DE NOVO DESIGN AND STRUCTURAL CHARACTERIZATION OF PROTEINS AND METALLOPROTEINS

William F. DeGrado and Christopher M. Summa
The Johnson Research Foundation and Department of Biochemistry and Biophysics, University of Pennsylvania Medical School, Philadelphia, Pennsylvania 19104-6059; e-mail: wdegrado@mail.med.upenn.edu; csumma@mail.med.upenn.edu

Vincenzo Pavone, Flavia Nastri, and Angela Lombardi
Centro Interuniversitario di Ricerca sui Peptidi Bioattivi and Centro di Studio di Biocrystallografia del C.N.R., University of Naples “Federico II”, Via Mezzocannone 4 I-80134 Naples, Italy; e-mail: pavone@chemna.dichi.unina.it; nastri@chemna.dichi.unina.it; and lombardi@chemna.dichi.unina.it

Key Words peptide, metal ligand, folding, computation, protein structure

Abstract De novo protein design has recently emerged as an attractive approach for studying the structure and function of proteins. This approach critically tests our understanding of the principles of protein folding; only in de novo design must one truly confront the issue of how to specify a protein’s fold and function. If we truly understand proteins, it should be possible to design receptors, enzymes, and ion channels from scratch. Further, as this understanding evolves and is further refined, it should be possible to design proteins and biomimetic polymers with properties unprecedented in nature.

CONTENTS

Introduction ................................................................. 780
Rules and Tools for De Novo Protein Design ......................... 781
Energetic Considerations .................................................. 781
Choosing a Target Fold .................................................... 782
The Combinatorial Problem of Inverse Protein Design ............... 783
Side-Chain–Repacking Algorithms ...................................... 784
Design of the Surface and Interfacial Residues ....................... 786
Towards Automated Design of Functional Proteins: Engineering
Metal Ion-Binding Sites .................................................... 786
INTRODUCTION

De novo protein design has been the topic of several recent reviews (1–5), and this term has been applied to a variety of activities. In its purest and most challenging form, de novo design involves the construction of a protein, intended to fold into a precisely defined 3-dimensional structure, with a sequence that is not directly related to that of any natural protein. The term “de novo design” also has frequently been applied to the design of sequences intended to adopt a given general fold, without any preconceived notions of the precise atomic-level details of the targets (6–9). It has also been applied to a variety of strategies involving the redesign of naturally occurring proteins. For example, a sequence-based strategy for protein design involves construction of “minimalist” sequences (2, 10, 11) that are simpler than their natural counterparts but, nevertheless, retain sufficient complexity for folding and function. Alternatively, automated, structure-based strategies, which begin with an experimentally determined 3-dimensional structure of a natural protein (12, 13), have been extremely useful for the design of proteins with enhanced stability or novel functions. In this review, we summarize principles and methods for designing proteins and highlight recent examples of structurally defined and functional proteins.
RULES AND TOOLS FOR DE NOVO PROTEIN DESIGN

Energetic Considerations

A designed sequence will fold into a unique, well-defined structure only if it satisfies two requirements. First, it must contain elements of positive design to thermodynamically stabilize the desired fold; the folded conformations of natural proteins are generally stabilized by \( \approx 4-10 \text{ kcal/mol} \) (a factor of 1000 to 10 million), relative to their unfolded states. Second, a sequence should contain elements of negative design to create a large energy gap between the native fold and any other folded conformation; otherwise the protein would adopt a molten globule-like ensemble of folds rather than a unique, native structure (14). Thus, successful design is particularly challenging in that it requires not only stabilization of the desired structure, but also destabilization of all possible alternatively folded structures that serve as kinetic and thermodynamic traps. An analogous situation would be to provide a person with directions for driving from New York to San Francisco. Although it is quite simple to specify the desired route, it is much more difficult to anticipate the multitude of intersections along the way and specify that the driver should not follow each of these alternative pathways. Indeed, early experimental studies (15, 16) and lattice simulations (17, 18) showed that negative design is a critical process, which requires one to anticipate and destabilize as many possible alternative low-energy structures as possible (19).

A large number of interactions contribute to the stability of a protein. Site-directed mutagenesis (20, 21) and studies with designed peptides (2) indicate that buried residues in the interior of proteins play particularly important energetic roles. Relative to Ala, a buried Leu or Phe residue contributes \( \approx 2-5 \text{ kcal/mol} \) to stability, and buried H-bonds or salt bridges further increase stability by as much as \( 3 \text{ kcal/mol} \), expressed on a per-residue basis. Short-range (in sequence) H-bonds between small polar side chains and backbone amides can be stabilizing by 1–2 kcal/mol, as in helix N caps (22). Exposed salt bridges are less stabilizing, contributing \( \approx 0.5 \text{ kcal/mol} \) under physiological conditions (23–26). Finally, each residue in a protein contributes its intrinsic conformational preference for adopting a given set of \( \phi, \psi \) torsional angles (27–32). With the exception of Pro, the energetic contribution of this effect spans a range of \( \approx 1 \text{ kcal/mol} \) for the commonly occurring amino acids. For folding to proceed, the sum of these favorable interactions must exceed the unfavorable loss in configurational entropy \( \approx 1-2 \text{ kcal/mol} \) destabilizing at room temperature (33)). Thus, the information required for folding is distributed throughout the chain in a network of diverse interactions, each with different energetic contributions. The most sophisticated approaches to protein design (34) include and optimize all types of interactions, and no single residue or interaction type can be said to determine the fold.

It is difficult to determine the origin of the large energy gap necessary for conformational specificity through the study of natural proteins alone. A lack of conformational specificity can have severe pathophysiological consequences (35),
hence natural proteins have evolved multiple, redundant mechanisms for defining this energetic gap. For example, the apolar cores of several natural proteins have been replaced with a single type or a random collection of apolar residues, with retention of activity (36–38). Evidently, residual interactions involving the nonmutated residues compensate for losses in packing of the core.

Studies with designed proteins and lattice models (18) have, however, highlighted the importance of negative design in providing conformational specificity. The need to strike the appropriate balance between apolar and polar interactions (39) represents a very important theme, which is reiterated with a very high frequency throughout the examples cited in the sections below. Whereas hydrophobic (Hb) interactions provide a powerful driving force for formation of the desired structure, polar interactions (40, 41) provide conformational specificity, often at the expense of thermodynamic stability (42). Examples of the importance of the use of polar residues in negative design include (a) the placement of hydrophilic (Hp) groups that are solvent exposed in the desired fold but buried in alternatively folded structures (43, 44); (b) inclusion of buried H-bonds and salt bridges that are less stabilizing than Hb interactions, but can be formed only in the desired structure (45–51a); (c) solvent-accessible electrostatic interactions that are very weakly stabilizing of the desired fold (23, 24, 26, 52–55) but strongly destabilize alternatives. Other elements of negative design include conformational restraints from residues such as Gly and Pro that are often used to interrupt elements of secondary structure. Shape complementarity and, more important, a lack thereof provide a third medium for negative design. For example, antiparallel 4-helix helical bundles have been engineered to have very poorly packed interiors in the undesired, parallel folding pattern, while maintaining excellent packing in the desired fold (19, 56).

Choosing a Target Fold

The first feature to consider when choosing a target 3-dimensional structure is the designability of the fold. In lattice simulations, many different sequences fold into the same low-energy structure, whereas other possible folds are rarely if ever found to represent the lowest-energy structures for any sequence (18, 57, 58). Similarly, some folds such as the triosephosphate isomerase (TIM) barrel (59, 60), 3-helix bundles (61), 4-helix bundles (7), and immunoglobulin (Ig) motif (62) recur frequently in proteins with highly divergent sequences. Such folds are not only highly designable, they also may be easily modified to introduce novel functions without greatly perturbing their 3-dimensional structures. If one wishes to take a structure-based approach to design, it is next necessary to create a 3-dimensional model of the desired backbone structure. If the aim is to redesign a natural protein, the coordinates of its crystallographic or NMR structure are used. However, in true de novo design, it is instead necessary to create a set of coordinates for a novel structure. This may be accomplished by using a rigorous mathematical parameterization of, for example, 222 symmetric 4-helix bundles (15, 63), β-sheets (64, 65), TIM barrels (66, 67), or coiled coils (68, 69). Alternatively, a backbone structure may be assembled from libraries of locally folded structures (70, 71).
The Combinatorial Problem of Inverse Protein Design

Once a backbone conformation has been selected, the next step involves “inverse design” (72, 73)—the specification of a sequence that will fold into the target structure. Because they form a very powerful driving force for folding (74, 75), the placement of apolar residues at buried positions in the predicted structure will provide a strong disposition for adopting the desired secondary and super-secondary structure (see Figure 1; 76–80). Thus, the pattern of Hb and Hp residues in a sequence is often defined at an early stage in protein design. Helices are stabilized with an Hb period of ~3.6 residues, whereas sheets require a 2-residue repeat. This Hb/Hp pattern limits the number of possible sequences significantly. For example, a 100-residue protein with 20 possible residues at each position has $10^{130}$ possible sequences; restricting the sequence to a given Hb/Hp pattern with only 5 different residues allowed at each position of the sequence reduces this number by a factor of $10^{60}$. Nevertheless, the number of combinations ($10^{70}$) remains astronomical.

