De Novo Proteins from Combinatorial Libraries

David A. Moffet and Michael H. Hecht*

Department of Chemistry, Princeton University, Princeton, New Jersey 08544

Received February 6, 2001

Contents

I. Introduction 3191
II. Randomly Generated vs Rationally Designed Sequences 3191
III. Improving a Preexisting Scaffold 3192
IV. The Central Core of Designed Combinatorial Libraries: Bury the Grease 3192
V. Uniquely Folded Proteins from Binary Patterned Libraries 3194
VI. Designed Combinatorial Libraries of de Novo β-Sheet Proteins 3195
VII. Beyond Binary Patterning: Selecting the Best 3196
A. Optimizing the Hydrophobic Core 3196
B. Selecting Stable Interfaces 3197
VIII. Cofactor Binding and Function 3199
IX. Computer-Assisted Design: Screening Libraries Prior to Synthesis 3200
X. Concluding Remarks 3202
XI. References 3202

I. Introduction

Recent advances in genomic sequencing have paved the way for the new field of proteomics in which researchers can explore the diversity, interactions, structures, and functions of every protein found in nature. As studies of natural proteins advance, and we develop an understanding of the detailed functions of individual proteins, as well as the myriad interactions between different proteins in the biosphere, one is tempted to look beyond proteomics. Rather than limiting our studies to the set of proteins existent on earth, one is tempted to ask what structures and functions might be unobserved but nonetheless possible. Studies that go ‘beyond proteomics’ are motivated by a new question: Instead of asking “what exists” such studies ask “what is possible”.

What is possible? One way to answer this question is to construct and characterize large libraries of proteins de novo. Such collections can serve as a ‘parallel universe’ in which the newly prepared proteins can be compared to the evolutionarily selected proteins that currently exist.

The hypothetical library of all possible amino acid sequences is enormous. For instance, a hypothetical library of all 100 residue sequences constructed from the 20 naturally occurring amino acids would contain 20^{100} ($>10^{130}$) sequences. If one could synthesize one molecule of each of these 100-mers and place them in a box, the volume of the resulting box would be larger than Avogadro’s number of universes.[1] Clearly, the complete library of all possible sequences has never been sampled by nature nor could it be prepared de novo. Sequence space is simply too vast.

In addition to the vast *quantity* of possible sequences, it is also worth considering the *quality* of such sequences. While it has been shown that collapsed structures can occasionally be isolated from collections of random sequences,[2−7] the occurrence of well-folded water-soluble structures is likely to be exceedingly rare. Sequences capable of folding into structures that are truly protein-like, i.e., the ‘high quality sequences’, may represent only a small neighborhood amidst the vastness of sequence space. Both the size and quality of sequence space will influence choices that are made in considering the construction of combinatorial libraries of de novo protein-like structures.

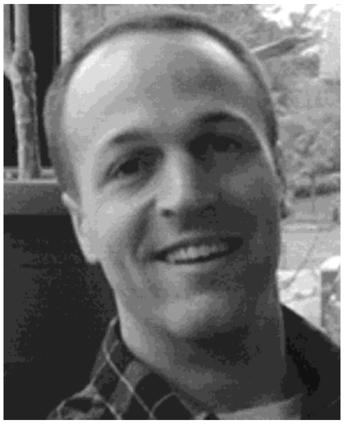
II. Randomly Generated vs Rationally Designed Sequences

Two global strategies for constructing new protein sequences are random sequence libraries and rational design. Both strategies have benefits and drawbacks associated with them. Due to the vast size of sequence space, strategies in which sequences are chosen at random will yield proteins with specifically desired structures or functions only at a very low frequency.

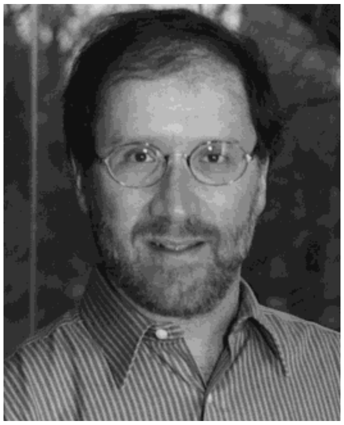
Rational design has shown great promise in creating novel proteins.[8−14] In this strategy, protein sequences are designed, residue by residue, to yield a sequence with a desired structure. This method has seen different levels of success in a wide variety of designs.[8−18] While rational design strategies can prove successful for designing proteins with desired structures, rational design typically does not explore extensive regions of sequence space and hence does not explore the question “what is possible?”. Also, while rational design can be useful for producing very good sequences, the question of whether a rationally designed sequence is in fact the best sequence for a desired trait is often left unanswered.

While both randomly generated and rationally designed sequences have appealing qualities, some of the most successful projects have blended aspects of both approaches. This blending typically results in the rational design of large libraries of related

10.1021/cr000051e CCC: $36.00
Published on Web 09/06/2001



David Moffet was born in New Cumberland, PA. He attended Shippensburg State University of Pennsylvania where he worked on inhibition studies of bovine cytochrome *c* oxidase with Dr. James Beres. He graduated from Shippensburg University in 1997 with his B.S. in Chemistry and a concentration in Biochemistry. Moffet is currently doing his graduate work at Princeton University under the guidance of Professor Michael Hecht. Moffet's research interests lie in the de novo design of functional and useful proteins. His favorite areas of work include the characterization of heme-binding de novo proteins and working with phage-displayed peptide libraries.



Professor Michael Hecht was born in Manhattan in New York City. He attended Cornell University, where he was introduced to the study of proteins through his undergraduate research in the laboratory of Professor Harold A. Scheraga. Hecht did his graduate work in the Department of Biology at MIT, where he received the first Ph.D. degree from the laboratory of Professor Robert T. Sauer. In his Ph.D. research, Hecht used mutagenesis to probe the structure and function of the λ repressor protein. After leaving MIT, Hecht spent a year traveling around the world. He then pursued postdoctoral studies under the guidance of Professors David and Jane Richardson in the Biochemistry Department at Duke University Medical School. With the Richardson's, he worked on the design of 'Felix', a four helix bundle protein, which was among the first proteins designed de novo. In 1990 Hecht joined the faculty at Princeton, where he is now an Associate Professor in the Chemistry Department with joint affiliations in both Molecular Biology and Materials Science. Professor Hecht's research at Princeton focuses on the design of de novo proteins from combinatorial libraries.

sequences, all targeted to have some desired trait. To be successful, the strategy must incorporate enough diversity to cover a significant part of sequence space while simultaneously incorporating enough rational design to limit exploration to those regions of sequence space most likely to yield sequences that possess the desired qualities. This review will focus on de novo proteins derived from large combinatorial libraries of sequences that have been guided by elements of rational design.

III. Improving a Preexisting Scaffold

Several techniques, including directed mutagenesis, protein engineering, directed evolution, in vitro recombination, and DNA shuffling, have become widely used for the modification of natural proteins.[19−23] These techniques are often used to introduce combinatorial diversity into preexisting natural proteins. Since the techniques are used on natural proteins, the resulting sequences are not truly de novo and will not be discussed in this review. For a recent review of several of these techniques, see Giver and Arnold (ref 24).

Another approach used to produce combinatorial diversity is through the construction of libraries of short peptides (typically less than 20 residues in length). This is usually done by solid-phase peptide synthesis or by phage display methods. These de novo sequences are certainly of interest to protein design. However, we will focus on sequences the size of natural globular protein domains. Good articles and reviews discussing libraries of small peptides may be found in refs 25−31, as well as in the article by Hoess in this issue.

IV. The Central Core of Designed Combinatorial Libraries: Bury the Grease

The underpinnings of many strategies used in protein design are derived from an examination of natural protein structures. When studying natural globular proteins, two dominant themes can be observed: (i) Natural proteins typically bury hydrophobic residues within the core of the protein while concurrently exposing hydrophillic residues to the solvent; (ii) Natural proteins contain an abundance of hydrogen-bonded secondary structure—α-helices and β-sheets. With both these features appearing almost ubiquitously in the well-folded structures of natural proteins, it seems reasonable to use these features as the cornerstones for designing libraries of novel proteins.

