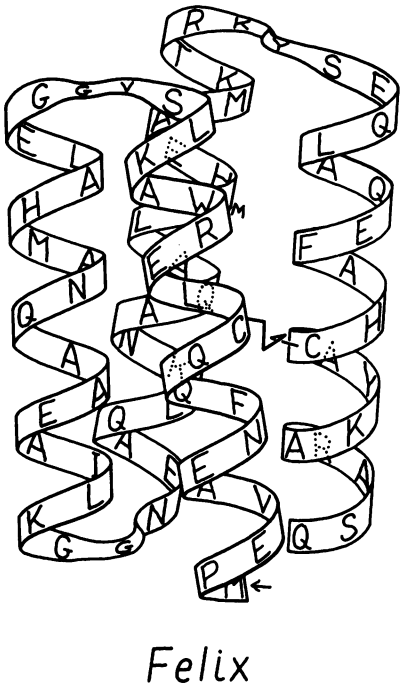


FIGURE 10 The designed amino-acid sequence of Felix (PDB files 1,3FLX; Hecht et al., 1990), shown on a ribbon drawing of the intended four-helix bundle tertiary structure. The bundle is a "left-turning" one (that is, it turns left at the top of the first helix), and there is a disulfide between helices 1 and 4.

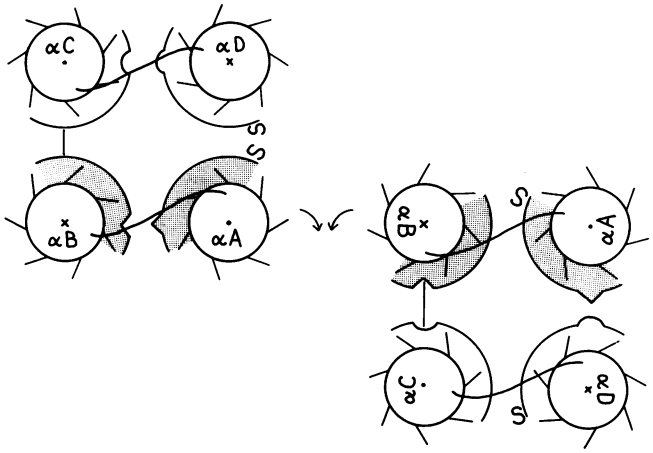


FIGURE 11 Stylization of the transformation between left-turning and right-turning helix bundles, with helices A and B as the reference pair, in end view. The hydrophobic half of their surfaces is shaded, with a triangular notch and projection to symbolize the A-B interaction in the original bundle, and SS for the disulfide. On the right-hand side of the figure, helices A and B have rotated inward like meshed gears; this brings their smooth contacts together and makes a hydrophobic surface for the rest of the bundle on the bottom rather than the top, so the sequence proceeds the other way around the bundle. In the right-turning version, the two Cys point outward from opposite sides of the molecule.

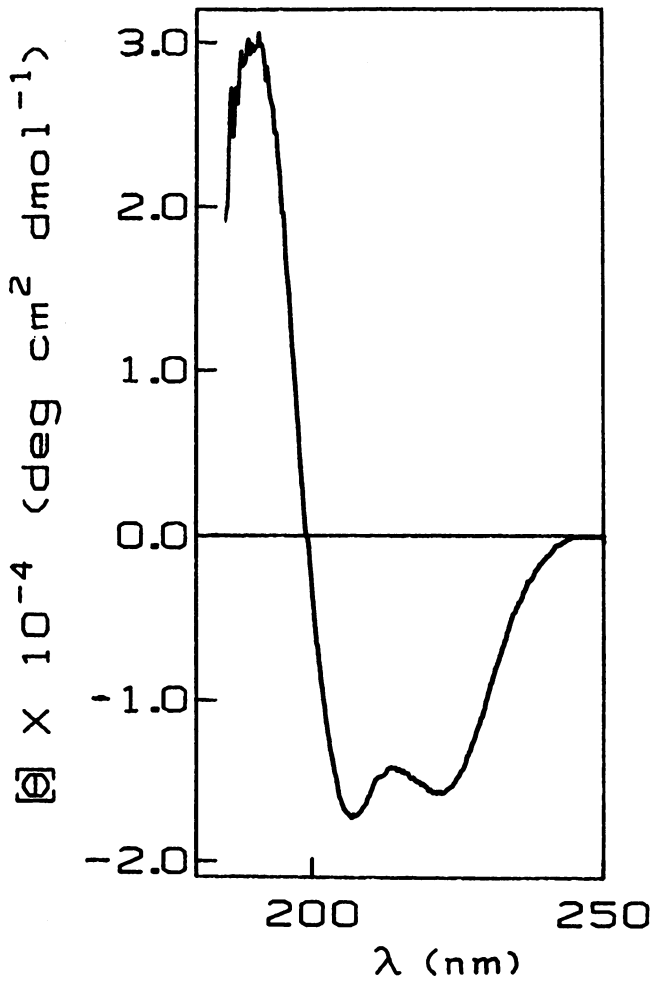


FIGURE 12 The circular dichroism (CD) spectrum measured for Felix, as produced by in vivo expression of the synthetic gene (Hecht et al., 1990). It shows the expected double minima characteristic of α-helix.

groups exposed in the first helical turn, which mimics very closely the helical H-bond geometry that the previous peptide along the chain might have made. In effect the Asn side-chain competes with the polypeptide chain for interacting at the previous open H-bonding site, thus making it more likely that the helix will begin at the Asn. Asp is also a good N-cap; it is almost as good as Asn for the H-bond, and it has the added advantage of a favorable interaction with the helix dipole (Shoemaker et al., 1987). The N-cap H-bond geometry does not work for Gln or Glu, or for any sidechain at the C-cap; however, the C-caps can use the special type of glycine interaction shown in plate 18. Both Pro and Glu, for quite different reasons, are highly favored at the N-cap + 1 position.

In general, the strongest preferences are seen when categories are defined narrowly; this means, unfortunately, that the database of known structures is still not nearly large enough for the purposes of protein design (or, one would suppose, for good prediction). But each time the

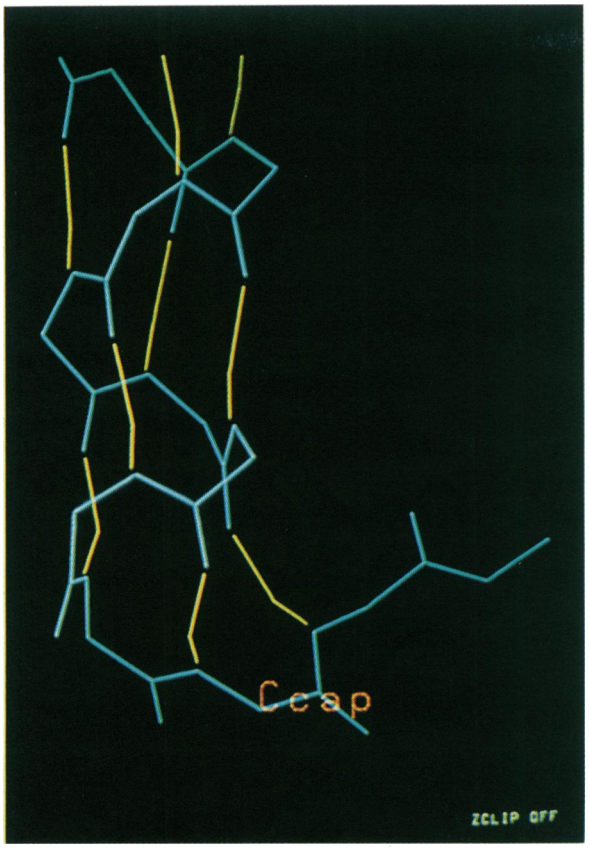
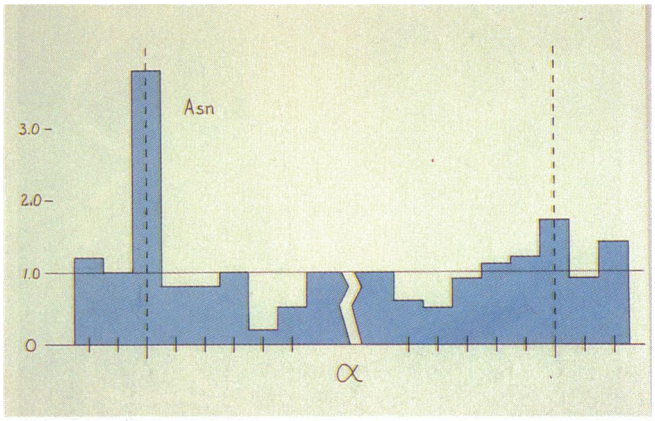


PLATE 18 The "C-cap" residue of an α-helix: the ambiguous residue at the COOH-terminal end that is half in and half out of the helix. Its Cα is in the right position, but its ϕ,ψ values are non-helical and it starts the chain off in a new direction. In this case the C-cap is a Gly in Lα conformation, which can make two H-bonds in the reverse order.

database doubles in size, we will be able to gather reliable statistics on twice as many specifically defined subcategories.