How does one deal with this explosion of sequences? First, phylogenetic studies suggest that a very large number of sequences fold into the same 3-dimensional structure (81). Therefore, it is necessary to design only one of a very large number of possible sequences capable of assuming the target fold (18). Second, one may use smaller sets of amino acids to provide a more manageable number of possible sequences (2, 10, 11). Lattice simulations have suggested that an Hb/Hp binary code is necessary and sufficient for folding (82 and references therein), although more recent simulations suggest that a diversity of amino acid side chains may be necessary to provide a unique fold (18). Indeed, a recent experimental study suggests that a minimum of two Hp and two Hb amino acids plus Gly may be required for nativelike, structured proteins (11).

Alternatively, one may use genetic methods to screen through vast numbers of proteins to find sequences that fold into a given 3-dimensional structure. Bowie et al (74) and Reidhaar-Olson et al (83) introduced cassette mutagenesis with random oligonucleotides as a method to generate a broad spectrum of sequences within a focused region of a protein, and Kamtekar and coworkers (6) have extended this approach to the construction of large libraries of sequences displaying defined Hb/Hp patterns. This was accomplished by using partially random oligonucleotides containing the DNA sequence NAN (where N is any base) to code for Phe, Ile, Leu, Met, or Val, and NTN to encode Hp residues, Glu, Gln, Asp, Asn, Lys, or His. Members of this or a related library were then individually screened for proteins with nativelike (84) or heme-binding properties (85). Random, combinatorial approaches are most powerful when combined with a selection for an appropriate function. This is often facilitated by displaying libraries of up to $10^{11}$ distinct sequences on the surface of a filamentous phage (86–88). Subsequent screening for binding to an immobilized receptor allows one to rapidly identify the highest-affinity binders in the entire population.

Before leaving the subject of core repacking, it is important to mention that buried H-bonded and electrostatic interactions are often very important (41, 45).
Figure 1  The positions of the hydrophobic residues help to determine the fold of an idealized \( \beta - \beta - \alpha \) motif (a) and a helix-turn-helix (b) motif. Hydrophobic residues (green) stabilize the supersecondary structure (78, 161).

for conformational specificity. Surprisingly, although there have been several excellent studies in which buried polar interactions have been removed from natural proteins (reviewed in 42, 89), there have been relatively few examples of designing them into novel structures (49). Side-chain-repacking algorithms may also become useful for designing H-bonded networks of polar, buried side-chains.

**Side-Chain–Repacking Algorithms**

The side chains in the solvent-inaccessible cores of proteins are generally tightly packed, providing a large van der Waals component to the thermodynamic stability of the fold. There are two limiting viewpoints concerning the process by which side chains may achieve their tightly interdigitated conformations and the concomitant implications for protein design (reviewed in 1, 82). At one extreme, the oil drop model posits that the interior side chains may have a very high degree of flexibility in finding a closely packed structure, allowing virtually any apolar sequence to pack into a native structure. At the other extreme, the jigsaw puzzle model hypothesizes that the core side chains are locked into place in precisely defined conformations and require very specific shape complementarity among the interacting residues. In fact, these are only limiting cases, and the number of sequences that can adopt a native fold is more limited than one might expect from an oil drop model, but less limiting than expected from the jigsaw puzzle analogy.
Combinatorial computational algorithms (12) provide a powerful complement to genetic methods for exploring sequence space. Here one begins with a given backbone structure and explores large numbers of side chains that can fit together to stabilize the fold (90). The combinatorial problem is even greater for computational algorithms, because each residue can exist in any of a number of distinct low-energy rotamers that differ in their side chain torsional angles (91–93). Most frequently, computational algorithms are used to select a sequence of only the buried, Hb residues, which contribute most to the energetics of protein folding. Nevertheless, the number of combinations can be unmanageable unless one uses an appropriate technique, such as a genetic algorithm (GA). The GA is a stochastic search algorithm—it does not systematically try every solution but, instead, evolves solutions by “mating” and “mutating” the sequence/structure information of successive generations of solutions (94). First applied to protein design by Jones (95), a GA has more recently been written that explicitly considers side-chain packing constraints (96, 97).

Although GAs randomly evolve toward a solution, the dead-end elimination theorem (DEE; 98, 99), a special case of the general search algorithm A* (100), facilitates the search for the best solution by systematically eliminating high-energy rotamers that may be rigorously excluded from the global minimum energy solution of the system. A rotamer \( i_r \) at a given position may be excluded if, when compared with another rotamer \( i_t \), the following inequality holds true:

\[
E(i_r) + \sum_j \min_s E(i_r j_s) > E(i_t) + \sum_j \max_s E(i_t j_s); \quad i \neq j, 1.
\]

in which \( E(i_r) \) and \( E(i_t) \) are rotamer/backbone energies, \( E(i_r j_s) \) and \( E(i_t j_s) \) are rotamer/rotamer pairwise energies between a rotamer at positions \( i \) and \( j \), and \( \min_s \) and \( \max_s \) are, respectively, functions that minimize and maximize the argument by cycling over all \( s \). Although the technique was initially developed to search side-chain conformers for homology modeling, recent modifications allow the identities of the amino acid to be varied (101). This method can reduce the search time by \( \approx 72 \) orders of magnitude, allowing one to find the globally optimum solution for small proteins (98).

An essential component of any computational search strategy is a potential energy function, which can range from simple Lennard-Jones potentials (102) to more sophisticated molecular-mechanics force fields (103–105). The force field of Dahiyat et al (34, 101, 106) also includes terms based on the degree of burial of polar and apolar groups and a residue’s secondary structure propensity (107). Theoretical papers have also appeared concerning protein design using amino acid pairwise potentials (95, 108), based on the statistical frequency with which pairs of amino acids interact in the known 3-dimensional structures of proteins (109–111). It should also be possible to use analogous atomic statistical potentials (112–114) in design.

A major problem surrounding the use of most core repacking algorithms is that they maintain a rigid backbone structure throughout the calculation (96, 97), whereas small adjustments to the main chain atoms might be necessary to
find the most stable sequence. One strategy for circumventing this problem involves scaling back the van der Waals radii of the side-chain atoms (115). Alternatively, one may repeat the repacking algorithm with an ensemble of related backbone conformations rather than a single structure. Such ensembles may be generated by restrained molecular dynamics calculations (116) or by parameterizing the backbone trace of a structure to allow systematic displacement of the helices or sheets (69, 117).

**Design of the Surface and Interfacial Residues**

One quarter or less of the side chains in proteins occur in their solvent-inaccessible cores; the remaining residues help stabilize the desired fold and destabilize alternatives (23, 24, 26, 44, 52–55). Automated methods have been developed for the design of the interfacial and fully solvent-accessible portions of a protein chain. The program AGADIR (118–121) predicts the stability of solvent-exposed α-helices, taking into consideration their intrinsic conformational preferences as well as interactions with neighboring groups within the helix. In a more general approach, Dahiyat & Mayo have extended their genetic algorithm to allow the design of the exteriors as well as the interiors of proteins (107). Clearly, this is a very important area for design that deserves future attention.

**Towards Automated Design of Functional Proteins:**

**Engineering Metal Ion-Binding Sites**

Metal-binding sites are essential components of many natural proteins (122). New metal sites can be introduced into designed proteins to add structural specificity (14, 123), to aid in the binding of organometallic cofactors (124), or to modulate the oligomeric state of the protein (125, 126). Metal sites can also be engineered into naturally occurring folds, often by using a computer search algorithm (90, 127). New sites forming a given ligation geometry are found by searching existing structures for backbone geometries consistent with the conformational needs of the ligating groups. Recent studies suggest that functional metal sites are accessible through design strategies (128–130), moving the field closer to the goal of novel enzyme design.

**DESIGN OF SECONDARY AND SUPER–SECONDARY STRUCTURES**

With very few exceptions (131, 132), peptides spanning the helices, turns, and β-hairpins of natural proteins show little secondary structure formation under physiological conditions. However, it has been possible to design model peptides
that fold into helices and β-hairpins, providing convenient minimal systems for studying the thermodynamics and kinetics of secondary structure formation.