A method pioneered by our laboratory for designing libraries of novel proteins relies on 'binary patterning' of polar (P) and nonpolar (N) amino acids. Binary patterning incorporates polar and nonpolar amino acids in accordance with the structural periodicity of the desired secondary structure. Such patterning allows the formation of secondary structure while simultaneously enabling the burial of nonpolar amino acids. Binary patterning is used to create amphipathic segments of secondary structure, where one face contains only polar residues while the other face contains only nonpolar residues. Binary patterning exploits the periodicities inherent in secondary structures. α-Helices have a repeating periodicity of 3.6 residues per turn, while β strands have an alternating periodicity (up−down−up−down−etc.). To design an amphipathic α-helix with one polar face and one nonpolar face, a binary pattern of P−N−P−P−N−N−P−P−N would be used. To design an amphipathic β strand with one face entirely hydrophobic and one face entirely hydrophilic, an alternating pattern of P−N−P−N would be used (see Figure 1). Binary

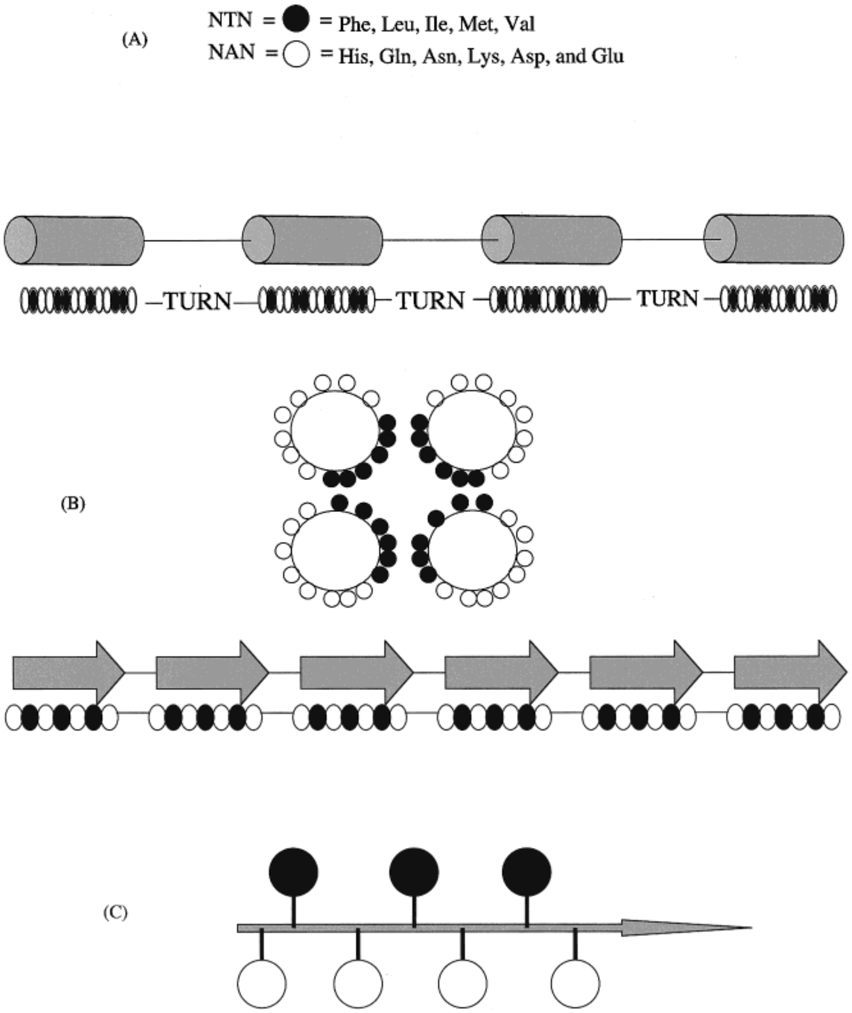

(A)
NTN = ● = Phe, Leu, Ile, Met, Val
NAN = ○ = His, Gln, Asn, Lys, Asp, and Glu
TURN
TURN
TURN
(B)
(C)

Figure 1. Designed combinatorial libraries based on binary patterning. (A) The degenerate DNA codons, along with the nonpolar (●) and polar (○) residues they encode. (B) Linear arrangement of polar and nonpolar residues within a combinatorial library of four-helix bundle proteins. Folding into a four-helix bundle would bury nonpolar residues in the core and expose polar residues to solvent.[34] (C) Linear arrangement of polar and nonpolar residues in a library of β-sheet structures. This pattern would cause one face of each strand to be polar and the opposite face to be nonpolar.[47]

patterning incorporates elements of rational design by explicitly specifying the polar/nonpolar ordering of the de novo sequences. At the same time, since the identities of polar and nonpolar residues are not specified explicitly, there is great potential for combinatorial diversity.

α-Helical structures are targeted most frequently in protein design. Indeed, the first de novo proteins were designed to fold into four-helix bundles.[32,33] These first helical proteins showed characteristic α-helical circular dichroism spectra but were not uniquely folded structures. Nonetheless, these early projects showed it was possible to design folded proteins ‘from scratch’.

The choice of helical proteins for initial forays into protein design is due, in part, to the intrinsic nature of α-helices. Helices form short-range, intramolecular hydrogen-bonding contacts with neighbors $i+4$ residues away in the primary sequence (see Figure 2). These are local interactions. On the other hand, β-sheet structures form hydrogen-bonding contacts with partners farther away in the linear chain. These nonlocal interactions can involve various possible partners at numerous loci within the primary sequence. Hence, β structures are typically more difficult to design.

The first reported combinatorial library of de novo proteins was based on a binary code design of amino acid sequences targeted to fold into four-helix bundles[34] (see Figure 1A and B). This library was prepared from a collection of synthetic genes expressed in bacteria. Combinatorial diversity was made possible by the organization of the genetic code. Residues designed to be nonpolar were encoded by the degenerate DNA codon NTN (where N is a mixture of all four nucleotide bases), which encodes Phe, Leu, Ile, Met, and Val. Conversely, residues designed to be polar were encoded by the degenerate DNA codon VAN (where V is a mixture of C, A, and G), which encodes His, Gln, Asn, Lys, Asp, and Glu. The designed sequences were 74 residues long with 24 combinatorially varied hydrophobic residues encoded by the degenerate codon NTN and 32 combinatorially varied hydrophillic residues encoded by the degenerate codon VAN. (The remaining 18 residues were designed to occur at helix ends and interhelical turns. These residues were held constant and were not combinatorially diverse.)

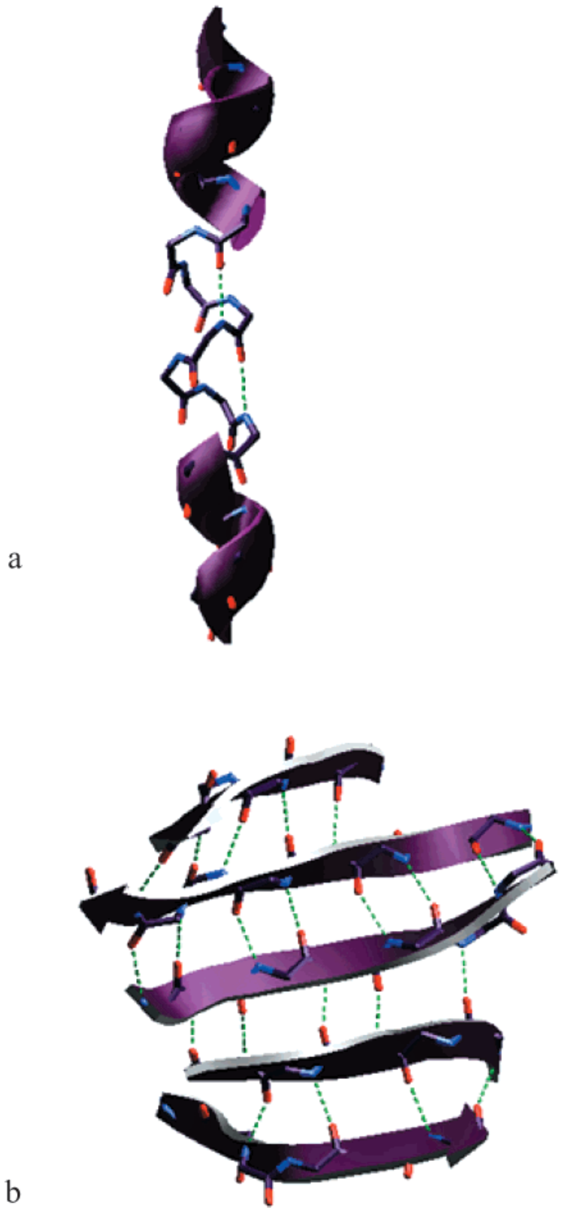


Figure 2. (A) Hydrogen-bonding interactions within a typical α-helical peptide. Hydrogen-bonding interactions occur between the backbone carbonyl of residue i and the backbone amide of residue i+4. (B) Hydrogen-bonding interactions in a β-sheet protein. Hydrogen bonding can occur between residues occurring at arbitrary distances from each other in the primary sequence.