The second general lesson we have learned about the design process is the necessity for what we think of as "negative design". The worst problem with any kind of molecular modeling is that showing a sequence fits well with one particular structure does nothing to prove there is not another structure it fits even better. Both modeling and design are still quite unreliable processes, as one can tell by the fact that no one ever builds a model of a given sequence into a certain structure and decides it does not fit. In making our designs, we now spend a lot of our effort trying to ensure incompatibility with the more obvious alternative structures. The most interesting part has been looking at the structures of natural proteins from this inside-out perspective and trying to figure out what they are doing to actively avoid structural alternatives. We are beginning to find some sequence choices that do nothing favorable for the native structure but are needed to keep the protein from doing something else all or part of the time.

One example of negative design involves trying to control the directionality of the four-helix bundle in Felix. A helix wheel, like the one in plate 17, shows that about half the surface of a helix is hydrophilic and half hydrophobic. However, for a four-helix bundle, the interior hydrophobic side actually consists of two sections, each of which forms a contact to one neighboring helix. Plates 20 and 21 show how this arrangement fits into the context of the bundle: each helix cylinder has a blue hydro-



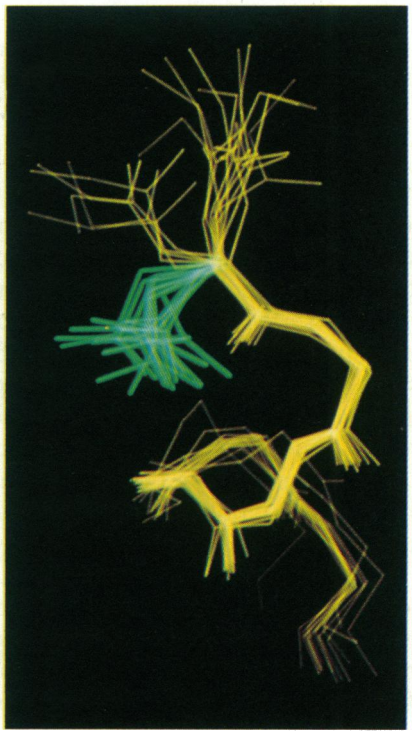


PLATE 19 Asn as a helix N-cap: (*a*) (*top*) Single-position helix preferences for Asn, referred to the N-cap and C-cap positions (*vertical dotted lines*). Asn has the strongest relative N-cap preference of any amino acid. (*b*) (*bottom*) Superimposed α-helices (*yellow*) with Asn N-caps (*green*), showing that the Asn side chain can H-bond to an exposed backbone NH and mimic the conformation of a helical peptide.

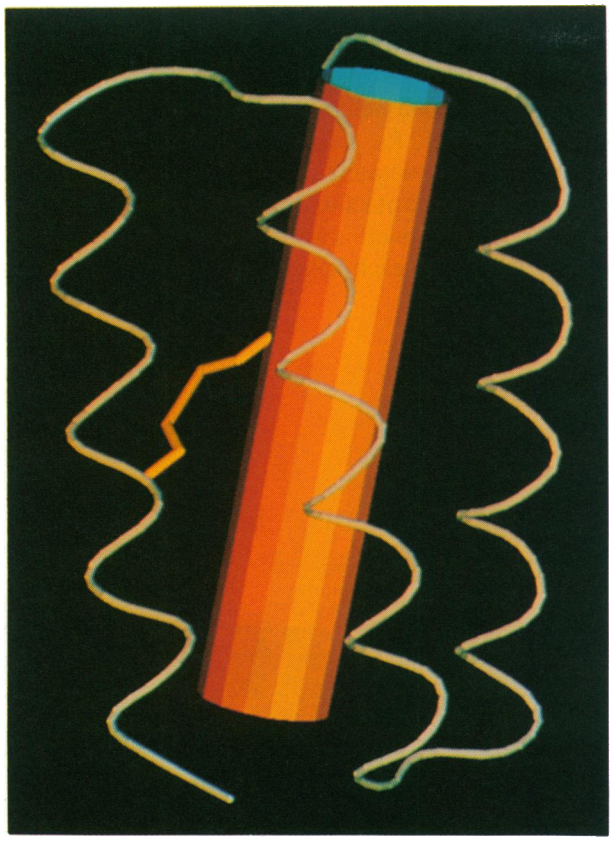


PLATE 20 Positioning of the two hydrophobic contact surfaces (*orange* and *yellow*) on one helix of a left-turning four-helix bundle.

philic half, and a yellow and an orange stripe to show the two hydrophobic contact surfaces. On each helical turn, the chain passes first through the orange stripe and then through the yellow stripe. Helices a and b (and c and d) put their yellow surfaces together and helices b and c (and d and a) put their orange surfaces together, and the chain turns to the left at the top of the first helix. This is only one of two possible arrangements, which a native protein never confuses although the contacts all look very much alike to people. In the alternative bundle topology, helices a and b put orange surfaces together and the chain turns to the right.

Fig. 11 is another way of visualizing this bundle-topology choice, using the a − b helix pair as reference. The contacts with bumps and cutouts correspond to yellow stripes and the smooth contacts to orange stripes. The left side of the figure shows the left-turning arrangement; imagine grabbing that a − b helix pair and rotating them 90° like two meshing gearwheels, ending up with their smooth surfaces together as on the right side. Now the leftover hydrophobic sides are on the bottom surface of the a − b pair rather than the top surface, so the other two helices must pack at the bottom and end up producing a right-turning bundle.

Thus, to produce a four-helix bundle with a specific topology we must somehow control which pairing of helix–helix contacts is preferred. For Felix we built models of both possible structures and tried to make the sidechains fit well in the left-turning form and badly in the right-turning form. We also tried to make the connections work better in the left-turning form. Most importantly, however, we designed a disulfide as a way of distinguishing which of the two structures is actually formed. Cys 11 and Cys 71 in Felix lie midway along the first and last α-helices, in good geometry to form a disulfide in the left-turning bundle. However, as seen in Fig. 11, for a right-turning bundle they point outward and are quite far apart. Therefore, if the disulfide forms easily and produces monomers the bundle must be left turning, while a right turning topology or a mixture should cross-link between molecules and make a mess. Since no one has yet obtained a three-dimensional structure of any complete designed protein, it is very valuable to have simpler probes of tertiary structure.

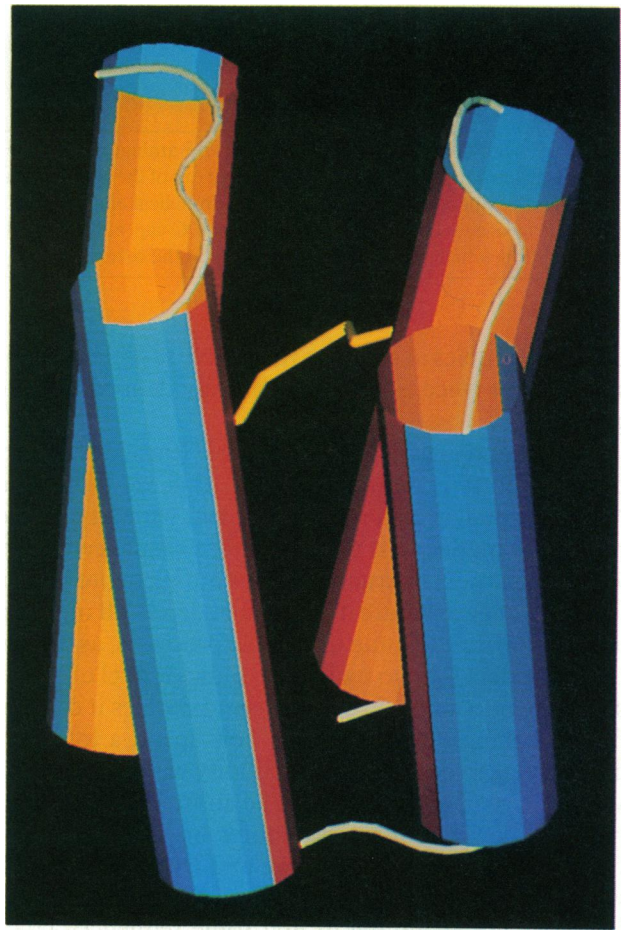


PLATE 21 Contact surfaces for all four helices in a left-turning four-helix bundle. Yellow surfaces interact front to back (helices a-b and c-d) and orange surfaces side to side (b-c and d-a).