α-Helices

The study of synthetic, monomeric helices has contributed significantly to our understanding of helix formation in proteins, and this subject has recently been reviewed (133, 134). Briefly, helices are stabilized by a variety of forces beginning with their intrinsic conformational preferences (135); Ala is the most helix-stabilizing amino acid, whereas Gly breaks helices, with an energy difference between the two of ~1 kcal/mol (136). Experimental scales for the helix-forming tendencies of the amino acids obtained by using synthetic peptides agree well with data from mutagenesis of natural proteins (137, 138), statistical surveys of helix formation (139), and theoretical investigations (140, 141). The helix-forming potential of the amino acids may be understood largely in terms of the conformational entropy, and—to a limited extent—the solvent-accessibility of the side chain and main chain in the random coil and helical states (121, 133). H-bonded interactions between side-chains and main-chain atoms must also contribute to helix propensity. For example, Asn is a helix breaker, whereas the nearly iso- steric Leu is a helix former, presumably because Asn is able to stabilize a number of nonhelical conformations through H-bonded interactions (40). Side-chain–side-chain interactions have also been implicated in stabilization of helices. Hb (121, 142, 143), aromatic (144), and electrostatic (145, 146) interactions between residues separated by one helical turn (3–4 residues) can stabilize helices by 0.5–1.5 kcal/mol. Stabilizing electrostatic interactions between charged residues and the partial charges associated with the N- and C-termini of the helix can stabilize a helix by ~0.5 kcal/mol (139).

Two groups have examined the conformational relaxation kinetics of Ala-rich, monomeric α-helical peptides, after a very brief laser-induced temperature jump. Williams et al. used the amide I′ IR absorption to monitor the kinetics of helix formation (147), whereas others have instead measured shifts in the fluorescence spectrum of an extrinsic fluorophore added to the N-terminus of the peptide (148). Interestingly, the relaxation kinetics differ by a factor of eight for the two systems: 160 ns from the IR measurements and 20 ns from the fluorescence measurements. Both values are considerably faster than the rate of formation of native structure in proteins, suggesting that helix formation may frequently be one of the first steps in folding (149).

Helix Capping and Interhelical Loops

Capping motifs play important roles in the grammar and syntax of protein folding, marking the beginning and end of α-helices and helping to define the geometric relationships between adjacent elements of secondary structure. The amide protons and carbonyl oxygens, respectively, project from the N-terminal
and C-terminal ends of the α-helix, presenting groups with unsatisfied hydrogen bonds. In the classical N-cap motif (Figure 2, center), the side chain of an Asn, Asp, Thr, or Ser residue hydrogen bonds to one or more of the exposed amide protons near the N terminus of the helix (40, 150, 151). The N-cap residue typically adopts a β-conformation (152), which projects the chain in a direction orthogonal to the α-helical axis. The residue preceding the N cap is often apolar and forms a stabilizing Hb interaction with the side chain of the fourth helical residue, in a stabilizing motif entitled the “hydrophobic staple” (153). N-capping motifs may also be reinforced by H-bonded interactions between the N-cap main chain amide and the side chain of a Glu or Gln at N3 (the SXXE box) (40, 152, 154). The C-terminal ends of helices frequently terminate in a Schellman (40, 155–157) motif (Figure 2, left) in which the chain doubles back on itself, hydrogen bonding to the exposed, helical, carbonyl oxygens. As in the Hb staple, apolar residues at positions before and after the C cap are able to interact in a stabilizing interaction (156–158). In this motif, the C-cap residue is frequently Gly, which adopts an αL conformation. A second helix C-capping motif places a Pro residue at a position that is 2 residues removed from the C-terminus of the helix (157, 159). Whereas helix N-capping motifs strongly stabilize monomeric helices by ≈2 kcal/mol, natural C-capping motifs do not stabilize monomeric helix formation, although a D-Arg at a C-cap position stabilized a monomeric helix by 1 kcal/mol (160). Thus, in proteins, C-cap motifs may primarily serve as elements of negative design, important for breaking helices.

Short interhelical loops are most frequently composed of simple combinations of common helix N-cap and C-cap motifs. For example, a very common 3-residue helical hairpin loop (161) has a conformation designated as helix-(αL-β-β)-helix, in which the first loop residue adopts a conformation from the left-handed α-helical region of the φ,ψ map (αL), and the next 2-loop residues adopt β-conformations. This loop actually contains a nested Schellman motif (helix-αL-β) and an Hb staple (β-β-helix) (Figure 2). A similar hairpin loop with a conformation Helix-(αL-β)-helix is found in numerous proteins, including the repressor of primer (ROP). The first residue in the loop has been mutated to all commonly occurring amino acids in this protein. The nature of the substitution modulated the stability of the protein over a range of 3.6 kcal/mol (162), and the stability of the mutants correlates with the ability of the residue to adopt the αL conformation.

### Turns in β-Hairpins

β-Hairpins consist of a pair of antiparallel β-strands connected by a short turn. The turn regions of β-hairpins generally lie on the surface of proteins and frequently mediate molecular recognition in peptide hormones, antibodies, enzymes, and receptors. Turns have been the topic of several excellent, exhaustive reviews (163–165). Here, we highlight design principles only for a subset of turns that occur frequently within β-hairpins.
Hierarchic design of a helix-loop-helix supersecondary structural motif from a Schellman/Rose C-cap motif and a hydrophobic staple N-cap motif. The Schellman motif (left) is characterized by an $\alpha_L$ conformation at the C-cap position (the first nonhelical residue), followed by a residue that adopts a $\beta$-conformation in 75% of 261 C caps examined by us. A hydrophobic interaction is often observed between the residue after the C cap and a side chain three to four residues before the end of the helix. The hydrophobic staple N-capping motif (center) often involves two residues in a $\beta$ conformation, preceding the N terminus of the helix. The first residue forms a hydrophobic interaction with one or more residues in the helix while the N-cap (in this case Asp) hydrogen bonds to exposed amides at the N-terminal end of the helix. Combining these two motifs provides a right-handed $\alpha$-$\alpha$ hairpin, as described by Efimov (307). Note that the nomenclature for C-cap and N-cap positions follows Gunasekaran et al (157) and differs from that seen in Aurora et al (156).

Until recently, the role of turns in determining the structure and stability of proteins was a topic of some debate. While some investigators suggested that turns contribute significantly to stability (162, 166–170), others suggested that they merely serve as semiflexible links between more rigid elements of secondary structure (171–173). Recent studies with circularly permuted proteins and mutants with altered turn sequences have shown that natural turn sequences stabilize the overall structure by ~2–5 kcal/mol, relative to a typical random sequence of similar length (169, 174–176). Thus, a typical protein with a
stability of -6 to -10 kcal/mol can tolerate major mutations to a single loop, and remain folded. However, a protein typically contains a large number of loops; simultaneous unfavorable mutations of all of the turns could therefore lead to an unfolded or misfolded protein. Thus, the sequence of turns represents an important consideration for de novo design.

A number of different nomenclatures have been proposed for designating turns (177, 178). Here, we refer to the residue immediately before the turn as position \( i \) and each subsequent residue as \( i + 1, i + 2 \), etc. Many turns within \( \beta \)-hairpins are delineated by a hydrogen bond between a carbonyl of a residue at the first \( (i) \) position and an amide NH at position \( i + 3 \) (\( \alpha \) or 2-residue turn), \( i + 4 \) (\( \alpha \) or 3-residue turn), or \( i + 5 \) (\( \pi \) or 4-residue turn) (177, 179; Figure 3). Two-residue \( \beta \)-turns are most abundant in hairpins and resemble the letter “U” but with one of the vertical lines twisted above the plane of the paper and the other line below this plane (177). The I and II subtypes of 2-residue turns have a right-handed twist, whereas I’ and II’ turns have the opposite handedness. \( \beta \)-Sheets show a right-handed twist, which is complementary to that of type I’ and II’ turns. Thus, these turns occur in \( \beta \)-hairpins with greater frequency than the most common type I turn, which generally occurs in other structural contexts.

Nature and protein designers use several strategies to specify a given turn type. Because turns frequently project towards solvent (163), inclusion of the appropriate hydrophilic residues will tend to encourage the formation of a turn. However, if optimal thermodynamic stability or a specific geometry is required, a turn subtype may be “dialed in” by including residues that are most compatible with the desired turn’s main chain torsion angles (\( \phi, \psi \)) (27–32). For example, the type I \( \beta \)-turn requires \( \phi \) and \( \psi \) angles that are sterically allowed for L-amino acids, and hence this turn occurs most frequently in proteins. This turn may be further stabilized by the appropriate inclusion of a Pro residue at position \( i + 1 \). The pyrrolidine ring restricts its own \( \phi \) angle to approximately \(-70^\circ\), while simultaneously restricting the \( \psi \) angle of the preceding residue to positive values (180, 181). The type II, II’, and I’ turns require one or two \( \phi, \psi \) combinations that are generally not allowed in L-amino acids, and Gly, Asn, or Asp (which frequently adopt conformations from the left-handed portion of the \( \phi, \psi \) map) generally occur at these positions in proteins. Introduction of a D-amino acid at appropriate positions can provide an even stronger constraint favoring a specific type of \( \beta \)-turn (163). In particular, heterochiral Pro-Pro dipeptides (or dipeptides containing a Pro followed by an N-methyl amino acid) have very strong tendencies to form type II or II’ turns (with L,D and D,L dipeptides, respectively) (182–186). Chemists have also devised a large array of C\( \alpha \)-dialkyl amino acids, organic templates, and bicyclic amino acids, which can structurally constrain or stabilize turns via hydrophobic interactions (5, 187, 188).