The theoretical diversity of this library would be $5^{24} \times 6^{32} = 4.7 \times 10^{41}$ different sequences.[34] While this is an enormous number of possible sequences, it is drastically smaller than the theoretical diversity of a fully unconstrained library of 74-residue sequences ($20^{74} > 10^{96}$). Thus, the binary code strategy constrains the diversity of the sequences being examined. In return for this constraint, what is gained?

All the binary code sequences examined thus far (>50) formed water-soluble proteins that fold into α-helical structures. In contrast, sequences chosen randomly would not have yielded such results. By using elements of rational design to constrain combinatorial diversity, the binary code strategy vastly increases the likelihood of finding the 'good' sequences.

Over 50 proteins from this initial combinatorial library have been purified and characterized. All the sequences examined thus far showed circular dichroism spectra of α-helical proteins, with negative minima at 208 and 222 nm. The collection also yielded some proteins with nativelike properties, such as NMR chemical shift dispersion,[35,36] cooperative chemical and temperature denaturations,[37] and slow H/D exchange rates.[38]

While some of the sequences within this designed library showed nativelike characteristics, many were more similar to molten globules. The molten globule state typically has considerable secondary structure but lacks a single, uniquely folded, ground-state structure. Instead, it interconverts between many related structures. Molten globules typically show little cooperativity in thermal denaturations, poor chemical shift dispersion in NMR studies, and fast H/D exchange rates. Most of the sequences from the originally designed 74-residue binary code library displayed at least some characteristics of molten globules.

V. Uniquely Folded Proteins from Binary Patterned Libraries

The binary code strategy for protein design successfully produced a large collection of de novo sequences that folded into soluble α-helical structures. However, the goal of producing large numbers of nativelike proteins was not fully achieved. Two possible reasons can be proposed to explain the abundance of molten globule-like proteins in this library: (i) The binary code strategy might not be sufficient to yield nativelike structures. Because of its combinatorial underpinnings, the binary code strategy cannot explicitly specify the core packing of a protein. One might expect that to achieve well-packed structures, hydrophobic cores must be designed explicitly, and without such residue-by-residue design, molten globules would be the default structures. (ii) Alternatively, one might argue that specific tertiary interactions need not be designed a priori; in the context of a stable structural scaffold, unique packing will occur nonetheless a posteriori. Support for this hypothesis can be drawn from numerous experiments demonstrating that natural protein structures can tolerate the simultaneous substitution of many (in some cases all) of their hydrophobic core residues.[39−43] For example, Fersht and co-workers replaced all the core residues of barnase with random hydrophobic side chains and found that a high proportion of these mutants (23%) retained activity.[39] In another example, Matthews and co-workers devised an explicit test of the "jigsaw puzzle model for protein folding" by replacing up to 10 residues in the core of T4 lysozyme with methionines.[40] They found that, in contrast to the predictions of the jigsaw puzzle model, the multiply substituted proteins were active and cooperatively folded.[40] These and other results[44] suggest that in the context of an appropriately designed structural scaffold, many different amino acid sequences can yield well-folded proteins. Clearly, however, for the binary code strategy to 'live up to its potential', it must be applied to

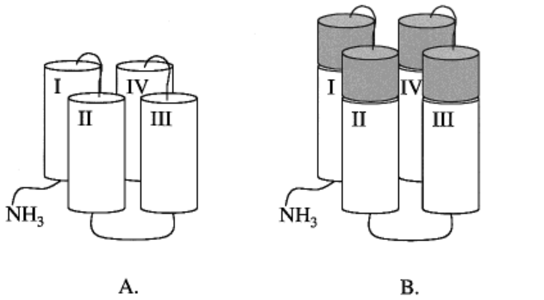


Figure 3. (A) Schematic representation of the original binary code library of Kamtekar et al.[34] Cylinders represent α-helices. Sequences in the original library were 74 residues long. (B) Second-generation library.[46] The darkened regions show elongation of the α-helices via combinatorial addition of residues in accordance with the binary pattern for amphiphilic α-helices.

a scaffold that is more robust than the original 74-residue design.

The template that served as the basis for the original binary code library may have suffered from significant shortcomings (both literally and figuratively). Most natural four-helix bundles are composed of >100 amino acids, with individual helices typically longer than 20 residues.[45] However, the template for the original library of four-helix bundle proteins was designed to be 74 amino acids in length, with each helix consisting of merely 14 residues.

To investigate the potential of the binary code strategy to encode collections of nativelike proteins, a second-generation library of binary-coded proteins was prepared[46] (Figure 3). This new library used protein #86, a preexisting sequence from the original 74-residue library, as the starting material. Minor changes were initially introduced to protein #86, including alterations to the turn regions and the addition of a tyrosine chromophore to aid in quantifying protein concentration. The major change to protein #86 was the addition of six combinatorially diverse residues to each of the four helices. These 24 additional helix-lengthening residues continued to follow the binary patterning with polar residues designed to be solvent exposed and nonpolar residues designed to be buried. In all, the second-generation proteins consist of 102 amino acids, comparable in size to natural four-helix bundle proteins.

From this second-generation library, five sequences were arbitrarily chosen for detailed analysis. All five sequences were found to be monomeric by size exclusion chromatography and highly helical by circular dichroism spectroscopy.[46] Chemical denaturation studies showed that the free energies ($\Delta G°$) stabilizing the second-generation proteins were typically 2- to 3-fold larger than the corresponding ΔG for the 'parental' protein #86. NMR spectroscopy was used to probe the structures of the proteins. In comparison to protein #86, the second-generation proteins yielded NMR spectra with enhanced chemical shift dispersion, sharper signals, and abundant NOE cross-peaks, all indicative of well-folded nativelike proteins. The success of this new library indicates that the binary code strategy is capable of yielding de novo proteins with structures resembling those of well-folded natural proteins.

VI. Designed Combinatorial Libraries of de Novo β-Sheet Proteins

The binary code strategy is not limited to the design of α-helical proteins. Libraries of de novo β-sheet proteins have also been constructed. Amphiphilic β strands have an alternating periodicity of P−N−P−N−P−N. On the basis of this periodicity, a combinatorial library of synthetic genes was created to encode β-sheet proteins.[47] Polar residues were designed to comprise one face of the β strands, with nonpolar residues on the opposing face (Figure 1C). The sequences in the library were designed to have six β strands with each strand having the binary periodicity P−N−P−N−P−N−P. Proteins from this library were expressed from a collection of synthetic genes cloned in *E. coli*. The proteins were expressed as inclusion bodies, solubilized in 6 M urea, and exchanged into phosphate buffer. All the proteins studied in this collection showed typical β-sheet secondary structure, having CD spectra with the characteristic minimum at 217 nm. The β-sheet proteins self-assemble into large fibrils visible by electron microscopy and atomic force microscopy.[47] The de novo fibrils resemble the amyloid fibrils found in several neurodegenerative diseases. Like natural amyloid, the de novo fibrils are composed of β-sheet secondary structure and bind the diagnostic dye, Congo red.