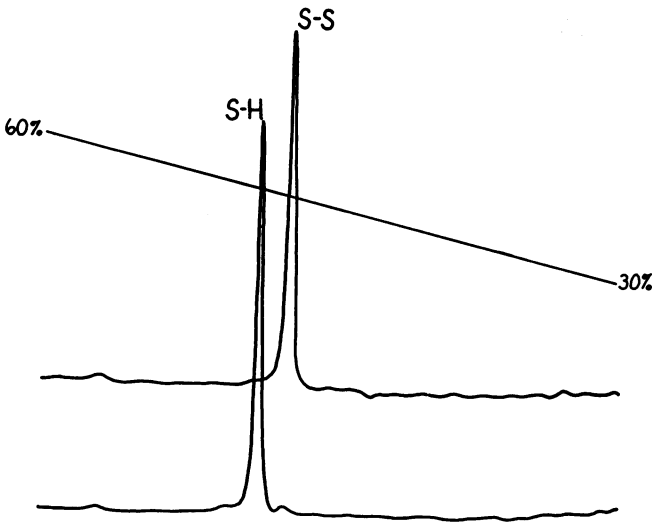


FIGURE 13 Comparison of elution profiles of Felix with the disulfide reduced and oxidized. At moderate concentrations all of the material can be oxidized with no formation of multimeric species. Since cross-linking would be expected for a right-turning bundle and most other alternative structures, the cleanly monomeric disulfide provides circumstantial evidence for correctness of the designed tertiary structure.

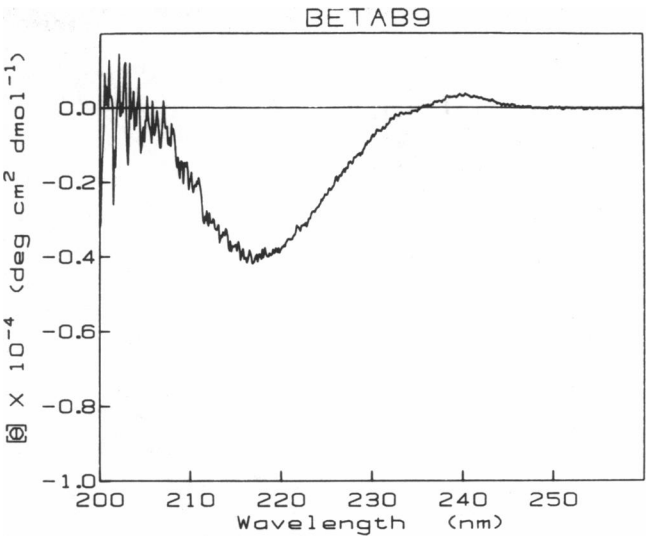


FIGURE 15 CD spectrum measured for Betabellin 9, as produced by peptide synthesis (Richardson and Richardson, 1989*a*), showing the shape characteristic of model peptides in β conformation.

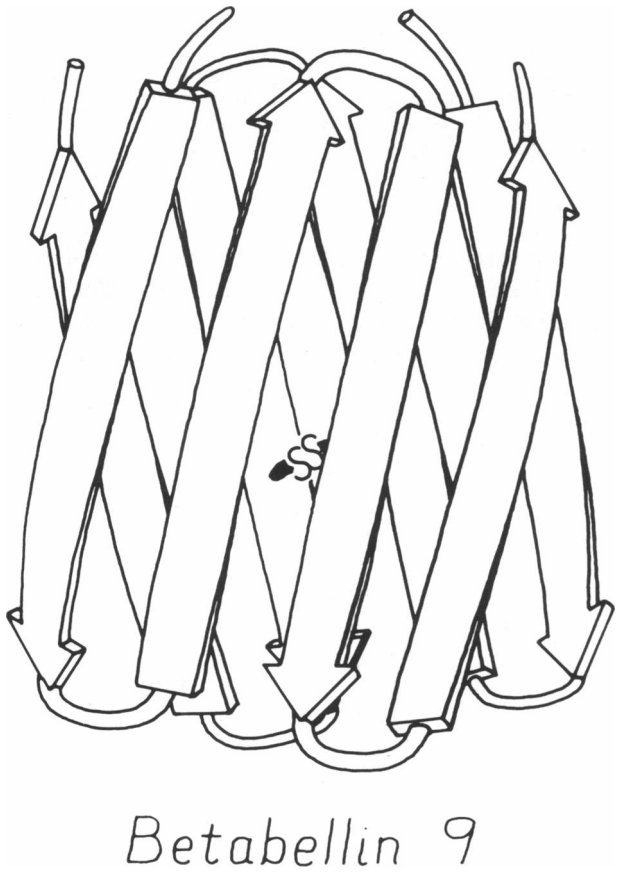


FIGURE 14 Ribbon drawing of the designed structure for Betabellins 9, and later, with a disulfide between the two identical β sheets but without the original cross-linker joining the two COOH-termini (see plate 16 for comparison).

The final result of the design process is just one word, like some sort of magical incantation. For Felix, it is: MPEVAENFQQCLERWAKLSVGGELAHMANQAAEAILKGGNEAQLKNAQALMHEAMKTRKYSEQLAQEFAHCAYKARASQ. Then one must produce and characterize the protein, in order to find out how good the design actually was. The answer, in general, is that they are quite surprisingly good, but still not quite good enough. Felix, for instance (Hecht et al., 1990), is helical (Fig. 12), soluble, seems to bury its tryptophan, and the disulfide forms cleanly within the monomer (Fig. 13), indicating that it forms some sort of helical bundle and has a left-turning rather than a right-turning topology, as intended. It shows cooperative unfolding with denaturant or temperature, but not very cooperative; it is only marginally stable. Its nuclear magnetic resonance (NMR) spectrum looks only partway between that of an unfolded and a native protein, presumably indicating disorder. We have designed and produced several minor changes to the Felix sequence, but their behavior is not demonstrably better.

Betabellin also seems to form approximately the right structure, but it took three cycles of redesign changes before it was soluble enough to study. The first change was to add ten more charges to the outside, which was all hydrophilic but originally had only four charges. β sheet peptides are notoriously insoluble, at least partly because the hydrophobic amino acids are almost all β formers while the charged residues prefer helix; however, native β proteins are quite soluble, and we need to learn how to emulate them. In this case the extra charges helped, but not as dramatically as we had hoped. The second change was to omit the cross-linker between the two halves of

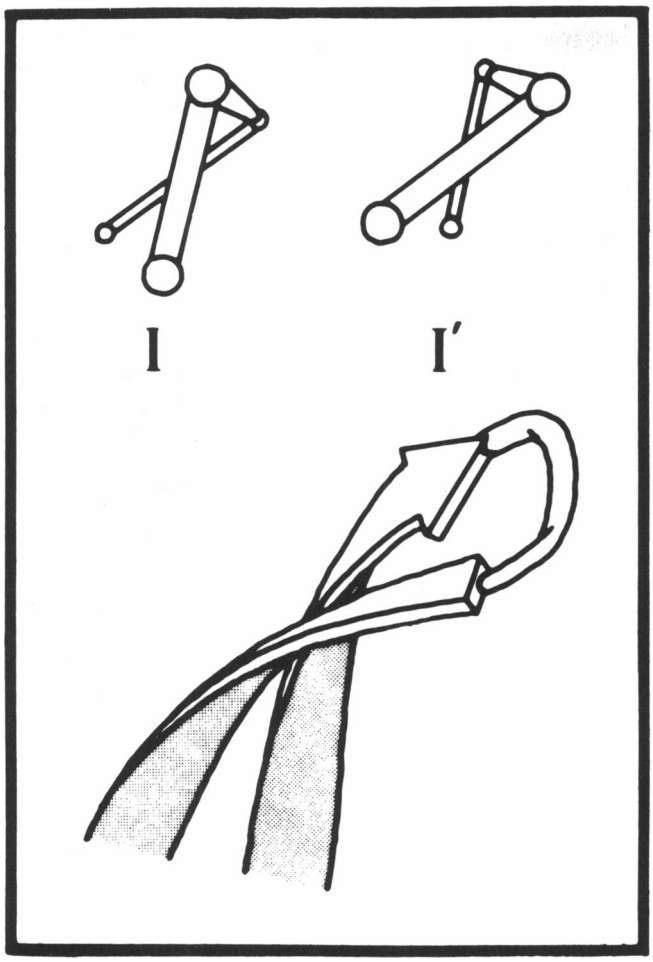


FIGURE 16 Simplified drawings of Type I vs. Type I′ tight turns, showing that Type I has a twist incompatible with the twist direction of β hairpins.