Turns are frequently stabilized by side-chain/main-chain interactions in a manner entirely analogous to helical N capping. The preferred main-chain conformation associated with N-capping interactions (152) is quite similar to the first 3 or 4 positions of most common turns. Thus, it should not be surprising that N-capping residues occur quite frequently at the analogous position of type I \( \beta \)-turns, as well as the \( \alpha \) and \( \pi \) turns illustrated in Figure 3. Similarly, Asp is highly favored at the
Figure 3  Common turn types as compiled by Gunasekaran et al (179). The conformations of the turn-forming residues (e.g. at positions $i$ and $i+1$ in the $\beta$ turn) are shown to the right in red. The symbol $\alpha$ refers to a residue with $\phi$, $\psi$ angles from the right-handed $\alpha$-helix portion of the Ramachandran plot, $\alpha_L$ is from the left-handed $\alpha$-helical portion, and $E'$ is the left-handed $\beta$ region.

$i+1$ position of a type $I'$ $\beta$-turn, where it forms a stabilizing interaction with a main-chain amide (189). Cys residues, also, often participate in N-capping interactions in 2-, 3-, and 4-residue turns, where they engage in H-bonded interactions that can specifically stabilize the thiolate form of the Cys side chain. This structural feature is frequently important for stabilizing thiylates for interaction with metal ions, as in the zinc fingers (Figure 4a, b). For example, attempts to replace a buried thiolate in a zinc finger with Hb groups resulted in a change in the turn structure (34).

Design of $\beta$-Hairpins and Monomeric $\beta$-Sheets

Figure 5 illustrates a typical four-stranded antiparallel $\beta$-sheet consisting of two $\beta$-hairpins (64, 190). In this conformation, side chains alternate project up and down as one progresses along the strand. The neighboring side chains across the sheet approach closely in one of two nonequivalent patterns of interaction defined by the H-bonding of the backbone atoms. In the smaller 10-atom H-bonded rings,
Figure 4  Structure of natural and designed zinc finger domains. The second zinc finger domain from ZIF268 protein (308) (a), and its β hairpin with H-bonds between the amide protons and a Cys thiolate (b). Dahiyat & Mayo (34) used a computer algorithm to fully redesign the sequence of the zinc finger domain to produce the FSD-1 protein (c). The removal of the Zn$^{2+}$-binding site caused the site of the turn to be altered, primarily because of the loss of specific H-bonded interactions between backbone amide nitrogens of the turn and the cysteine thiolates (b).

the distance between adjacent Cα atoms is 5.5 Å, whereas the 14-atom H-bonded ring’s position the Cα atoms within ~4.5 Å (Figure 5, bottom). The two side chains projecting from neighboring Cα atoms interact strongly (191–193), with energetics that differ between the two distinct rings. The commonly occurring amino acids also have a range of intrinsic energetic preferences to adopt an antiparallel β-conformation, which have been elucidated from statistical studies as well as site-directed mutagenesis of β-sheets (194, 195), and the β-hairpin of a zinc finger (196). These parameters provide excellent guides for construction of β-sheets.

With very few exceptions (132, 197), peptides spanning the sequences of β-hairpins in proteins are not folded in aqueous solution. Therefore, an important early goal in peptide and protein engineering was to design linear peptides that
Figure 5  The structure of a four-stranded antiparallel β sheet [Protein Data Bank (PDB) entry 1QAT (309)]. *Top:* The typical right-handed twist, the pleating of the backbone, and the up/down periodicity of the side chain atoms. A residue whose Cβ is pointing “down” (yellow) is immediately followed by a residue whose Cβ is pointing “up” (blue). Bulge residues (red) cause this regular periodicity to be disrupted. *Bottom:* The cross-strand residue interactions can be classified as belonging to either a small or a large H-bonded ring. A bulge occurs when the normal H-bonding pattern is disrupted. Carbon atoms are shown in green, nitrogen atoms in blue, and oxygen atoms in red. H-bonds are represented as dashed lines.
adopt this conformation in water. The first such peptide consisted of a 9-residue segment from tendamistat, with an idealized type I beta turn, NPDG, replacing the protein’s native turn (198). This tetrapeptide represents a consensus sequence for a type I β-hairpin (199, 200); in proteins, Asp at the i position (as defined in Figure 3) may form an N-cap interaction, Pro and Asn further stabilize the structure via their intrinsic conformational preferences, and the Gly adopts an unusual conformation that counteracts the normal right-handed twist of a type I turn (facilitating the connection to the antiparallel β-strand). The NPDG sequence, however, may not be an ideal choice for de novo design, because it can also adopt a 3-residue turn, designated a type I-G1 bulge (201). Indeed, insertion of this sequence into a 16-residue peptide from a β-hairpin of ubiquitin gave rise to a peptide that partially populated a type I-G1 bulge conformation, even though this disrupted the native registry of the two β-strands (202). Similarly, two de novo-designed decapeptides containing NXDG turn sequences showed a mixed population of 2-residue and 3-residue loops (203).

The type I’ turn is the most frequently occurring loop in β-hairpins and has been reliably engineered in short peptides by inclusion of an Asn-Gly dipeptide into the turn. For example, a designed 12-residue peptide containing this motif is partially folded into the desired β-hairpin structure (204).

In 1998, four groups published the design of three-stranded β-sheets, an accomplishment with very significant implications for protein design. Schenck & Gellman (205) described a three-stranded structure with two D-Pro-Gly turns, compatible with forming a type II’ β-turn (206). NMR studies of this peptide indicated that it folded into the desired three-stranded structure, and the chemical shifts of the Cα protons suggested that the folded conformation was substantially—but most probably not fully—populated in solution. Das and coworkers also described a three-stranded β-sheet stabilized by D-Pro-Gly type II’ turns. This peptide was too hydrophobic to dissolve in water, but adopted the desired conformation in organic solvents (207).

Kortemme and coworkers have also prepared a 20-residue three-stranded β-sheet containing Asn-Gly type I’ beta turns (see Figure 6; 208). Key to their success was the incorporation of an aromatic cluster along one face of the sheet, patterned after the β-sheet from the WW domain (209). NMR and thermal denaturation studies indicated that the peptide adopted the desired three-stranded structure, with a stability of ~0.6 kcal/mol. This value corresponds to an equilibrium constant of 2, indicating that the peptide is ~65% folded at room temperature. In a related paper, Sharman & Searle described a similar three-stranded sheet with an Asn-Gly at the turn sequence (210). Although this peptide formed the desired fold in 50% methanol, it was not structured in aqueous solution. This result illustrates the importance of the sequence of the β-strands in defining the conformation of the sheet. All four of these three-stranded structures are more completely folded than a comparable two-stranded hairpin, indicative of a small but significant degree of cooperativity, which might arise from a reduction in the entropy of folding for the second hairpin, once one of the hairpins has already formed.
DESIGN OF GLOBULAR PROTEINS

The de novo design of globular proteins has been the subject of several in-depth reviews (211–215). The following sections focus primarily on those designs that have been structurally characterized at high resolution.

Characterization of Designed Proteins

The defining feature of a native protein is a well-defined 3-dimensional structure, stabilized by the formation of a tightly packed Hb core. Initially, the de novo design of proteins containing this feature was surprisingly difficult, often resulting in a more dynamic ensemble of loosely folded molten globules (2, 44, 216). The molten globule is a nonnative folded state with a broad thermal unfolding transition, characterized by low per-residue values of $\Delta H^0$ and $\Delta C_p$ (217, 218). It is, however, important to stress that there is a continuum of folded states available to a protein, and the terms native, molten globule, and unfolded states represent convenient but somewhat imprecisely defined boundaries between various extremes. For example, native proteins generally show regions of high mobility (see 219 and references cited within), and the unfolded state is far from a featureless random coil (220). Thus, it is important to perform multiple solution measurements to determine where a designed protein lies along this continuum of states.

Along these lines, a number of tests for the molten globule state have been proposed (221), including the measurement of $\Delta C_p$ and $\Delta H^0$ for unfolding, the ability to bind Hb dyes, the degree of dispersion in an NMR spectrum, the

Figure 6  Solution structure of a de novo–designed three-stranded $\beta$ sheet. Two views of a designed, 20-residue $\beta$ sheet (208) are shown, including two stabilizing aromatic residues.
cooperativity of the thermal unfolding curves, and the rates of exchange of the amide protons with solvent (HDX). Of these parameters, we have found that $\Delta C_p$ is most objective and discriminative. Its value is generally low for molten globules but invariably large for native proteins (10–15 cal degree$^{-1}$ residue$^{-1}$) (222). A second important distinguishing feature is the rate of HDX. Native proteins, but not molten globules, show a subset of amide protons that exchange only in the fully unfolded state (223, 224). The protection factors (the ratio of the rate of HDX for a given position in the random coil to the rate in the native state) of these amides therefore reflect the global stability of the protein, and these two values provide an important test for nativelike behavior. Ultimately, the determination of the high-resolution structure of a designed protein by NMR or X-ray crystallography provides the most critical evaluation of a design. However, one cautionary note is in order; several designed helical-bundle peptides and coiled coils that adopt multiple conformations in solution have been crystallized in only one of the multiple conformations found in solution (216, 225–227). Care must, therefore, be taken in the interpretation of structural data gained through the use of X-ray crystallography without the benefit of corroborative solution data. The work of Schafmeister et al (28) is a particularly thorough example of the use of both techniques to unambiguously characterize a de novo designed protein.