The de novo proteins isolated from the β-sheet binary code library have significantly different properties from those isolated from the α-helical library. Why did the first library yield α-helical structures that fold intramolecularly into small globular domains while the second library yielded β-strands that assemble intermolecularly into large aggregates resembling amyloid? What accounts for the differences in physical properties between these two libraries? Both libraries are composed of the same binary-coded amino acids. Hence, the different properties are not due to differences in amino acid composition. The differences are also not due to differences in sequence length. Several libraries of different lengths have been examined for both the α-helical pattern and the β-sheet pattern. Irrespective of the length, sequences with P−N−P−P−N−N−P periodicity form soluble proteins with α-helical secondary structure while sequences with P−N−P−N−P−N−P periodicity (examined under the same experimental conditions) form β-sheet secondary structures that self-assemble into amyloid-like fibrils. The key difference between these two libraries of sequences is the binary patterning itself.

After discovering that alternating polar/nonpolar patterns predispose de novo sequences to form amyloid-like structures, we became curious about the occurrence of such patterns in the sequences of natural proteins.[48] Analysis of a database of 250514 protein sequences for all possible binary patterns of polar and nonpolar residues revealed that alternating patterns (...P−N−P−N−P...) occur significantly less frequently than other patterns with the same composition.[48] This under-representation was apparent for all 'windows' from 5 to 10 residues in length. The statistical under-representation of alternating binary patterns in natural proteins, along with the observation that such patterns promote amyloid-like structures in de novo proteins, suggests that this alternating pattern is inherently amyloidogenic and disfavored by evolutionary selection.

VII. Beyond Binary Patterning: Selecting the Best

A. Optimizing the Hydrophobic Core

Binary patterning of hydrophilic and hydrophobic amino acids provides a foundation for the design of combinatorial libraries of de novo proteins. However, this strategy alone does not attempt to optimize amino acid interactions within the de novo proteins. To find optimal packing interactions, screens must be devised to sort through the various combinatorial possibilities.

Screens and selections in vivo are based on biological phenotypes. However, proteins designed de novo do not have biological phenotypes. Moreover, even for natural proteins the occurrence of optimal packing interactions is only loosely correlated with an observable phenotype. Therefore, screening combinatorial libraries of de novo proteins for well-packed hydrophobic cores is a challenging task.

We have attempted to meet this challenge by devising new screens to facilitate the hunt for well-packed structures among combinatorial libraries expressed in bacteria. Because the de novo proteins do not have observable biological phenotypes, these screens must be based on biophysical properties assayed in vitro. Biophysical properties, however, have historically been measured using only highly purified samples and thus typically have not been suitable for high-throughput screening. To facilitate rapid screening, it is essential that methods be effective without the necessity for laborious and time-consuming protein purifications. Therefore, we developed methods to enable the isolation of semi-pure protein samples from bacterial cultures using only temperature shifts and centrifugation.[49] The simplicity of these methods facilitates preparation of many protein samples in parallel. Although the resulting samples are not 100% pure, they are quite suitable for screening for nativelike biophysical properties.

We developed two novel screens to search for well-packed structures amidst libraries of de novo proteins. Both screens assess biophysical properties that are hallmarks of nativelike structures but absent in fluctuating molten globules. The first screen uses one-dimensional ^{1}H NMR spectroscopy to probe for chemical-shift dispersion and sharp peaks among collections of semi-pure samples of de novo proteins expressed in *Escherichia coli*.[35] The second screen uses electrospray mass spectrometry to monitor the hydrogen−deuterium (H−D) exchange kinetics in expression libraries of de novo proteins.[38] Since protection of amide protons from exchange depends on the existence of a stably folded and nonfluctuating structure, this screen can be used to identify nativelike proteins from combinatorial libraries containing both nativelike and molten globule-like structures. Both screens can be applied to semi-pure samples that do not require extensive protein purification, and thus can be used for rapid screening of large combinatorial libraries.

Biophysical methods have also been used by Dutton and co-workers to search through regions of sequence space for well-folded de novo proteins. They began with a rationally designed sequence called $[H10H24]_2$ as their prototype protein maquette.[50] This de novo protein is composed of four identical α-helices, each containing 31 amino acids. The 31-mers have N-terminal cysteines, which when oxidized yield [H10H24], a disulfide-linked 62-residue peptide. These 62-mers then dimerize noncovalently to form $[H10H24]_2$, a four-helix bundle protein. $[H10H24]_2$ is molten globule-like and displays poor NMR chemical shift dispersion.

Dutton and co-workers used iterative protein redesign, followed by NMR screening, to convert the original $[H10H24]_2$ sequence into a well-folded nativelike structure.[51] The original 31-residue α-helix in $[H10H24]_2$ had leucine residues in all of its three core heptad 'd' positions. The sequence was renamed protein "LLL" to describe this composition. In their iterative redesign, Dutton and co-workers modified these three core positions by introducing point mutations from among the nonpolar amino acids, isoleucine, valine, and phenylalanine. With four residues allowed at three sequence positions, a library of 64 different sequences was possible.

The first round of redesign produced the nine possible single-point mutations of LLL (LLI, LLV, LLF, LIL, LVL, LFL, ILL, VLL, and FLL). Each of these nine sequences was analyzed by NMR, and the resulting chemical shift dispersion for each sequence was compared to LLL. Three sequences (ILL, VLL, and LFL) showed enhanced chemical shift dispersion in their NMR spectra and were therefore selected for a second round of redesign. Single-point mutations were made in each of these three sequences. Five of the resulting doubly mutated sequences (IIL, IVL, IFL, VIL, and VVL) showed improved NMR spectra.

A third round of redesign was carried out beginning with protein IFL. This protein was mutated to IFI, IFV, and IFF. All three of these third-generation sequences showed a complete loss of conformational specificity and were less stable than IFL.[51] Iterative redesign and NMR screening had selected IFL as the best sequence from the library. The structure of protein IFL was then determined by NMR spectroscopy and shown to be predominantly nativelike.[52] These experiments demonstrate that without explicitly searching through all combinatorial possibilities, iterative redesign of a protein core can be accomplished using biophysical screens to optimize sequences.

Case and McLendon used a very different approach to search sequence space for optimal packing arrangements.[53] They constructed a peptide library that relied upon self-assembly processes to find the best packed three-helix bundle sequences. Three different 20-residue peptides were synthesized. Each peptide possessed the same water-exposed hydrophilic residues, but the residues designed to pack within the core were varied. Peptide sequence αpL contained four Leu residues targeted to pack within the core,

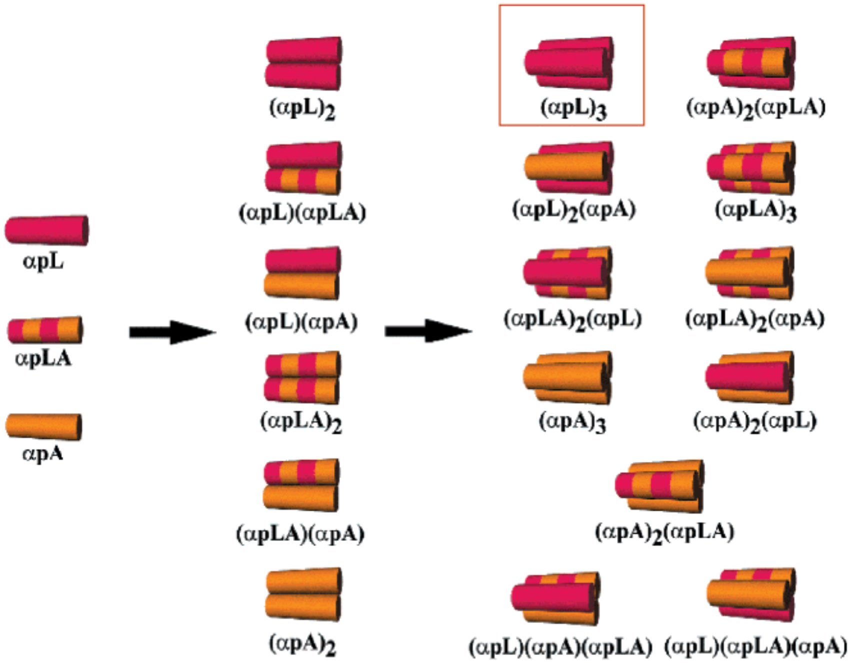

$(\alpha pL)_2$
$(\alpha pL)_3$
$(\alpha pA)_2(\alpha pLA)$
$(\alpha pL)(\alpha pLA)$
$(\alpha pL)_2(\alpha pA)$
$(\alpha pLA)_3$
αpL
αpLA
$(\alpha pL)(\alpha pA)$
$(\alpha pLA)_2(\alpha pL)$
$(\alpha pLA)_2(\alpha pA)$
$(\alpha pLA)_2$
$(\alpha pA)_3$
$(\alpha pA)_2(\alpha pL)$
αpA
$(\alpha pLA)(\alpha pA)$
$(\alpha pA)_2(\alpha pLA)$
$(\alpha pA)_2$
$(\alpha pL)(\alpha pA)(\alpha pLA)$
$(\alpha pL)(\alpha pLA)(\alpha pA)$