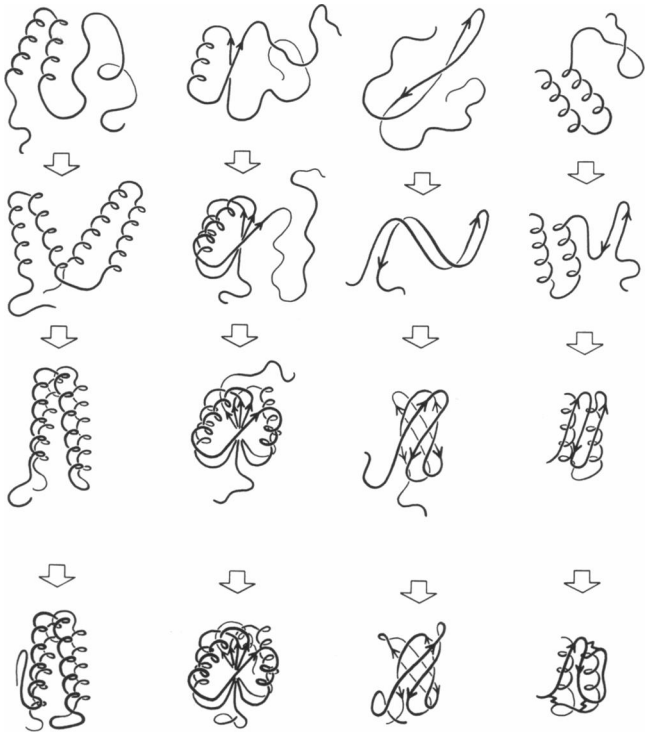


FIGURE 18 Hypothetical scheme of stages in the process of protein folding, with the formation first of partial secondary structure, then of approximate tertiary structure, and finally the minor but crucial rearrangements into a unique and well ordered native structure. Protein designs achieve approximately correct structures, but have not yet managed that final step.

Betabellin (see Fig. 14), for which there was no empirical database of structural information. This allowed the individual sheet peptides to be purified more easily, the disulfide formed well, the solubility improved signifi-

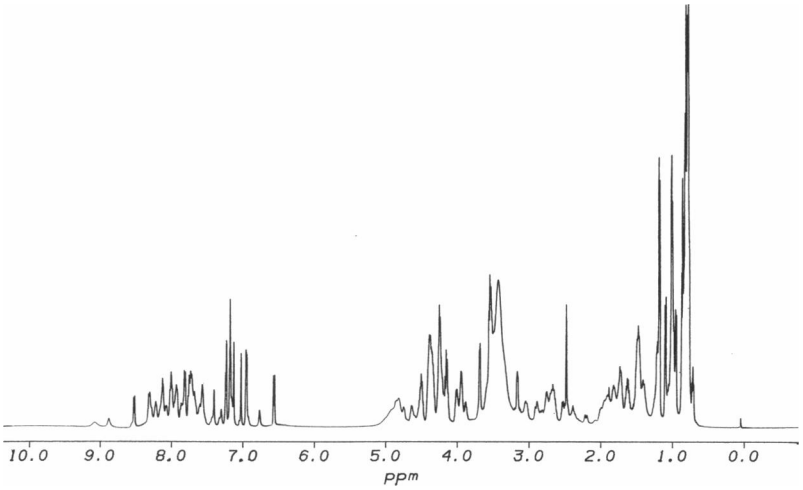


FIGURE 17 One-dimensional NMR spectrum of Betabellin 12, which uses D-Pro,D-Asp at the turns (McClain et al., 1992). Taken in deuterated DMSO at 25°C, on a Bruker 600 MegaHertz spectrometer at Glaxo, Inc. (see plate 22 for a two-dimensional spectrum).

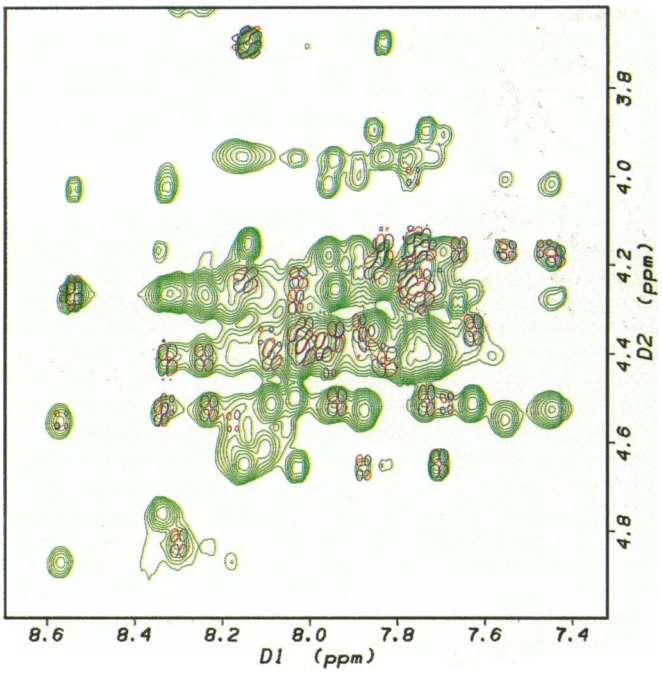


PLATE 22 Two-dimensional NMR of Betabellin 12: superimposed NOESY (*red contours*) and COSY (*blue* and *green tetrad peaks*) spectra in the NH – CαH "fingerprint" region (25°C, in deuterated DMSO; chemical shift reference: DMSO at 2.500 ppm). Dispersion of resonances was sufficient for sequential assignment.

cantly, and the CD spectrum of Betabellin 9 (Fig. 15) looked like that for a model β peptide. The third change was in the tight turns. After our original Betabellin design, Sibanda and Thornton (1985) studied turns for the narrowly defined subset of tight β hairpins and found an overwhelming preference for the otherwise rare Type I′

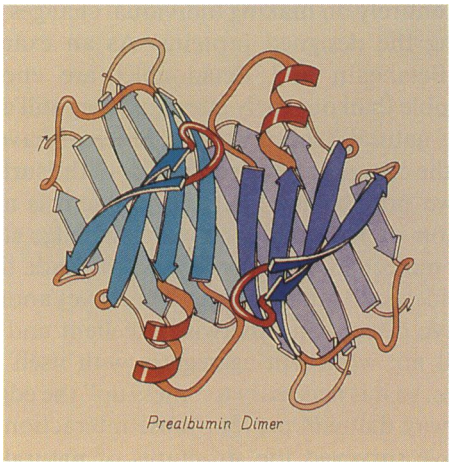


PLATE 23 Ribbon schematic of the prealbumin dimer (PDB file 2PAB; Blake et al., 1978). The blue and green subunits do H-bond at their inner β sheet edges, like the sort of interaction we wish to prevent in Betabellin; however, interaction at the outer edges is prevented with a strongly curved strand in the top sheet and a very short strand in the bottom sheet.