Monomeric Helical Hairpins

Perhaps the simplest model for a folded protein is an antiparallel pair of hydrophobically interacting amphiphilic $\alpha$-helices, connected by a covalent loop. Two groups have designed monomeric helix-loop-helix motifs, using only sequence patterns and schematics as guides (8, 229–231). Both peptides assumed the desired fold, suggesting that Hb interactions alone may be sufficient for folding. However, the structures showed marginal stability, and, as illustrated in Figure 7 (left), the conformations of buried Hb side chains at the helix-helix interface were poorly defined. Thus, additional conformational specificity may require a more extensive set of specific interactions.

Coiled Coils

The $\alpha$-helical coiled coil (68, 232) classically consists of a bundle of parallel or antiparallel $\alpha$-helices with a left-handed superhelical twist. Coiled coils generally contain a 7-residue (heptad) repeat whose positions are designated $a$–$g$, with the $a$ and $d$ positions directed towards the interior of the structure (Figure 8). A large number of coiled coils have been engineered, and structures are now available for designed parallel trimers (225, 233–237), antiparallel trimers (238), and parallel tetramers (48). Because the design of coiled coils has been recently reviewed (2, 239–242), we highlight findings that are most pertinent to de novo protein design.

Hb interactions between residues at $a$ and $d$ are essential for mediating the association of $\alpha$-helices in coiled coils (79, 243). However, more precise details affect
the resulting topology, as illustrated in the encompassing contributions of Harbury et al (48, 233) on mutants of the two-stranded coiled coil GCN4. For example, derivatives of GCN4 with Leu at position $a$ and Ile at $d$ (designated Leu $a$ Ile $d$) form tetramers, whereas Val $a$ Leu $d$ forms a mixture of dimers and trimers, and Ile $a$ Leu $d$ forms dimers (48, 233). Crystallographic studies indicate that the side chains in these peptides pack efficiently in the center of the structure and adopt well-defined, low-energy conformations. Thus, the steric interactions of apolar side chains dictate the fold of these peptides (69, 244), rather than vice versa.

A buried, polar H-bonded interaction between two pseudosymmetrically disposed Asn side chains is important for maintaining the dimeric conformation of GCN4 (48). This is an excellent example of “negative design”; the H-bonded interaction is actually less stabilizing than a corresponding Hb interaction. However, it more strongly destabilizes alternatively folded states, which would require full or partial burial of the polar Asn side chains in less favorable environments. Additionally, electrostatic interactions between partially exposed residues that interact across the helix-helix interface are important for conformational specificity. Under physiological conditions, these residues are more important for destabilizing alternatively folded structures than for stabilizing the desired fold (48).

Hodges and coworkers have pioneered the de novo design of coiled coils, and they have prepared a host of coiled coils, including structures that self assemble in response to metal ions. These peptides, far too numerous to enumerate in the present account, are the subject of an excellent recent review (245). Ogihara et
al (237) and Lovejoy et al (238) have also designed and crystallographically characterized a series of trimeric coiled coils that either adopt antiparallel [coil-Ser, which has Leu at $a$ and $d$ (238)] or parallel [coil-$V_aL_d$, with Val at $a$, Leu at $d$ (237)] geometries depending on the steric interactions of the side chains in the core. Interestingly, coil-$V_aL_d$ undergoes a pH-dependent transition from a trimeric coiled coil at neutral pH to a dimeric structure at low pH, which minimizes the electrostatic repulsion between the peptide chains (126). The introduction of a buried polar interaction between Asn side chains (as in GCN4) also switches this peptide from a trimer to a dimer (216).

From Coiled Coils to Globular Proteins

Coiled coils and self-associated helical bundles (15, 246) can serve as convenient starting points for the design of more complex, nativelike structures. For example, Schafmeister et al (228, 247) have designed a single-chain, 4-helix bundle based on the structure of a 24-residue membrane-interactive peptide that serendipitously had been found to crystallize as an antiparallel 4-helix bundle. The packing of the side chains in this tetramer showed extensive shape complementarity in the arrangement of small Ala and large Leu side chains, similar to that of the antiparallel four-stranded coiled coil, ROP. Schafmeister et al next connected four identical copies of the helix with three Gly-rich loops, resulting in a single-chain, stable, folded protein ($\Delta G_{\text{fold}} = -9.3$ kcal/mol) that satisfied all the above-mentioned thermodynamic criteria for a native protein. These solution experiments strongly suggested that the protein folded into a well packed structure, and the crystal structure of the protein, solved to 2.9-Å resolution, is indeed

Figure 8  Schematic of the heptad repeat of an antiparallel coiled coil. The positions of the residues are assigned the letters $a$ through $g$ corresponding to a seven-residue repeat (heptad). The $a$ and $d$ residues are $H_b$ and well packed in the interior of the coiled coil. In this example, which is based on the crystal structure of a designed four-helix bundle, shape complementarity between large residues at $a$ and small residues at $d$ may help stabilize the antiparallel tetrameric conformation and destabilize alternatively folded structures (adapted from 240).
nearly identical to the starting, unlinked helical bundle. Unfortunately, perhaps because the four helices are identical, a crystallographic twofold axis was directed along the central axis of the bundle. Thus, it was impossible to observe significant electron density from the loops or to determine whether the bundle’s overall topology was clockwise, counterclockwise, or a mixture of these topologies [looking down the axis of the first helix, each additional helix may be added in a clockwise or counterclockwise manner (56)].

Bryson and coworkers (248) have similarly built a single-chain 3-helix-bundle protein, beginning with the crystal structure of coil-Ser, a previously designed antiparallel, three-stranded coiled coil with Leu at most of the $a$ and $d$ positions (135). This peptide crystallizes into a single, well defined structure, but in solution it shows less conformational specificity (216). With the crystal structure as a starting point, a uniquely folded 3-helix bundle protein could be engineered by (a) shortening the helices to a length typical of those found in helical bundles; (b) designing loops containing N-cap motifs to favor folding in a unique topology; (c) introduction of electrostatic interactions between the helices to stabilize desired helix/helix pairings and destabilize alternatives; and (d) repacking the core to contain a diverse collection of uniquely packed apolar residues by using the computer program ROC (96). The resulting 73-residue protein, $\alpha_3D$ (248; 249) contains 19 of the commonly occurring amino acids, has a stability typical of globular proteins (-5.1 kcal/mol), and behaves like a native protein by all criteria described above. Further, the NMR structure of $\alpha_3D$ is well defined and agrees reasonably well with the design, showing an overall rms deviation of 1.8 Å for the backbone atoms. As in the design, the protein adopts a counterclockwise topology rather than the clockwise topology observed in protein A and Z domains (250, 251).

Many of the features included in the design contribute to the singular fold of $\alpha_3D$. Interfacial Glu, Arg, and Lys side chains appear to be particularly important. They shield the Hb core from solvent, while also forming the designed interhelical electrostatic interactions. Further, the N-capping motif included to bias the conformation of the loops was observed in one of the interhelical turns. Finally, 16 of the 18 Hb side chains adopt unique conformations in the protein interior. Thus, it has been possible to engineer a uniquely folded protein, beginning with a rather nonspecific aggregate of amphiphilic helices. Efforts are currently underway to evolve the sequence of $\alpha_3D$ to allow binding of a variety of small and large ligands.

The structure of coil-Ser (Figure 9, center) has been used as a template for the design of a helix-loop-helix motif that binds to a complex naturally occurring $\alpha$-helical peptide, the calmodulin-binding domain of calcineurin (CN) (252). Two of the helices in the trimer served as templates for the receptor, whereas the third was a template for the bound peptide. The sequence of the CN peptide was threaded onto the backbone of one of the coil-Ser helices, and the remaining two helices were connected with a covalent loop (Figure 9, left). The sequence of the receptor was then varied to create a surface that would be
complementary to that of the bound CN helix. Polar side chains were introduced at solvent-accessible positions, and the buried apolar side chains were chosen by using the program ROC (96). The designed hairpin bound to the target with an affinity of ~1.0 μM and was highly specific, because it did not bind to a related calmodulin-binding peptide. In subsequent designs (253) a parallel three-helix bundle (237) served as a template, and the receptor helices were connected via a disulfide bond, allowing construction of multiple receptors by combinatorial recombination of different peptides. Receptors with affinities ranging from the low nanomolar range to >0.1 mM were obtained. Those with the poorest affinities did not fold, whereas the higher-affinity receptors showed a conserved sequence motif. Interestingly, many high-affinity complexes were not predicted to be stable by the ROC program, because their Hb side chains were longer than could be accommodated by the rigid backbone model of the starting structure. However, by allowing systematic variations in the geometry of the backbone template, all experimentally observed high-affinity receptors could be predicted.