Figure 4. Library of three-helix bundles designed by Case and McLendon.[53] Cylinders represent the three peptides: αpL, which has Leu residues in each of the four hydrophobic core positions; αpLA, which alternates Leu and Ala at the core positions; and αpA, which has Ala residues in each core position. Each peptide is attached at its N-terminus to 2,2′-bipyridyl-5-carboxylic amide, which is capable of coordinating to Fe^{2+} or Ru^{2+} in a ratio of 3 peptides to 1 metal. With a library of three different peptides, 11 different three-helix bundle proteins are possible. Introduction of Fe^{2+} results in the formation of an exchange-labile population of three-helix bundle proteins. The boxed protein, $(\alpha pL)_3$, was found to be the most thermodynamically stable three-helix bundle and hence found in the highest concentration. (Figure kindly provided by M. Case and G. McLendon).

sequence αpA contained four Ala residues, and sequence αpLA alternated Ala, Leu, Ala, Leu within the core (see Figure 4). Each peptide was attached at the N-terminal to Bpy (2,2′-bipyridyl-5-carboxylic amide), which coordinates to metals such as ruthenium(II) and iron(II) in a stoichiometry of three peptides to one metal. The coordination of Bpy to Fe(II) is not strong and forms an exchange-labile species.[53] Two scenarios are possible with an exchange-labile library: (i) Mixing of the three peptides with Fe(II) might yield an equal concentration of all of the 11 possible three helix bundle complexes. (ii) Alternatively, if the packing of the helices is thermodynamically coupled to metal binding, the final equilibrium mixture will favor the most stable protein complexes and these will be present in the highest concentrations.

Case and McLendon found that packing and metal binding are thermodynamically coupled.[53] In the case where 2/9 equiv of Fe(II) was added to 1 equiv of protein, the homotrimer, $(\alpha pL)_3$, was found in higher abundance than any of the other possible three-helix bundles. Denaturation studies of each of the 11 possible trimeric peptides (bound covalently to Ru(II) which is not exchange labile) showed that the αpL trimers were in fact the most stable of the 11 possible three-helix bundles. These experiments showed that self-assembly of a virtual library of peptide sequences can guide the design and selection of core residues in de novo sequences. While this example was a proof of principle experiment, it could be used with much larger libraries of combinatorially diverse peptide sequences.

B. Selecting Stable Interfaces

While core packing is considered a key ingredient for the formation of stable protein structures, the interfaces between units of secondary structure and/or in protein oligermization are also of great importance. Arndt et al. described a “protein-fragment complementation” selection system to identify stable heterospecific interfaces between the α-helices of coiled coil proteins.[54] They constructed two combinatorial libraries of sequences: Library A was based on the coiled coil region of the proto-oncogene c-Jun, and library B was based on the coiled coil region of the proto-oncogene c-Fos.[54] Both libraries held the solvent-exposed b, f, and c positions constant (see Figure 5). The core position, ‘d’, was held constant as leucine for all sequences, and the core position, ‘a’ was valine for all positions except for a 50% mixture of valine and asparagine in the central ‘a’ position. Combinatorial diversity was introduced at the e and g interfacial positions of the coiled coil proteins of both libraries. These libraries were designed to have an equal mixture of glutamine, glutamate, lysine, and arginine in the e and g interface positions (see Figure 5).

Genetically fused to each helical sequence in the library was one of two pieces of the enzyme murine dihydrofolate reductase (mDHFR). This enzyme is necessary for the survival of bacteria growing on minimal media plates containing trimethoprim (a selective inhibitor of prokaryotic DHFR). For mDHFR to be active in this system the two segments of mDHFR must be brought together by “cognate in-

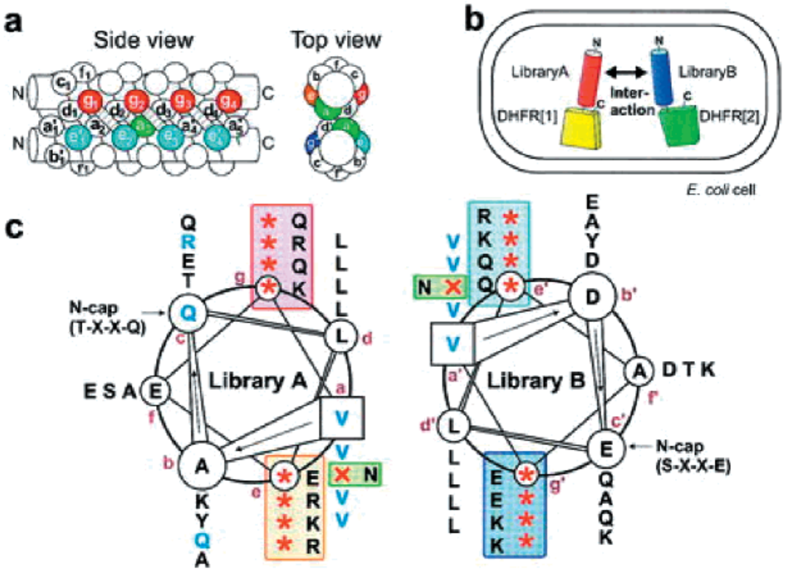


Figure 5. (a) Schematic representation of a parallel dimeric coiled coil. Side view: the helical backbones are represented by cylinders, the side chains by knobs. The path of the polypeptide chain is indicated by a line wrapped around the cylinders. For simplicity, supercoiling of the helices is not shown. Residues at positions a and d make up the hydrophobic interface, and residues at positions e and g pack against the hydrophobic core. They can participate in interhelical electrostatic interactions between residue i (g position) of one helix and residue i′+5 of the other helix (e′ position, belonging to the next heptad), as indicated by the hatched bars. Top view: arrangement of the heptad positions. (b) Schematic representation of the protein-fragment complementation assay. Each library is genetically fused to one of two mDHFR fragments. Interaction between the two library peptides restores enzyme activity, which is crucial for cell survival under selective conditions. (c) Overview of the library design depicted as α-helical wheel plot from the N to the C terminus (inside to outside). The black residues correspond to the original residues from c-Jun (library A) and c-Fos (library B). Changes introduced by the design are in blue. The randomized positions are in red (*, equimolar mixture of Q, E, K, R; ×, equimolar mixture of N, V) and are boxed with the same colors as used in part a. The selected residues from the best sequences, named WinZip-A1B1, are next to the randomized positions in the respective boxes. (Reprinted with permission from ref 54. Copyright 2000 Academic Press.)

teraction between members of both libraries".[54] Plasmids containing genes from library A and plasmids containing genes from library B were cotransformed into *E. coli*. The bacteria were then plated on selective media. Only those bacteria whose protein products from libraries A and B interacted to form coiled coil dimers (and therefore reassembled the two parts of mDHFR) would survive. By altering the stringency of the selections, the authors were able to select more stable dimerization interactions, ultimately selecting one very stable coiled coil protein, WinZip-A1B1. This heterodimeric coiled coil protein was found to have a K_d of 24 nM and a t_m of 55 °C.[54] For comparison, the natural homodimeric coiled coil APC (adenomatus polyposis coli) protein has a t_m of 46 °C despite the fact that it is longer than WinZip-A1B1.[55]

A method for selecting both an optimized hydrophobic core and a stable interprotein interface was investigated by Braisted and Wells.[56] They used combinatorial libraries displayed on phage to minimize a natural three-helix bundle protein, the Z-domain of protein A, into a stable two-helix protein.[56] To accomplish this goal, both the hydrophobic core of the new two-helix bundle and its interface with an antibody had to be optimized and selected.