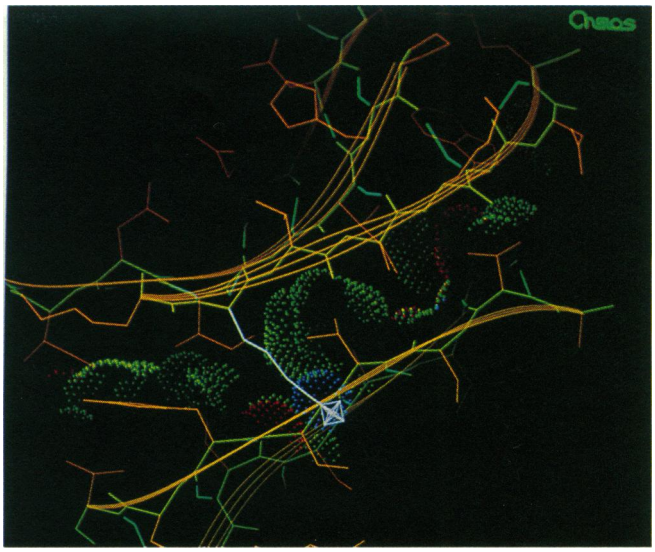


PLATE 24 The outer edges of the β sandwich in concanavalin A (PDB file 2CNA; Reeke et al., 1975), with a single inward-pointing lysine sidechain (marked in white) to help prevent edge-to-edge aggregation. Backbone is emphasized with a three-strand yellow ribbon, and the contacts of inward-pointing sidechains are shown with small-probe dots.

and Type II′ conformations, which fit the β sheet twist nicely (see Fig. 16) but include one or two positive φ angles which are unfavorable for non-glycine L-amino acids. For Betabellin 12, we took advantage of peptide synthesis to put a D-Pro,D-Asp sequence at each of the turns to favor a Type I′ conformation, and indeed this material was better than any previous version (McClain et al., 1992).

The most impressive improvement for Betabellin 12 was in its NMR spectrum, which looks rather native-like, with good spectral dispersion (Fig. 17) and large numbers of NOE (nuclear Overhauser) cross-peaks. NMR is a very valuable tool for assessing designed proteins, although it mainly tells you things you did not want to hear, such as that your material is not pure or that there is not a unique conformation. In this case, however, we were able to make a sequential assignment of almost all the proton resonances. Plate 22 shows a superposition of two-dimensional COSY (through-bond) and NOESY (through-space) spectra in the NH-CαH fingerprint region. The short range NOEs show that regions of sequence designed as β strands are consistent with extended strands, and there are definitely turns in the right places. However, there are somewhat more COSY peaks than there should be, which is indicative of more than one conformation. Even more disturbingly, it turned out that none of the NOEs are "long range" in the sequence, or represent tertiary-structure interactions rather than local conformation. One way of stating the problem is that there are three aromatic residues in the unique half-sequence, and none of those ring protons

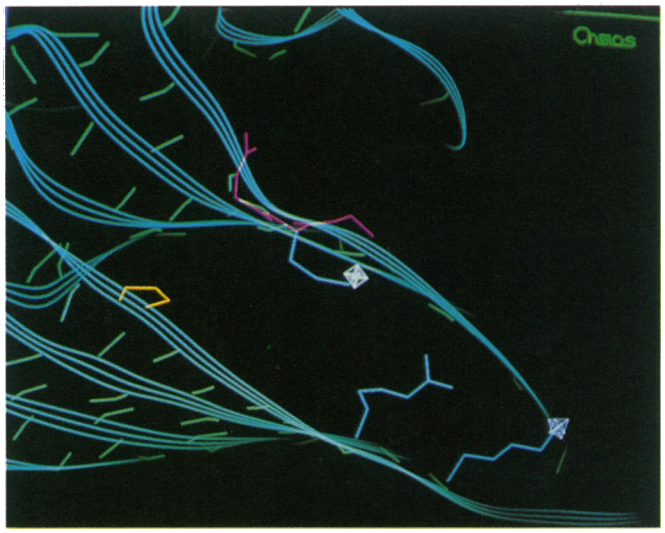


PLATE 25 The outer edge of the β sandwich in satellite tobacco necrosis virus (PDB file 2STV; Jones and Liljas, 1984), with a proline (*yellow*), a β bulge (*purple*), and three inward-pointing charged sidechains (2 Lys and an Arg; *blue*) to prevent edge-to-edge aggregation.

have any NOEs to another residue. At least one of the tyrosines should be on the inside, and if the Betabellin structure were unique and native-like it should have many NOEs to other parts of the protein. Betabellin forms good β structure as judged by CD and Raman evidence (McClain et al., 1992), and it must put the hydrophobic sides together most of the time to be this soluble. Presumably, what the NMR data tell us is that there are multiple ways of putting together those β strands and positioning the side chains, probably in very rapid equilibrium, so that too many conformations are being averaged to show any long range NOEs.

Betabellin, then, Felix, and all other designed proteins that have been characterized to this extent, seem to behave more like "molten globules" (Ohgushi and Wada, 1983), or partially folded proteins, than they do like natural proteins with a unique native structure. Kinemage 6 shows two different computer models that were built for Felix. They were built by different people and slightly different strategies, but both incorporate all the design criteria and have been carefully adjusted and minimized. They are very similar, but the animation between them shows that the helix angles change a bit and some side chains move quite significantly. If the real Felix protein can sample conformations this different, then it would indeed appear rather like a molten globule.

In summary, the field of protein design has recently come out to a rather unexpected general result: it is apparently well within the present state of knowledge to design and make entirely new proteins that fold up to approximately the intended three-dimensional structure, but no one yet knows enough to achieve a unique, native-like structure. Initial expectations were wrong about which was the hard step: it is near the end rather than near the beginning. If Fig. 18 represents an approximation of the folding process, then these designs have managed nucleation and initial structure formation but are currently stuck partway through the process; the ends are not tucked in and there are multiple conformations on the inside. We had initially expected that once one got this close, it would be easy to improve the structure by making individual "mutations" and accumulating the best ones. However, it now appears that there are probably problems in several parts of our designs, and the degree of disorder will not change appreciably if we manage to fix just one of them. The natural proteins apparently still know some tricks we don't know about successfully avoiding alternative structures and settling into a dynamic but unique conformation.

Protein design: future directions

Given the conclusions above, our own research program in protein design is now responding with developments in three different directions. First of all, we have not given up entirely on making individual changes aimed at improving the designed proteins. As an example, although Betabellin and Betadoublet are enormously more soluble than our early attempts, they still are not as soluble as natural β sheet proteins. Since we have already ensured the charge and hydrophilicity of the surface sidechains, we now believe that the problem is apt to be aggregation by H-bonding between the edge strands of different molecules. In other words, we made the structure too idealized and too neat. This is yet another case of negative design: an otherwise excellent and stable β sheet will not work if it aggregates with itself or other molecules, so it is necessary to "mess up" the edge strand in some way that will prevent those interactions.

We have surveyed the structures of natural β sheet proteins (especially the open edges of β sandwiches) to see how they deal with this edge problem, and always find at least one blocking feature on each exposed edge. The prealbumin in plate 23 illustrates both the kind of edge-to-edge dimerization we wish to avoid, and also the strong twist and very short edge strand that help prevent