Dimeric 4-Helix Bundles

Dimeric 4-helix bundles occur in nature as noncovalently self-assembled proteins such as ROP. One early designed dimeric 4-helix bundle, α₂B (15), was modeled as a pair of interconnected helices with Leu, Glu, and Lys as the sole amino acids in the helix sequences. Although this peptide dimerized to form an exceedingly stable bundle, it adopted a molten-globule-like conformation. One possible cause for this behavior is that such dimers may adopt four possible topologies with left- or right-turning loops on the same or opposite sides of the bundle (56); alternatively, the loops can cross over the bundle diagonally (254). With identical sequences for helices 1 and 2 and a flexible loop, these different topologies should be energetically degenerate. Thus, the sequence of α₂B was varied to remove the sequence degeneracy of the helices and impart a unique fold. In particular, approximately half of the Leu residues of α₂B were replaced by side chains with aromatic and β-branched apolar side chains, creating the next generation peptide, α₂C (255). Although these changes resulted in some increase in the natively like behavior, a fully native structure was obtained only after changing three of the interfacial apolar residues to polar residues (44). The resulting protein, α₂D, showed most hallmarks of a native protein, although it was thermodynamically less stable than α₂B. Thus, the substitutions that occasioned the appearance of natively like properties represent elements of negative design, as they destabilize alternatively folded structures rather than stabilizing a single, folded state.

The 3-dimensional structure of α₂D [Figure 7 (center) and Figure 10] was determined at high resolution (0.28 Å rms deviation for backbone atoms) by NMR (254). It features a twofold symmetrical pair of helix-loop-helix motifs with the loops draped diagonally across the ends of a 4-helix bundle, helping to
sequester its apolar core from solvent. Similar motifs have been observed in both monomeric and dimeric α-helical and β-sheet proteins. Because of its symmetry, α₂D contains three distinct helix-helix interfaces. The first is quite similar to the interfaces originally designed in α₂B and resembles an antiparallel Leu zipper with interdigitated Leu residues shielded by electrostatically interacting Glu and Lys residues (Figure 10a). The substitutions required for the native fold segregate to the other two helix-helix interfaces. One is quite rich in interacting aromatic side chains (Figure 10b) and also shows a pair of H-bonded His residues, whereas the other consists of a well-packed collection of diverse hydrophobes (Figure 10d).

In an elegant series of papers, Broo et al (256–258) described a series of dimeric 4-helix bundles that catalyze the hydrolysis of 4-nitrophenyl esters, with a rate enhancement of ~2 to 3 orders of magnitude relative to the small-molecule catalyst 4-methylimidazole (258). An interfacial His residue with a depressed pKₐ acts as a nucleophilic catalyst. A second, positively charged group on the protein stabilizes the anionic transition state for the reaction. The depression of the pKₐ of the active His is a result of its location within a partially Hb environment proximal to a positively charged residue separated by one turn of the α-helix (257). NMR studies provided evidence for the designed structure (258); long-range nuclear Overhauser effect cross peaks

**Figure 9**  From simple coiled coils to native functional proteins. A single-chain, three-helix bundle, α₃D (right), was designed by shortening the helices of coil-Ser and building interhelical loops. At left, a helix-binding peptide (SR1, blue) was engineered to bind the calmodulin-binding helix (red) from calcineurin.
Figure 10  Solution structure of the 4-helix bundle, α2D. The helical interface (a) shows a series of tightly interdigitated Leu residues. A second interface (b) displays a series of interacting His residues, whereas a third interface (d) is very rich in aromatic residues. Finally, the loop residues drape over the hydrophobic core (c) and sequester it from solvent. Adapted from Reference 254.

(NOEs) indicate the presence of one antiparallel helix-helix interface within the bundle. Interestingly, the NOEs involve a cluster of aromatic residues that appear to be particularly important for engendering a high degree of conformational order. However, attempts to calculate a 3-dimensional structure from NMR-derived distance constraints failed to produce a uniquely folded structure. Thus, these peptides may have hybrid properties of native and molten globule states.
β and α-β Proteins

One of the first attempts at de novo protein design focused on engineering β-sandwich proteins, betabellins, composed of two hydrophobically interacting four-stranded β-sheets (3, 16). Although the peptides initially showed rather poor solubility, subsequent redesign of the loops and incorporation of more polar residues resulted in structures that were monomeric and adopted the desired secondary structure in aqueous solution (259, 260). However, their NMR spectra were not sufficiently dispersed to allow structure determination, suggesting that they may retain some of the features of the molten globule (261).

REDESIGN OF NATURAL, FOLDED PROTEINS

Recently, a number of natural proteins have been redesigned, with changes that are sufficiently radical to be of interest to de novo design. For example, Tramontano and coworkers have miniaturized the Ig fold to a six-stranded structure, using the resulting “minibody” as a scaffold for molecular recognition (262). In another approach to protein redesign, large portions of a protein chain are replaced with sequences of minimal complexity (263–266). Recently, Riddle and coworkers have replaced nearly three-fourths of the sequence of an SH3 domain with a mixture of Ile, Lys, Glu, Ala, and Gly, with retention of its binding activity (11).

Three additional strategies have proven very successful for protein redesign. These approaches are based on genetic selection, computational repacking of the cores of natural proteins, and the construction of consensus sequences. In the following section, we provide individual examples of these three approaches.

A Sequence-Based Approach to the Design of a Soluble β-Protein

Mayo et al (267) and Ilyina et al (268) have designed a tetrameric three-stranded β-sheet, based on the sequence of cytokines PF4 and IL8. A peptide spanning the antiparallel three-stranded sheet from IL8 is partially folded in solution, so its sequence was modified to improve its water solubility and sheet-forming potential. The resulting peptide, β4, had 50% sequence homology with IL8 and formed tetramers with properties of both the native and molten globule states. Like the corresponding sequence in IL-8, β4 formed an antiparallel three-stranded β-sheet. The second and third strands of β4 are nearly identical to IL-8 in both amino acid sequence and structure. However, the more radically redesigned first strand adopted two distinct conformations, which differed in the registry of the sheets. The monomers assemble into C2-symmetric six-stranded sheets, with an extended Hb surface that packs against a pair of helices in IL-8. In β4 these surfaces instead interact in a number of dynamically interconverting conformations as assessed by the lack of intersheet NOEs. These pioneering studies, together with recent work on monomeric β-sheets (205–208), encourage one to believe that the de novo design of uniquely folded β-proteins is within reach.
Genetic Selection of an Fc-Binding Helix-Loop-Helix Motif

In a particularly elegant contribution, Braisted & Wells (269) and Starovasnik et al (270) used phage display to “minimize” an IgG-binding three-helix bundle protein (a “Z domain”) down to the size of a helical hairpin. The Z domain uses its first two helices to bind to the Fc region of IgG (271). The third helix plays a purely structural role and hence was genetically jettisoned. Although the resulting protein did not fold or function without this third helix, the method of phage display could be used to recover both folding as well as high-affinity binding. This was accomplished by first randomizing the residues in helices 1 and 2 that packed together with helix 3 in the wild-type Z domain and next selecting for functional proteins. Interestingly, five hydrophobes in helices 1 and 2 were converted to more polar groups in the highest-affinity binders. Biophysical studies of the redesigned miniprotein (270) showed that it was only partially folded, but that its stability could be increased by introduction of an interhelix disulfide bond. The NMR structure (Figure 7, right; 270) of the resulting 34-residue hairpin (Z34C)—in contrast to the above-mentioned designed α-t-α hairpins (Figure 7, left; 229, 230)—showed a very well defined Hb core between the two helices. This core is stabilized by interactions with many of the side chains introduced in phage selection (Arg-13, Lys-31, and Arg-36). The apolar portions of these residues pack against the core, whereas their polar groups are exposed to solvent. It is interesting that the original, Hb side chains at these positions in the wild-type sequence could also sequester the core residues. However, they might permit too many nonspecific associations, leading to an ensemble of partially folded structures. By contrast, the amphiphilic side chains of Lys and Arg provide fewer nonproductive associations and favor a uniquely folded structure.