The Z-domain of protein A is a stable, 59-residue three-helix bundle that binds the F_c portion of IgG.[57,58] Inspection of the X-ray[59] and NMR[60] structures show that only residues from helix 1 and helix 2 of protein A contact IgG. Helix 3 makes no contacts to IgG, yet helix 3 is necessary to stabilize the structures of helices 1 and 2. Upon removal of helix 3 of protein A, helices 1 and 2 become unstructured, thereby decreasing the binding affinity to IgG by $>10^5$-fold.[61] The goal of the project initiated by Braisted and Wells[56] was to find a sequence that would yield a two-helix bundle that would both retain stability and bind to IgG.

To isolate a stable two-helix bundle the authors used an iterative approach. Rather than randomizing all the residues in the sequence simultaneously (which would yield a theoretical library far surpassing the size of the phage display library), the protein was randomized in three separate steps corresponding to three structural regions. The first step involved removing helix 3 and preparing a library of "exoface" mutants capable of binding IgG. (Braisted and Wells defined the exoface as the residues of helices 1 and 2 that form the hydrophobic core with helix 3 in the intact Z-domain).[56] The exoface library had four residues randomized among all possible 20 naturally occurring amino acids, yielding a theoretical library of 20^4. In the wild-type domain, these residues are nonpolar and make contact with helix 3 of the original protein A. The library of exoface mutants was selected for IgG binding, and two mutations were found that rendered the sequence capable of binding IgG. The other two residues were maintained as wild type. The two mutations, Leu20Asp and Phe31Lys, placed polar residues on the new solvent-exposed surface of the protein.

The second iterative step took the best sequence from the exoface library as the starting point for production of a second library. This library, the "intraface" library, had all five of the nonpolar residues between helices 1 and 2 randomized, generating a theoretical library of 20^5. (The intraface was defined as the hydrophobic core region between helix 1 and helix 2 of the intact Z-domain).[56] This library was selected for IgG binding. Three non-wild-type residues were selected, Ala13Arg, Ile17Ala, and Leu35Ala, with the other two residues retained as wild type.

The third step took the best sequence from the exoface/intraface library and prepared five "interface" libraries. (The interface was defined as the region of contact between the Z-domain and the F_c portion of IgG).[56] Each interface library randomized 4 of the 19 residues of protein A known to contact IgG. The majority of the residues in these libraries were conserved. The newly discovered two-helix variants were examined in binding studies. The best variants, which included the exoface/intraface mutations along with seven or eight additional interface mutations produced stable two-helix sequences with only a 10-fold loss in IgG binding ability. The new sequence, Z38, was only 38 residues in length, cut down from the natural three-helix 59-residue protein. Of those 38 residues, 13 were altered to yield a new protein sequence capable of binding IgG.

Wells and co-workers solved the NMR structure of sequence Z38 in solution.[62] The protein folds into a two-helix structure "essentially indistinguishable" from helices 1 and 2 of the three-helix bundle of the wild-type domain of protein A.[62] However, Z38 was only marginally stable.[62] Using the NMR structure as a guide, two residues of Z38 were changed to cysteines, with the intent of incorporating a stabilizing disulfide bond. Also, four residues not involved with the binding of IgG were removed. This final, disulfide containing 34-residue sequence, named Z34C, had greater thermal stability and a 9-fold improvement in binding affinity relative to Z38. Moreover, Z34C has approximately the same binding affinity as the natural three-helix protein. These experiments demonstrate how a combinatorial approach, coupled with rational design and iterative selections, can be used to produce new sequences with desired functions.

VIII. Cofactor Binding and Function

Among the most important goals of protein design are the incorporation of binding and catalysis into de novo amino acid sequences. Nature uses two general strategies to produce functional proteins. In the 'purist' strategy, nature uses only the 20 amino acids (and the polypeptide backbone) to generate binding and catalytic sites. In the second strategy, nature 'cheats' by relying on nonprotein cofactors to accomplish catalysis. Cofactors, which range in size and complexity from a single zinc atom to a large heme macrocycle, can be thought of as 'preorganized activity modules' capable of performing a variety of chemical reactions. Bound cofactors present the opportunity for a range of activities that may have been difficult or impossible to achieve using the polypeptide sequence alone. It has been estimated that over one-half of naturally occurring enzymes harbor a metal and/or other cofactor.[63]

Nature's success in using cofactors suggests that these preorganized activity modules might also be useful in generating de novo enzymes. Rational design of binding sites has been attempted for both small and large cofactors.[9,64−72] In some cases, achieving the required binding geometry is quite difficult. For example, the rational design of a type II copper site into a preexisting protein scaffold required great precision and the absence of competing reactions.[73] In other cases, however, cofactor binding is relatively easy to achieve. For example, heme binding has been accomplished in several different design contexts[50,74−75] and has even been observed for the backward or 'retro' version[76] of a designed sequence.

Both the 'difficult' and the 'easy' cases of cofactor binding are attractive targets for the combinatorial approach. Even for difficult cases, the sequence diversity of combinatorial libraries coupled with the use of powerful selections or screens can enhance the likelihood of finding sequences capable of ligand binding.

Initial work using combinatorial methods to isolate cofactor binding proteins has focused primarily on the binding of heme by α-helical proteins. Haehnel and co-workers described a system for screening libraries of designed four-helix bundles in search of heme-binding and heme-based functionality.[77] They constructed a cyclic decapeptide scaffold upon which α-helices were covalently attached (see Figure 6). Two different libraries of helices were designed and synthesized. Proteins in library A_i were 15 residues long, with five residues targeted to pack within the core. These core residues were combinatorially randomized among Gly, Ala, Val, Leu, Ile, Phe, Tyr, and Gln. The second library, B_j, contained peptides 16 residues long with the four core residues comprised of the same combinatorially randomized amino acids used in A_i. However, the fifth core residue in the B_j library was histidine, which was included to ligate heme.

The four-helix bundles were prepared by attaching two identical helices from the A_i library and two identical helices from the B_j libraries to a decapeptide scaffold. To ensure an antiparallel configuration, the A_i library proteins were attached to the decapeptide scaffold via an N-terminal linkage while the B_j library proteins were attached via a C-terminal linkage. The resulting scaffold-bound proteins were targeted to fold into A_2B_2 antiparrallel four-helix bundles capable of bis-histidine heme binding.

To screen for heme binding, the protein solutions were spotted onto a cellulose membrane and the spots were incubated with a hemin solution. The immobilized spots were analyzed by UV−vis spectra to determine the extent of heme binding. Midpoint redox potentials were then estimated by viewing the UV−vis spectra of samples in their oxidized form and reduced form and at a potential of −95 mV. Midpoint potentials for nearly 400 samples were extrapolated using those three spectra. The ranges of potentials fell between −90 and −150 mV (versus the standard hydrogen electrode).[77] This strategy of manufacturing scaffold-assisted heme proteins can be used for

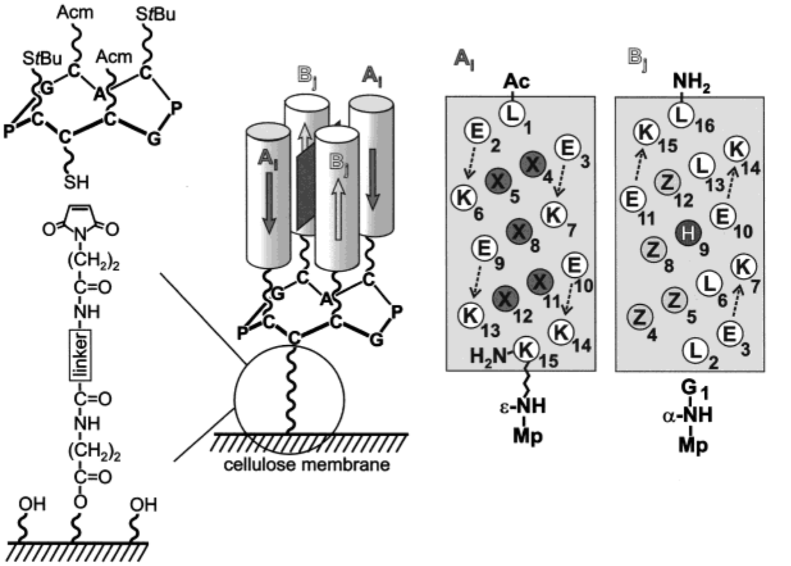


Figure 6. Stepwise assembly of the cellulose-bound library of synthetic heme proteins. Varied residues in peptide library A_i are labeled as X_k, and those in peptide library B_j are labeled as Z_l. H_9 is the heme-ligating histidine in library B_j. (Reprinted with permission from ref 77. Copyright 2000 WILEY−VCH.)