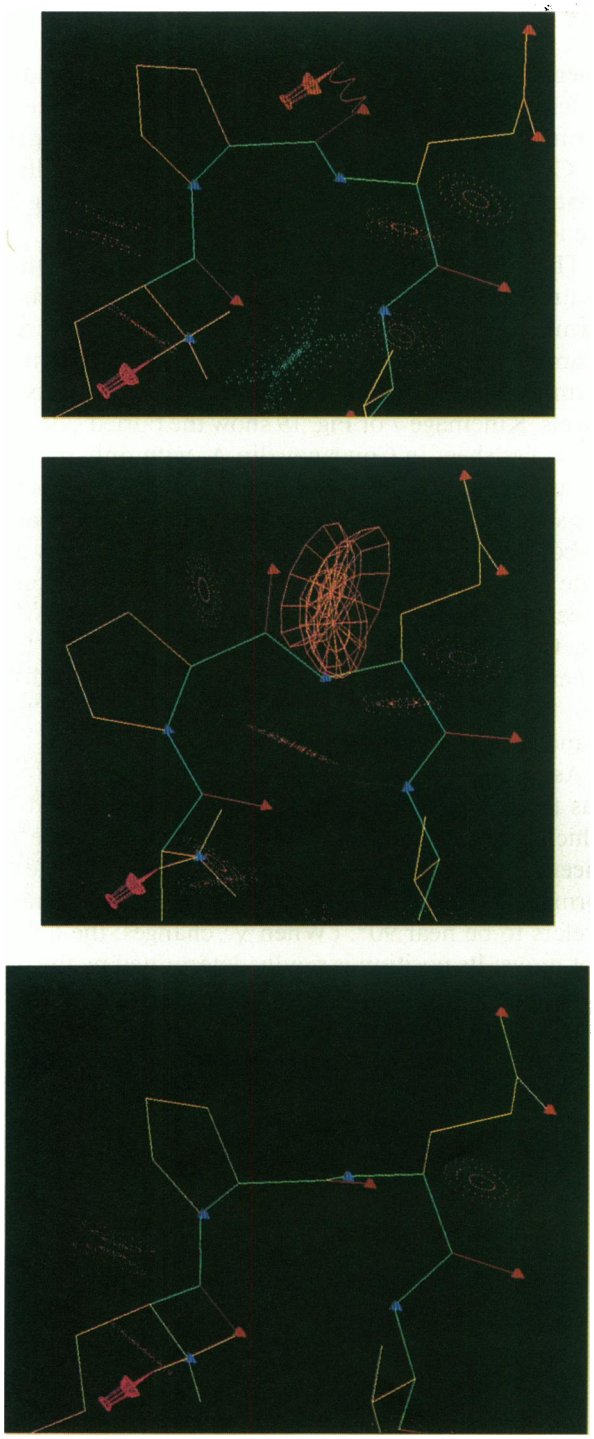


PLATE 26 A simple interactive operation using SCULPT: (*a*) (*top*) the starting model is a Type I tight turn; (*b*) (*middle*) the user grabs the carbonyl O and tugs on it to rotate the peptide, passing slowly through a region with unfavorable bumps (shown as *pink sectors*); (*c*) (*bottom*) it then settles into the Type II conformation.

further aggregation at the outer edge of the prealbumin β sheets. Concanavalin A also forms a β sheet dimer; the outer edges of its sheets are quite regular, as seen in plate 24, but there is an inward-pointing lysine on one of those strands that would mean burial of two like charges if it interacted with another Con A dimer. Plate 25 shows the edge of the satellite tobacco necrosis virus, which uses every trick in the book to block β H-bonding: an exposed proline ring to get in the way, a β-bulge to kink the strand, and three inward-pointing positive charges. Other proteins use a very pronounced or unusual twist, or a flap that covers the exposed strand edge. The most reliable of these strategies to use in a designed protein is probably the inward-pointing charged sidechain, which we now plan to incorporate into our β sheet designs. That will not help the uniqueness problem, but it may give us material better suited for biophysical measurements.

The second new direction involves capitalizing on the fact that design of approximate structure has been shown possible. For instance, if we can design a left-turning helix bundle then we can presumably design minimal changes that will convert it into a right-turning helix bundle. That would be very worthwhile because it would allow us to test alternative hypotheses about the factors most influential in determining that choice between to-

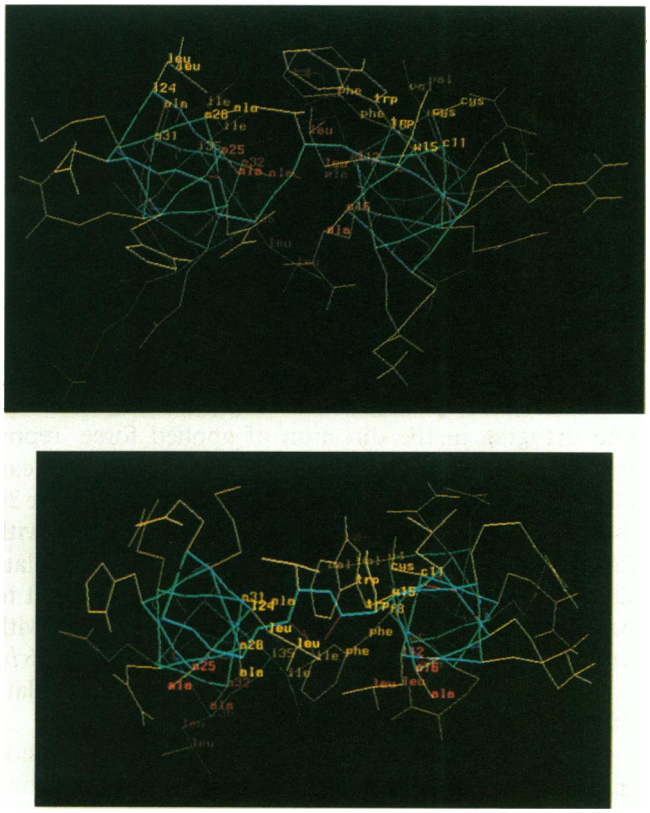


PLATE 27 SCULPT was used to turn the Felix model "inside out", to go from a left-turning bundle to a right-turning bundle. Plate *27*, *a* (*top*) and *b* (*bottom*), shows the beginning and ending states for one pair of helices, which rotate like meshed gears from making orange contacts to making yellow contacts.

FIGURE 19 Stick figure, in stereo, for part of the inner side of a large antiparallel β sheet from concanavalin A (PDB file 2CNA; Reeke et al., 1975), taken from Kinemage 7. Three of the Phe side chains have the "perpendicular" χ_1 conformation that puts them over the nearest-neighbor side-chain across a narrow pair of H-bonds.

pologies. This effort is closely tied in with a collaborative project to improve the tools for doing de novo modeling in general, by developing what we call a protein "sculpting" tool.

So far, the process of molecular modeling has involved two completely distinct aspects: (*a*) purely geometrical, interactive modeling in real time that allows one to twist bonds but also to move one atom on top of another; and (*b*) batch calculations that incorporate the physical forces but cannot be steered. The SCULPT program lets the user apply tugs to the model while a real-time force calculation follows along, letting atoms either move out of the way if bumped or move closer together if they interact favorably. Tugs are added by clicking on atoms and dragging in the direction of applied force, represented on the graphics screen by a coiled spring. One can also tie down parts of the molecule with "nails". Plate 26 is a simple sequence illustrating an operation with SCULPT: the starting model is a Type I tight turn (plate 26 *a*); the user grabs the carbonyl O and tugs on it to rotate the peptide, passing slowly through a region with unfavorable bumps (shown as pink sectors in plate 26 *b*) and then settling into the Type II conformation (plate 26 *c*).

SCULPT was used to turn the Felix model "inside out", to go from a left-turning bundle to a right-turning bundle. Plate *27, a* and *b,* shows the beginning and ending states for one of the helix–helix contacts, which rotates in the gear-wheel fashion of Fig. 11 from orange contacts to yellow contacts. The connecting loop unwinds from one helix and winds up onto the other by about one residue. The loop conformation and some sidechains were improved by some hand tugging and the model extensively minimized. We are now using this pair of models to choose appropriate (but minimal) sequence changes for the right-turning bundle, including a different disulfide for that form. We plan to make four forms (each sequence with the right and the wrong pairs of Cys) and use this system to test systematically the relative influence of various factors in determining bundle topology.

The third direction for future work is a serious attempt to understand internal packing interactions and the determinants of unique versus multiple structures. As an example, we will look at what determines side chain conformations, especially for aromatics, on the surface of β sheets. Kinemage 7 or Fig. 19 show the buried side of the large, flat sheet in Concanavalin A, with only the backbone, the H-bonds, and sidechains on the interior side of the sheet included. Since this is an antiparallel sheet, the H-bonds are perpendicular to the strands, and they alternate a narrow pair with a more widely spaced pair between any two neighboring strands, as we saw above for the two-stranded β ribbons in BPTI or STNV. Every other sidechain along a strand points inward, and their $C\alpha - C\beta$ vectors are approximately perpendicular to the plane of the sheet.