Redesign of Zinc Fingers

The 2-His,2-Cys zinc finger motif, which consists of an α-helix packed against an antiparallel β hairpin, contains a small Hb core adjacent to a zinc-binding site. This small, 25- to 30-residue motif requires Zn(II) for folding (272). Recently, two groups have designed peptides that adopt the zinc finger fold, even in the absence of Zn(II) (34, 106, 273–275). These studies were particularly significant because the designed peptides were about half the critical length previously thought to be required for folding. In both cases, the Zn(II)-binding ligands were replaced with residues that extended the Hb core of the peptide motif. Struthers and coworkers’ initial design involved two key modifications: a constrained D-Pro-Ser dipeptide sequence was substituted into the hairpin to strongly stabilize a type II’ conformation, and a large Hb 3-(1,10-phenanthrol-2-yl)-L-Ala residue was introduced at the former metal-binding site (274). Subsequently, these workers showed that this large, bulky group could be replaced with a collection of natural amino acids, with retention of structure (273). Further, they demonstrated that the polar nature of residues on the exterior of the hairpin was essential for maintaining a uniquely folded structure in aqueous solution (275).
Dahiyat & Mayo (34) have used their side-chain repacking algorithm to simultaneously repack the entire 28-residue sequence of a ββα motif, in the first example of fully automated protein design. Beginning with the coordinates of the second zinc finger from the natural DNA-binding protein Zif268, they designed a peptide that was intended to fold in the absence of metal ions. The peptide formed a marginally stable structure with somewhat broad unfolding transitions, as would be expected from its very small size. The solution NMR structure of the peptide is remarkably similar to the design target (Figure 4), showing the expected elements of secondary structure, an N-capping interaction, and unique and predetermined apolar side-chain conformations. The only area of serious deviation lies along the 4-residue β-turn and an adjacent β-strand. Interestingly, in the wild-type protein this turn is stabilized by an interaction between two of its amide protons with a buried thiolate of a Cys Zn(II) ligand. The side-chain repacking algorithm replaced the space occupied by this buried Cys residue with Hb side chains, resulting in a destabilization of the wild-type turn conformation. This finding speaks to the difficulty of designing buried polar interactions, which often serve very important structural and functional roles.

METALLOPROTEIN DESIGN

Proteins use a very limited set of metal ions to mediate a variety of different processes. Often the same metal center can serve a number of roles; for example, binuclear iron centers engage in reversible binding of oxygen, electron transfer, and the catalysis of hydroxylation reactions. How a protein matrix is able to finely tune the properties of a metal ion to effect such different functions is an important question that may now be addressed by de novo protein design (13, 276). Metal-binding sites in proteins may be classified as serving either structural or functional roles. Structural sites generally stabilize or direct the folding of the protein and exhibit common, coordinately saturated geometries that are well preceded in small-molecule/metal ion complexes. By contrast, functional metal-binding sites often show more unusual ligation geometries, which are largely preorganized in the folded apo-protein.

Zinc-Binding Proteins

Zn²⁺ plays an important structural and functional role in a variety of proteins, including hormones and enzymes (277). Before discussing the design of metalloproteins, it is instructive to consider some of the features required for the activity of a simple metalloenzyme. Figure 11 illustrates a schematic diagram of the Zn²⁺-binding site from carbonic anhydrase, which catalyzes the hydration of CO₂ (278). The catalytic Zn²⁺ is coordinated by three His ligands in a tetrahedral geometry, with the fourth position available for catalysis. The His imidazole rings are positioned by H-bonded interactions with the nonchelating nitrogen atom. These second-shell effects are important for modulating metal ion affinity, specificity, and
function. Further, the ion-binding site lies deep within an Hb pocket in close proximity to a general acid/base and other residues important for shuttling protons in and out of the active site. By comparison with even this relatively simple site, the metal ion-binding sites designed to date are quite primitive. Generally, only first-shell effects have been considered, and the positioning of the metal-binding site within a cavity containing additional functional groups presents an important future challenge. However, as described below, the state of the art is advancing quite rapidly, and the design of highly sophisticated sites is well within reach.

In early work, 3-His carbonic anhydrase-like Zn\textsuperscript{2+} -binding sites were introduced into the de novo designed four-helix bundle $\alpha_4$ (123), and the minibody (279), a redesigned antibodylike protein. The designed proteins were found to bind Zn\textsuperscript{2+} with affinities in the nanomolar to low-micromolar range, and NMR spectroscopy indicated that the $\alpha_4$-derived protein indeed formed a three-coordinate site consisting of three His side chains. The binding of Zn\textsuperscript{2+} served to increase the degree of conformational specificity in the $\alpha_4$ protein, decreasing some of its molten globule-like characteristics and forcing it to adopt a unique topology (280). The $\alpha_4$ fold was also used as a template for engineering a closed-sphere Cys\textsubscript{2}His\textsubscript{2} Zn\textsuperscript{2+} -binding site, similar to the structural site in His\textsubscript{2}Cys\textsubscript{2} zinc fingers (281). The direct contribution of each ligand to the stabilization and coordination geometry of a metal site was also investigated (282) through the synthesis of derivatives lacking one of the chelating groups.

Much work has focused on the use of automated computational algorithms for the introduction of four-coordinate Zn\textsuperscript{2+} -binding sites into natural protein frameworks of known structure. The program MetalSearch (127) was used to introduce a His\textsubscript{3}Cys Zn(II)-binding site into a B1 domain of streptococcal protein G, a 56-residue domain composed of a four-stranded $\beta$-sheet crossed by a single $\alpha$-helix. Three different variants were designed to stabilize the desired ligand geometry to different extents. UV-visible, circular dichroism (CD), and NMR spectroscopy showed that each variant bound Zn\textsuperscript{2+} in a tetrahedral geometry with dissociation constants in the low-nanomolar range; relatively small differences were found between the three variants, highlighting the flexibility of the protein core. Similarly, the program Dezymer (90) has been used to prepare a family of seven different Cys\textsubscript{2}His\textsubscript{2} Zn\textsuperscript{2+} -binding sites located at different portions of the Hb core of thioredoxin (283). Five proteins bound Zn\textsuperscript{2+} with nanomolar to micromolar affinities, with binding constants that appear to be modulated by steric interactions between the buried, designed coordination sphere and the surrounding matrix.

**Mercury-Binding Peptides**

Dieckmann et al (125, 126) have used Cys-containing Hg\textsuperscript{2+}-binding peptides to investigate how the protein matrix can dictate a somewhat unusual ligation geometry on a metal ion. While Hg\textsuperscript{2+}-thiolate complexes generally prefer linear 2-coordinate geometries, the MerR protein instead binds this metal ion in a three-coordinate trigonal arrangement (284). The requirements for adopting this
Figure 11  Primary and secondary Zn\(^{2+}\)-chelating ligands in the carbonic anhydrase active site. A schematic of the zinc-binding site of carbonic anhydrase is shown. The site consists of groups falling into three tiers of importance—those groups that directly ligate the metal (green) and those groups that modulate the metal ligation through secondary (red) and tertiary (black) effects. Adapted from Reference 278.

geometry were explored by placing a Cys residue at the \(a\) or \(d\) position of a series of three-stranded, parallel coiled coils. A three-coordinate mercury/thiolate complex was formed when the Cys residue occupied the \(a\) position—but not the \(d\) position—of a coiled coil. Modeling studies were consistent with this finding and showed that the Cys ligands converged into a well defined trigonal site only when they were placed in an \(a\) position. It is noteworthy that this designed peptide is the first water-soluble and stable model capable of mimicking the proposed geometry found in the metalloregulatory MerR protein. Interestingly, these peptides also formed dimers rather than trimers at low pH. Under these conditions, both the \(a\)- and \(d\)-substituted peptides formed Hg\(^{2+}\) complexes with spectroscopic features similar to those observed for linear two-coordinate complexes.

Iron-Sulfur Clusters

Iron-sulfur proteins are involved in a large number of different electron transfer processes. Cuboidal Fe\(_4\)S\(_4\) sites exist in nature in a +1, +2, or +3 oxidation state, with redox potentials ranging from -700 mV to +450 mV depending on the redox couple and the structure and composition of the surrounding protein matrix (285). Several cuboidal Fe\(_4\)S\(_4\) centers have now been designed, providing simple systems for future investigations into how the protein matrix is able to tune the properties of this cluster over such a wide range.
An Fe₄S₄ center has been introduced into a helix-loop-helix peptide known to associate to form a four-helix bundle (286). A 16-residue peptide derived from a consensus motif of natural ferredoxins was incorporated into the loop region of this peptide, resulting in the formation of a dimeric Fe₄S₄ center. The helical regions of this model four-helix bundle are also able to bind up to four heme centers, with bis-His ligation. Although the detailed overall structure of the Fe₄S₄-heme peptide is not known, this work demonstrates the feasibility of developing highly complex model proteins containing multiple redox cofactors.

The designed protein α₄ (287) has also been used as a template to create a mimic of the F₄ iron-sulfur center found in photosystem I (288). The Cys residues in this cluster are located in two different subunits, giving rise to a coordination requiring long-range tertiary interactions between two loops. The decapeptide sequence corresponding to the F₄ domain was thus inserted into the interhelical loops 1 and 3 of the α₄ motif situating a cluster at the apex of the model protein. The resulting protein formed a low-potential [Fe₄S₄]²⁺/₁⁺ cluster, with a redox potential more positive than that of the natural F₄ cluster. Thus, features lying outside of the spliced loops must be important for defining the potential of this center.