constructing many different de novo sequences in parallel. The immobilization of these proteins onto addressable spots facilitates rapid screening for redox potentials and could also be used for large-scale screening for numerous heme-based functionalities. Moreover, it can also be used to screen for the binding of other cofactors, as demonstrated by a recent study in which Haehnel and co-workers used their method to screen for copper binding in combinatorial libraries of immobilized de novo four-helix bundles.[78]

While the libraries of Haehnel and co-workers were explicitly designed to bind heme (or copper in their more recent work), work in our lab has shown that it is possible to isolate de novo heme proteins even from libraries that were not explicitly designed to bind heme. The original binary code α-helical library included histidine and methionine in the combinatorial mix.[34] These residues are frequently found as heme ligands in natural proteins. Although heme binding was not a consideration in our original binary code strategy, the incorporation of His and Met residues presented the possibility for yielding a collection of de novo heme proteins. Screening the original library of α-helical sequences showed, surprisingly, that approximately one-half bound heme.[79] These hemeproteins yielded bright red solutions with UV−vis and resonance Raman spectra similar to those of natural heme proteins.[79]

The availability of this binary patterned library of de novo heme proteins provided for the possibility of heme-based functionality. The novel heme proteins were assessed both for their function as small molecule binders (analogous to myoglobin or hemoglobin) and for their ability to carry out redox catalysis. The binding experiments focused on the kinetics and affinity of carbon monoxide binding. The de novo heme proteins were found to bind CO with kinetic rates of association and dissociation similar to those of natural heme proteins.[80] This indicated that the heme groups were partially buried within the core of the protein, rather than remaining solvent exposed. Resonance Raman spectroscopy of the CO adducts of the novel heme proteins supported these findings.[80]

To evaluate the catalytic potential of the binary code heme proteins, the collection was screened for peroxidase activity. Several proteins exhibited catalytic rates nearing those of natural peroxidases.[81] The best peroxidase isolated from this collection had a catalytic turnover rate only 3.5 times slower than that of horseradish peroxidase.[81]

When comparing the binary code proteins to natural or synthetic peroxidases, it should be emphasized that the binary code peroxidases were *not* subjected to genetic selections for heme binding or peroxidase activity. Moreover, they were *not* explicitly designed to bind heme. They were isolated from a library of sequences designed by binary patterning of polar and nonpolar amino acids to fold into α-helical bundles. Among this unselected collection of binary code proteins, (i) all of the purified proteins form α-helical structures,[34,37] (ii) approximately one-half bind heme,[79] (iii) several function as peroxidases,[81] and (iv) at least one protein exhibits a rapid catalytic turnover.[81]

The organization of the genetic code suggests a role for polar/nonpolar patterning in the evolution of protein structure and function. Our earlier findings with libraries of de novo sequences showed that binary patterning plays a key role in dictating protein structure.[34,36,47] Our recent finding that several proteins from a small sampling of a binary code library bind heme and accomplish catalysis suggests that binary patterning coupled with binding to preorganized activity modules (heme or other cofactors) may have provided a facile route toward the evolution of functional enzymes.

IX. Computer-Assisted Design: Screening Libraries Prior to Synthesis

Two main challenges for combinatorial protein design are (i) choosing which regions of sequence space are most likely to yield productive sequences and (ii) screening the resulting libraries for proteins with desired traits. An increasingly powerful method for meeting both of these challenges involves computer-assisted library design. In essence this approach aims to use modeling and prediction to sift through a huge number of sequences prior to any wet lab work. Ideally, the sequences 'most likely to succeed' can be identified a priori, thereby guiding library design. In the best case scenario, not only are the 'most likely to succeed' sequences identified, but moreover the absolute winner (e.g., most thermostable or most active) can be identified in silico prior to any synthetic laboratory work.

Reviews of algorithms for computational design can be found in refs 82−85. Here we will concentrate not on the algorithms themselves, but on their applications as presynthesis screens of combinatorial diversity.

Considerable effort has focused on the computationally assisted redesign of natural proteins—usually with the goal of repacking the core residues. One of the earliest attempts to repack the hydrophobic core of a natural protein was carried out by Desjarlais and Handel, who used two programs to repack the core of phage 434 Cro protein.[86] The first program produced a library of rotomer structures for the amino acid side chains. The second program used this rotamer library as input to screen through the combinatorial possibilities for low-energy packing sequences. Guided by their computational results, Desjarlais and Handel constructed stable versions of 434 Cro in which up to eight core residues were mutated. The algorithm found one sequence, D5, which repacked the core with five mutations and had a higher T_m than the wild-type protein.[86] As a control, the authors showed that random mutations generated unfolded sequences. These experiments demonstrate the power of computational methods to analyze sequence space and select good sequences prior to constructing new proteins.

An alternative algorithm, entitled CORE, was developed by Farid and co-workers to repack the buried residues of native proteins by computationally searching through sequence space prior to protein synthesis.[87,88] In their method, residues found to be less than 10% solvent exposed in the wild-type structure were considered core residues and could be altered while all other residues were left unchanged. Overall, CORE aimed to accomplish three goals: (i) To ensure there were no 'bumps' or steric hindrances of residues with the backbone structure; (ii) To maximize the change in heat capacity (ΔC_p) between the folded protein and the unfolded state; (iii) To minimize the change in conformational entropy associated with folding (ΔS_{conf}).[87]

CORE was used to repack the interior of four natural proteins with the goal of producing hyperthermophilic variants.[89] Jiang et al. used CORE to repack the interior of the B1 domain of protein G (Gβ1).[89] Their analysis yielded hundreds of sequences expected to show greater thermostability than wild-type Gβ1. One such sequence, ranked sixth by Farid and co-workers, was the sequence α90, first discovered by Dahiyat and Mayo[90] and experimentally determined to be hyperthermostable relative to the wild-type sequence.

Farid and co-workers also used CORE to repack the Cro protein from bacteriophage 434.[89] Their in silico search through sequence space led to several sequences expected to possess higher thermal stability than the wild-type protein. The only sequence selected by both CORE and the programs of Desjarlais and Handel was sequence D5. This was also the only sequence experimentally determined by Desjarlais and Handel to have a higher T_m than the wild-type protein (as discussed above and in refs 86 and 89). Moreover, CORE did not suggest the other, less thermostable, sequences examined by Desjarlais and Handel.

CORE has also been used by Farid and co-workers for the design of a de novo hyperthermostable four helix bundle.[88] Their novel protein was composed of two peptide chains held together by a disulfide bond between the C-terminal cysteine from chain A and the N-terminal cysteine of chain B. This protein unit was designed to dimerize into a homodimeric four-helix bundle protein. Solvent-exposed residues were Lys and Glu, which were positioned to provide favorable electrostatic interactions. The core residues were then selected by the CORE packing program from among the various combinatorial possibilities. The structure of the resulting protein has not been determined; however, chemical and thermal denaturation data show this protein, HYP-1, is very stable. At pH 7, the protein was estimated to be only 10% denatured at 98 °C. The chemical denaturation data revealed that the free energy of unfolding of the dimer was 13 kcal per mol.[88]

The first successful fully automated computational sequence selection was reported by Dahiyat and Mayo.[91] These authors took the known backbone structure of the zinc finger domain of Zif268 and used an algorithm based upon the dead end elimination theorem[92.93] to search for the lowest energy arrangement of residues consistent with this backbone structure. For the 28 residue sequence, the combinatorial diversity of possible sequences is $20^{28} = 3 \times 10^{36}$. To computationally model each of these sequences in a productive manner would require that several rotomers be considered for each amino acid at each position. Thus, the number of possible sequence and rotomer combinations is 1.1×10^{62} (ref 91). Even with fast computers, this level of combinatorial diversity precludes analysis of every possible sequence and rotomer combination. To meet this combinatorial challenge, the algorithm of Dahiyat and Mayo used dead end elimination to discard any residue/rotomer combination that is not part of the global minimum energy conformation (GMEC).[91] Implementation of these computational methods to redesign Zif268 converged on one optimal sequence, named full sequence design (FSD-1).[91] Dahiyat and Mayo synthesized this sequence and showed by NMR that the backbone fold of FSD-1 is very similar to the target structure, with an rms deviation between FSD-1 and the targeted structure of only 1.98 Å (and less than 1 Å for residues 8−26).