As suggested schematically in plate 28, each side chain has three choices of χ^1 angle. For Phe, Tyr, and Val, which are common and influental on the inside of β sheets, that threefold choice describes their possible conformations very well to the first approximation, since χ^2 prefers to be near 90°. (When χ^1 changes, the $C\beta$ does not move. Its position is rigidly determined by the backbone, and any bumps it makes must be relieved by adjustments of the backbone conformation. If it were not for glycine, we could more reasonably consider $C\beta$ part of the main chain rather than part of the side chain.) For Trp, Leu, and Ile, χ^2 is also a major factor, while in all cases there can be departures from the canonical values where necessary. However, high resolution structures have shown that these departures are much smaller and

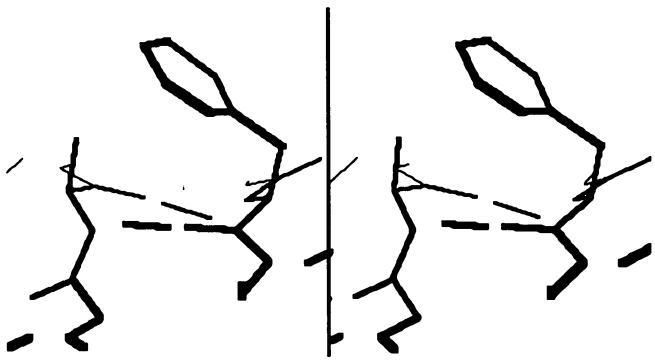


FIGURE 20 Closeup, in side view, of the Phe 195-Ala 50 pair from the Con A β sheet. There is extensive contact between the Ala methyl and the Phe ring, and the Phe $C\alpha$-$C\beta$ bond leans backward slightly.

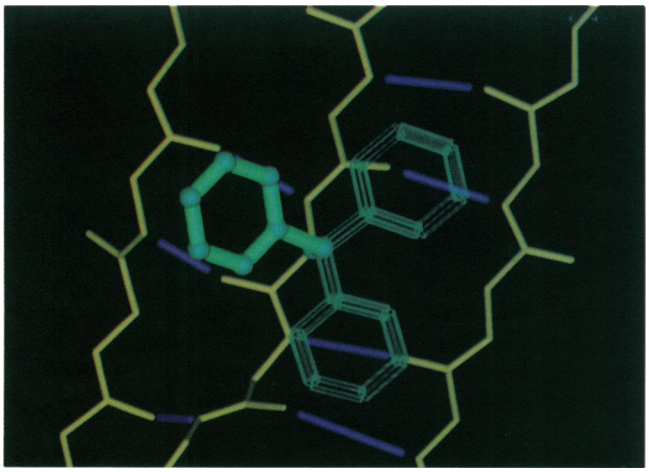


PLATE 28 The three major sidechain conformational choices, illustrated for one phenylalanine on a β sheet from concanavalin A. The native conformation ($\chi 1$ near +60°) is shown in solid green and the other two alternatives as open outlines.

less common than one would have supposed (Ponder and Richards, 1987; Richardson and Richardson, 1989*b*). Therefore, to consider packing and uniqueness for a β sheet interior, we will begin by analyzing the threefold choice of each side chain and the patterns of how those choices influence one another.

The three staggered χ^1 conformations are: (*a*) −60°, with Cγ opposite the mainchain CO; (*b*) 180°, with Cγ opposite the mainchain N; (*c*) +60°, with Cγ opposite the CαH, perpendicular to the backbone. (Since we will concentrate on aromatics here, we do not need to worry about the choice of reference atom for Cβ-branched residues.) For sidechains in general, the "a" χ^1 conformation is preferred, "b" a bit less so, and "c" is relatively rare (Janin and Wodak, 1978). For aromatic sidechains on the buried sides of β sheets, however, the perpendicular "c" conformation is quite common (Richardson and Richardson, 1989*b*).

The Con A β sheet of Kinemage 7 has 3 buried aromatics in the perpendicular χ^1 conformation, each interacting with a different amino acid as partner across the narrow H-bond pair: Phe 197 with Gly 48, Phe 195 with Ala 50, and Phe 128 with Phe 111. In that χ^1 orientation, the aromatic ring lies almost directly on top of the Cα and Cβ of the residue on the neighboring strand, across a narrow pair of H-bonds (see plate 28). Looking at Phe 195 along the plane of the β sheet (see Fig. 20 or Kinemage 7), we can see that the Cα − Cβ vector leans back somewhat from vertical, pushed away by the methyl group of Ala 50. For Phe 197, whose partner is a Gly, the Cα − Cβ vector leans forward somewhat in order for the Phe ring to touch the backbone of the Gly.

These van der Waals contacts can be shown explicitly with small probe dots (Kinemage 7). If the dot contacts are calculated with implicit H (the usual method, which increases each heavy atom radius to produce the correct total volume for the atom plus its hydrogens), they show a large contact on each face of the ring but almost nothing around the edges; the Phe-Ala contact overlaps a little and the Phe-Gly looks sparse. This is because the aromatic ring H's and the Gly CαH have asymmetrical positions around their heavy atoms and are not at all well represented by radially symmetric, implicit H. Dot contacts calculated with explicit H atoms (Kinemage 7) show excellent, detailed fit all around the ring, with the large flat surface of the Ala methyl or the Gly CαH fitting right up against the side of the ring and other groups touching the H atoms around the edge. This emphasizes again the necessity of including the H atoms in any serious analysis either of internal packing or of binding sites.

One reason, then, for the $\chi^1 = +60°$ conformation for aromatics on buried sides of β sheets is that with a small adjustment of the backbone one way or the other they can make a good contact with either a Cα group of Gly or a Cβ group on some other side chain. In fact, for a small residue like Gly or Ala, the only surrounding position on the same sheet that can occupy the space above it is a large sidechain with $\chi^1 = +60°$ across the narrow pair of H-bonds. Statistically, on buried sides of β sheets, an aromatic which is across a narrow H-bond pair from a Gly or Ala will almost invariably adopt the $\chi^1 = +60°$ conformation.

Let's look now at Phe 128 across from Phe 111 to see what happens if the partner side chain is longer than Ala (Kinemage 7). Phe 111 has $\chi^1 = 180°$, pointing away from 128, and the contact of the Phe 128 ring with the Phe 111 CβHs looks almost identical to what we saw for Ala. However, when a buried aromatic is across from a side chain longer than Ala, it only adopts the $\chi^1 = +60°$ conformation half to one-third of the time. The reason for this difference is obvious, of course: the partner side chain also has a choice of conformation, and if it lies toward the aromatic, the aromatic must move out of the way into one of the other χ^1 positions. This emphasizes that the entire set of sheet residues are making conformational choices and those choices all interact with their neighbors. This system is rather like a spin glass in physics, where each lattice point has a choice of up or down spins, but the spins all interact with their neighbors.

We need to find out, then, not only what is the best set of conformations for any given sequence on a sheet, but we also need to learn the rules for which arrangements of side chains on a sheet produce a unique best conformation and which produce multiple equivalent choices. This understanding is presumably necessary before we can design new proteins with unique, truly native-like conformations, and it also will tell us a lot about how natural proteins design their structures, either during evolution or during folding.

Finally, our ideas about looking at proteins can be summarized by one trivial statement and a few outrageous ones:

(*a*) Viewpoint and selectivity are both crucial: to make sense of complex things, you need to look at them in the right way, and add or delete parts.

(*b*) What makes proteins interesting as structures is that they are inherently both one-dimensional and three-dimensional at the same time.

(*c*) β-carbons are part of the backbone.

(*d*) Contacts are all hydrogens.

(*e*) Negative design, to actively block alternatives, is just as important to protein structure as positive design.

(*f*) The hardest part of protein folding, or protein design, is the last little bit.

We would like to thank the Biophysical Society officers and meeting organizers, the Keck Center for Computational Biology, AV Associates, and Evans and Sutherland, whose extensive help made possible the big-screen interactive graphics used for the Biophysical Society 1992 National Lecture, on which this article is based.

Graphics credits are as follows: Plates 20, 21, and 28: VIEW, by Larry Bergman, on a Silicon Graphics Iris 340GTX. Plates 26 and 27: SCULPT, by MCS, on a Silicon Graphics Iris 340GTX (SCULPT and VIEW are projects of the GRIP molecular graphics group in the Computer Science Department at UNC Chapel Hill, led by Fred Brooks). Plates 7, 15, 17–19, 24–25, and Fig. 1: CHAOS, by DCR, on an Evans & Sutherland ESV (Salt Lake City, UT). Plate 6, Figs. 19, 20, and all Kinemages: PREKIN and MAGE, by DCR, on a Mac IIci. Plate 17 and Fig. 22: NMR data plotted with FELIX, by Dennis Hare. Plates 9, 13, 23, and Figs. 3–16 and 18: hand-drawn by JSR. Plates 4 and 16: drawn by JSR and air-brushed by Mike Zalis. Plate 14: stained glass by Karen Williams, Alpine, TX. The schematic drawings are copyrighted by J. S. Richardson.