The program Dezymer has also been used to introduce a buried cuboidal Fe₄S₄ cluster into the structure of thioredoxin, which normally does not bind transition metal ions (289). The sequence and secondary structure of the designed site did not resemble known ferredoxin motifs, demonstrating the feasibility of designing such sites entirely from first principles. The resulting protein shows a high-potential [Fe₄S₄]²⁺/₃⁺ center with properties consistent with its location in an Hb cavity without stabilizing second-shell H-bonding effects. The thioredoxin scaffold was also used for engineering a mononuclear iron-sulfur center (130). The conversion of a disulfide bridge of the thioredoxin active site into two coordinating Cys residues, together with the insertion of two additional Cys side chains, gave rise to an FeS₄ mononuclear center. The resulting protein binds Fe³⁺ and Co²⁺ in a 1:1 complex with electron paramagnetic resonance and electronic spectra analogous to natural rubredoxin. The above designs considered only the geometry of the primary coordination shell around the metal ion, illustrating that simple geometrical considerations are sufficient for achieving the desired stoichiometry and geometry. It will now be interesting to use these ferredoxin and rubredoxin mimics as models to determine how solvent accessibility, hydrogen bonding to the ligands, and surface charge affect the function of the centers.

**Heme-Protein Models**

Heme proteins serve a variety of functions, including reversible O₂ binding, O₂ activation, decomposition of peroxides, and electron transfer. Recently, a number of model heme proteins have been prepared that should allow one to directly determine how the protein matrix tunes the properties of the heme to evoke such a wide variety of activities. Choma et al (124), Robertson et al (128), and Rabanal et al (290) have described a number of four-helix-bundle proteins that spontaneously
fold in aqueous solution to form heme-binding sites. Initial work focused on a
derivative of a dimeric four-helix bundle, \( \alpha_2 \beta \) (15), which was engineered to cre-
ate an Hb cavity to accommodate a bis-His–ligated heme (124). The peptide
indeed assembled to form the desired complex, and the redox potential and spectral
properties of the heme derivatives indicated that they resided within the
desired cavity but were more accessible to solvent than originally anticipated.

Next, a series of model peptides based loosely on the B and D helices of
cytochrome \( bc_1 \) were constructed (128, 290–293). The prototype (128) is made
up of a 62-residue parallel \( \alpha_2 \) system, in which the two helices are cross-linked by
a disulfide bridge. Each helix contains two His side chains separated by 14
residues, and the \( \alpha_2 \) dimers associate to form a four-helix bundle, resulting in four
bis-His heme-binding sites per four-helix bundle. Their spectral, heme-binding,
and electrochemical properties, including heme-heme redox interactions, closely
resemble those of native proteins and are fully consistent with the working model
for their structures. However, both the apo-peptides and the heme complexes
showed dynamic behavior that precluded structure determination by NMR. To
improve the properties of the prototype and to obtain functional molecules with
more nativelike structures, a series of parallel and antiparallel, heme-binding four-
helix bundles have been synthesized. These variants differ from the prototype in
the peptide chain composition and/or in the porphyrin moiety (291–293). Unfortu-
nately, the heme-bound states of these proteins have significant dynamic charac-
ter, precluding structure determination by NMR. However, considerable progress
has been made in designing altered sequences that fold into relatively unique struc-
tures in the absence of heme (292). Similar improvements in the properties of the
heme-ligated forms of these proteins should be possible and will provide better
characterized, simple systems for examining heme/protein interactions.

In an effort to determine the stringency of design necessary to obtain a suit-
able heme binding site, Rojas et al (85) have screened for heme binding with a
semi-random library of peptides generated with a specific Hb/Hp pattern. Of the
30 sequences tested, 15 bind heme with a range of affinities and show the charac-
teristic red color of 6-coordinate Fe\(^{3+}\)-heme. This result is particularly encour-
gaging because it suggests that very diverse sequences can be designed which are
capable of binding heme. The challenge still rests on the designer, however, to
fine-tune heme-binding affinity through thoughtful manipulation of the side
chains in the heme-binding pocket.

Other investigators have focused on the design of porphyrins with covalently
bound peptides. The first example, developed by Sasaki & Kaiser (294, 295), fea-
tured four identical 15-residue peptide chains covalently linked to the four propi-
onic groups of coproporphyrin I. Surprisingly, this peptide was reported to have
mild aniline oxidase activity, using only O\(_2\) as the oxidant. More recently, Nastri
et al (296, 297), D’Auria et al (298), & Lombardi et al (299) adopted
a minimalist approach for the design of small peptide-porphyrin conjugates named
mimochromes. Mimochromes are much simpler than the above systems and con-
tain a single heme sandwiched between two \( \alpha \)-helices. The basic structure of
mimochrome I is a deuteroporphyrin IX covalently linked to two nonapeptides via amide bonds between the propionic groups of the macrocycle and the Lys side chains. The prototype molecule, mimochrome I, is patterned after the \( \beta \)-chain of human deoxyhemoglobin (300) and contains a central His residue to ligate the heme iron. Hb Leu residues are also positioned at positions \( i-4 \) and \( i+4 \) relative to the His, to hydrophobically interact with the heme macrocycle. The spectroscopic and structural features of mimochrome I cobalt and iron complexes (296, 298) indicate that they bind heme iron in a low-spin bis-His–ligated complex as envisioned in the design. Interestingly, two diastereomeric complexes differing in the configuration about the metal ion are formed, and the two exchange-inert Co(III) complexes can be separated. The NMR structures of these isomers have been determined (Figure 12), providing the first example of the structure determination of a designed heme-peptide complex. As expected, in one isomer, the helices lie antiparallel to one another and parallel to the porphyrin plane.

Whereas mimochrome I required the addition of trifluoroethanol (TFE) for maximal helical content, more recent designs have focused on models with longer, more stable helices that show considerably greater helix content in water. These designs also adopt only one of the two possible enantiomers at the metal center. Thus, these simple, structurally defined systems provide an excellent opportunity for investigating how changes in the electrostatics, polarity, and solvent accessibility of the heme site affect the thermodynamics and rates of electron transfer. In addition, five-coordinate heme mimochromes have been prepared, providing an empty site available for catalyzing various oxidation reactions.

In an independent series of contributions, Benson and coworkers (301) have also prepared a series of very similar peptide-sandwiched mesoheme complexes, containing two identical \( \alpha \)-helices covalently attached to Fe-mesoporphyrin II and IX. Interestingly, a Trp residue located at position \( i+4 \) from the His residue is able to enhance the helical content of the complexes as assessed by CD spectroscopy, possibly owing to stabilizing edge-to-face aromatic interactions (302). Further, these workers have designed a series of disulfide bond-stabilized two-helix peptides, which noncovalently bind to heme to form a bis-His–ligated complex (303).

**Metal Ions in Peptide Self-Assembly Processes**

The introduction of coordinating groups into de novo designed peptides is quite useful for metal ion-assisted folding of peptides into a self-assembled architecture with a unique aggregation state. Ghadiri and coworkers (304, 305) used this approach, in the development of three- and four-helix-bundle metalloproteins, which undergo intermolecular self-assembly on formation of exchange-inert Ru(II) complexes with pyridyl or bipyridyl ligands. The ruthenium ion was chosen for its ability to form very stable and kinetically exchange-inert complexes. Sakamoto et al (305a) have recently shown that titration of heme can induce helical structure and self-assembly into a His-containing disulfide-linked parallel dimer in the presence of TFE. A similar peptide (305b) can be induced to
tetramerize upon Fe$^{3+}$-mesoporphyrin addition. These molecular architectures have also formed the basis for the design of protein mimics containing an organic electron donor, allowing flash-activatable electron transfer (306). Such bifunctional systems provide a class of molecules for testing the effect of conformation on electron transfer in proteins and in peptides.

OUTLOOK

It has now been ~10 years since the first serious attempts were made to design proteins from scratch. Since then, de novo protein design has proven to be a powerful method for addressing questions in protein stability and folding. The features required for stabilizing secondary structure have been elucidated, and the gross features that direct the collapse of the protein chain into a compact globular structure have been demonstrated. More recently, uniquely structured proteins have been designed and structurally characterized. There is no single formula for obtaining a uniquely structured state versus a molten globule. However, in several cases designed proteins were converted from molten-globulelike structures to native proteins by fine tuning the steric properties and the placement of Hp versus Hb amino acids in the sequence. In some cases, aromatic or H-bonded interactions have been demonstrated to be important for conformational specificity.
Automated methods for structure and sequence prediction are having an enormous impact on the field, allowing one to design uniquely packed structures and metal-binding sites within proteins. The extension of these methods to the design of ligand-binding sites and catalytic sites should be relatively facile. Further, genetic approaches and methods of combinatorial synthesis should allow one to more rapidly optimize initial designs. Thus, de novo protein design should be an increasingly useful approach—not only for testing the rules of protein folding and function but also for the design of novel proteins with practical applications towards catalysis, pharmaceuticals, and diagnostics.

ACKNOWLEDGMENTS

The authors thank GR Dieckmann and RB Hill for critical readings of the manuscript.

LITERATURE CITED