The research summarized here was supported principally by NIH GM-15000, and also in part by ONR, NASA, the Life Sciences Foundation, the MacArthur Foundation, Merck, Glaxo, and other NIH grants.

Received for publication and in final form 3 August 1992.

REFERENCES

Abad-Zapatero, C., J. P. Griffith, J. L. Sussman, and M. G. Rossmann. 1987. Refined crystal structure of dogfish M4 apo-lactate dehydrogenase. *J. Mol. Biol.* 198:445–467.

Arnone, A., C. J. Bier, F. A. Cotton, V. W. Day, E. E. Hazen, Jr., D. C. Richardson, J. S. Richardson, and A. Yonath. 1971. A high resolution structure of an inhibitor complex of the extracellular nuclease of Staphylococcus aureus: experimental procedures and chain tracing. *J. Biol. Chem.* 246:2303–2316.

Banner, D. W., A. C. Bloomer, G. A. Petsko, D. C. Phillips, C. I. Pogson, I. A. Wilson, P. H. Corran, A. J. Furth, J. D. Milman, R. E. Offord, J. D. Priddle, and S. G. Waley. 1975. Structure of chicken muscle triose phosphate isomerase determined crystallographically at 2.5 Ångstroms resolution using amino acid sequence data. *Nature (Lond.)*. 255:609–614.

Bernstein, F. C., T. F. Koetzle, G. J. B. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. 1977. The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* 112:535–542.

Blake, C., M. Geisow, S. Oatley, B. Rerat, and C. Rerat. 1978. Structure of prealbumin: secondary, tertiary, and quaternary interactions determined by Fourier refinement at 1.8Å. *J. Mol. Biol.* 121:339–356.

Carson, M., and C. E. Bugg. 1986. Algorithm for ribbon models of proteins. *J. Mol. Graphics* 4:121–122, 207.

Chothia, C. 1973. Conformation of twisted β-pleated sheets in proteins. *J. Mol. Biol.* 75:295–302.

Connolly, M. L. 1983. Solvent-accessible surfaces of proteins and nucleic acids. *Science (Wash.)*. 221:709–713.

DeGrado, W. F., Z. R. Wasserman, and J. D. Lear. 1989. Protein design: a minimalist approach. *Science (Wash. DC)*. 243:622–628.

Hecht, M. H., J. S. Richardson, D. C. Richardson, and R. C. Ogden. 1990. De novo design, expression, and characterization of Felix: a four-helix bundle protein of native-like sequence. *Science (Wash. DC)*. 249:884–891.

Jacobson, R. H., M. Matsumura, H. R. Faber, and B. W. Matthews. 1992. Structure of a stabilizing disulfide bridge mutant that closes the active-site cleft of T4 lysozyme. *Protein Sci.* 1:46–57.

James, M., A. Sielecki, G. Brayer, L. Delbaere, and C.-A. Bauer. 1980. Structure of product and inhibitor complexes of Strep. griseus protease A at 1.8 Å resolution: a model for serine protease catalysis. *J. Mol. Biol.* 144:43–88.

Janin, J., and S. Wodak. 1978. Conformation of amino acid side-chains in proteins. *J. Mol. Biol.* 125:357–386.

Jones, T. A., and L. L. Liljas. 1984. Structure of satellite tobacco necrosis virus after refinement at 2.5 Ångstroms resolution. *J. Mol. Biol.* 177:735–767.

Lederer, F., A. Glatigny, P. H. Bethge, H. D. Bellamy, and F. S. Mathews. 1981. Improvement of the 2.5 Ångstrom resolution model of cytochrome b562 by redetermining the primary structure and using molecular graphics. *J. Mol. Biol.* 148:427–448.

McClain, R. D., Y. Yan, R. W. Williams, M. E. Donlan, and B. W. Erickson. 1992. Protein engineering of betabellin 12. *In* Peptides: Chemistry and Biology (Proceedings of the 12th American Peptide Symposium). J. A. Smith and J. E. Rivier, editors. ESCOM, Leiden. 364–365.

Ohgushi, M., and A. Wada. 1983. 'Molten globule state': a compact form of globular proteins with mobile side-chains. *FEBS Lett.* 164:21.

Ponder, J. W., and F. M. Richards. 1987. Tertiary templates for proteins: use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.* 193:775–791.

Quinn, T. P., N. B. Tweedy, J. S. Richardson, and D. C. Richardson. 1991. A de novo designed β sheet protein. Abstract for Protein Society meeting, Baltimore.

Rao, S. T., and M. G. Rossmann. 1973. Comparison of super-secondary structure in proteins. *J. Mol. Biol.* 76:241–256.

Reeke, G. N., Jr., J. W. Becker, and G. M. Edelman. 1975. The covalent and three-dimensional structure of concanavalin A: IV. Atomic coordinates, hydrogen bonding, and quaternary structure. *J. Biol. Chem.* 250:1525–1547.

Richardson, J. S. 1976. Handedness of crossover connections in β sheets. *Proc. Natl. Acad. Sci. USA.* 73:2619–2623.

Richardson, J. S. 1977. β sheet topology and the relatedness of proteins. *Nature (Lond.)*. 268:495–500.

Richardson, J. S. 1981. The anatomy and taxonomy of protein structures. *Adv. Protein Chem.* 34:167–339.

Richardson, J. S. 1985. Schematic drawings of protein structures. *In* Diffraction Methods for Biological Macromolecules. H. W. Wyckoff, C. H. W. Hirs, and S. N. Timasheff, editors. *Methods Enzymol.* 115b:359–380.

Richardson, J. S., and D. C. Richardson. 1987. Some design principles: Betabellin. *In* Protein Engineering. D. Oxender and C. F. Fox, editors. Alan R. Liss, New York. 149–163.

Richardson, J. S., and D. C. Richardson. 1988. Amino-acid preferences for specific locations at the ends of α-helices. *Science (Wash. DC)*. 240:1648–1652.

Richardson, J. S., and D. C. Richardson. 1989a. The de novo design of protein structures. *Trends Biochem. Sci.* 14:304–309.

Richardson, J. S., and D. C. Richardson. 1989b. Principles and patterns of protein conformation. *In* Prediction of Protein Structure and Principles of Protein Conformation. G. Fasman, editor. Plenum Press, New York. 1–98.

Richardson, D. C., and J. S. Richardson. 1992. The Kinemage: a tool for scientific communication. *Protein Sci.* 1:3–9.

Rubin, B. H., and J. S. Richardson. 1972. The simple construction of protein α-carbon models. *Biopolymers.* 11:2381–2385.

Salemme, F. R. 1983. Structural properties of protein β-sheets. *Prog. Biophys. Mol. Biol.* 42:95–133.

Shoemaker, K. R., P. S. Kim, E. J. York, J. M. Stewart, and R. L. Baldwin. 1987. Tests of the helix dipole model for stabilization of α-helices. *Nature (Lond.).* 326:563–567.

Sibanda, B. L., and J. M. Thornton. 1985. β hairpin families in globular proteins. *Nature (Lond.).* 316:170–174.

Surles, M. C. 1992. An algorithm with linear complexity for interactive, physically-based modeling of large proteins. SIGGRAPH '92 "Computer Graphics". 26:221–230.

Tainer, J. A., E. D. Getzoff, K. M. Beem, J. S. Richardson, and D. C. Richardson. 1982. Determination and analysis of the 2 Ångstrom structure of copper, zinc superoxide dismutase. *J. Mol. Biol.* 160:181–217.

Unson, C. G., B. W. Erickson, D. C. Richardson, and J. S. Richardson. 1984. Protein engineering: design and synthesis of a protein. *Fed. Proc.* 43:1837.

Wistow, G., B. Turnell, L. Summers, C. Slingsby, D. Moss, L. Miller, P. Lindley, and T. Blundell. 1983. X-ray analysis of the eye lens protein γ-II crystallin at 1.9 Å resolution. *J. Mol. Biol.* 170:175–202.

Wlodawer, A., and L. Sjolin. 1983. Structure of ribonuclease A. Results of joint neutron and x-ray refinement at 2.0 Å resolution. *Biochemistry.* 22:2720–2728.

Wlodawer, A., J. Walter, R. Huber, and L. Sjolin. 1984. Structure of bovine pancreatic trypsin inhibitor: Results of joint neutron and x-ray refinement of crystal form II. *J. Mol. Biol.* 180:301–329.