In addition to using computational methods to search through alternative packing arrangements, algorithms can also be devised to search for combinations of side chains that facilitate the incorporation of novel binding sites into preexisting protein scaffolds. Hellinga and co-workers developed an automated design algorithm, DEZYMER, which identifies regions in a natural protein structure that might be capable of accommodating metal-binding ligands in a predefined geometry.[94] DEZYMER sorts through the various possible combinations of side chain substitutions and selects those most likely to be compatible with a new metal binding site. Guided by the DEZYMER algorithm, Hellinga and co-workers constructed several different metal sites in natural proteins previously lacking such sites. They also showed that several of their novel metalloproteins are catalytically active.[69,95−98]

Most of the first-generation computational algorithms required that protein backbone structures remain fixed throughout the calculations. This requirement typically forced researchers to focus attention on natural proteins with known 3-dimensional structures. Consequently, the earliest computationally driven designs tended to be ‘redesigns’ of natural proteins rather than de novo designs of novel structures.

To address the limitations imposed by keeping the main chain fixed, Harbury et al.[99] set out to design novel coiled coil structures in which the conformation of the protein backbone was allowed to vary in accordance with “a family of parametric curves”.[100] The regular and symmetric pattern of coiled coils allowed the authors to use only three parameters, supercoil radius, supercoil frequency, and the **a**-position orientation angle, to describe the backbone conformation.[100] Their first use of this algorithm successfully predicted the structures of GCN4 coiled coil variants previously known to form dimeric, trimeric, or tetrameric structures.[100,101] Next they used their algorithm to search for de novo sequences that would fold into dimeric, trimeric, and tetrameric α-helical bundles having the never-before seen right-handed super-helical twist.[99] Rather than using the 7-fold repeating pattern of natural, left-handed bundles, the authors used a novel 11-fold repeating pattern aimed to yield a right-handed twist. This novel pattern contained three hydrophobic core residues per undecatad repeat. The algorithm searched through a combinatorial mix containing six different nonpolar amino acids, each with several low-energy rotomer conformations, to find sequences compatible with dimeric, trimeric, and tetrameric α-helical bundles. The best sequences, as scored by the algorithm, were synthesized and analyzed by sedimentation equilibrium. The novel sequences were found to oligomerize as intended. The crystal structure of RH4, the sequence targeted to form a tetrameric α-helical bundle, was solved and compared to the calculated structure. RH4 not only folded into a four-helix bundle with a right-handed super-helical twist, but the experimentally determined structure “matched the designed structure in atomic detail”.[99]

X. Concluding Remarks

As the era of proteomic research continues to accelerate, it is apparent that a new era, which progresses ‘beyond proteomics’, is already under way. Research into the structures and functions of proteins need not be limited to ‘only’ all proteins that exist on earth. It is already possible to venture beyond proteomics and to devise vast combinatorial libraries of novel proteins not found in nature. Recent work by Szostak and co-workers has shown that even libraries of random sequences—not constrained by elements of rational design—can occasionally yield functional polypeptides.[102,103] Keefe and Szostak used an mRNA display system to screen a library of 6 × 10^{12} sequences (containing 80 contiguous randomized residues) for rare sequences that bind ATP. Following several rounds of selection, four new ATP binding sequences were discovered. The ability of these novel sequences to fold into protein-like structures was not determined. Nonetheless, the finding that specific binding molecules can be isolated from random libraries indicates that if libraries are large enough and if selection methods are stringent enough, even random libraries can yield novel proteins with interesting properties.

As described in the preceding sections, by using rational design to guide the search through sequence space, a variety of structural and functional de novo proteins can be isolated—even without the use of selections. These rationally designed combinatorial libraries have already yielded proteins that are (i) α-helical, (ii) β-sheet, (iii) monomeric, (iv) capable of self-assembly into ordered arrays, (v) well packed and nativelike, (vi) hyperthermostable, (vii) capable of binding cofactors, and (viii) catalytically active.

Future work on de novo proteins from combinatorial libraries will incorporate all of the different approaches outlined in this review. Ultimately the most successful projects will probably use *rational methods* to constrain *large libraries* that are then subjected to *stringent selections* for desired properties.

The current revolution in proteomics research is providing the foundation for significant advances in biotechnology. As protein research progresses beyond proteomics, applications will no longer be limited to proteins derived from biological systems. Future advances in biotechnology will likely rely increasingly on combinatorial libraries of proteins designed entirely de novo.

XI. References

(1) Beasley, J. R.; Hecht, M. H. *J. Biol. Chem.* **1997**, *272*, 2031.
(2) Davidson, A. R.; Sauer, R. T. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 2146.
(3) Davidson, A. R.; Lumb, K. J.; Sauer, R. T. *Nat. Struct. Biol* **1995**, *2*, 856.
(4) Mandecki, W. *Protein Eng.* **1990**, *3*, 221.
(5) Rao, S. P.; Carlstrom, D. E.; Miller, W. G. *Biochemistry* **1974**, *13*, 943.
(6) Anufrieva, E. V.; Bychkova, V. E.; Krakovyak, M. G.; Pautov, V. D.; Ptitsyn, O. B. *FEBS Lett.* **1975**, *55*, 46.
(7) Katchalski, E.; Sela, M. *Adv. Prot. Chem.* **1958**, *XIII*, 243.
(8) DeGrado, W. F.; Summa, C. M.; Pavone, V.; Nastri, F.; Lombardi, A. *Annu. Rev. Biochem.* **1999**, *68*, 779.
(9) Lombardi, A.; Summa, C. M.; Geremia, S.; Randaccio, L.; Pavone, V.; Degrado, W. F. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 6298.
(10) Walsh, S. T. R.; Cheng, H.; Bryson, J. W.; Roder, H.; DeGrado, W. F. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 5486.

(11) Kohn, W. D.; Kay, C. M.; Sykes, B. D.; Hodges, R. S. *J. Am. Chem. Soc.* **1998**, *120*, 1124.
(12) Broo, K.; Brive, L.; Lundh, A−C.; Ahlberg, P.; Baltzer, L. *J. Am. Chem. Soc.* **1996**, *118*, 8172.
(13) Schafmeister, C. E.; LaPorte, S. L.; Miercke, L. J. W.; Stroud, R. M. *Nat. Struct. Biol.* **1997**, *4*, 1039.
(14) Severin, K.; Lee, D. H.; Kennan, A. J.; Ghadiri, M. R. *Nature* **1997**, *389*, 706.
(15) Raleigh, D. P.; DeGrado, W. F. *J. Am. Chem. Soc.* **1992**, *114*, 10079.
(16) Raleigh, D. P.; Betz, S. F.; DeGrado, W. F. *J. Am. Chem. Soc.* **1995**, *117*, 7558.
(17) Hill, R. B.; DeGrado, W. F. *J. Am. Chem. Soc.* **1998**, *120*, 1138.
(18) Lovejoy, B.; Choe, S.; Cascio, D.; McRorie, D. K.; DeGrado, W. F.; Eisenburg, D. *Nature* **1993**, *259*, 1288.
(19) Altamirano, M. M.; Blackburn, J. M.; Aguayo, C.; Fersht, A. R. *Nature* **2000**, *403*, 617.
(20) Iffland, A.; Tafelmeyer, P.; Saudan, C.; Johnsson, K. *Biochemistry* **2000**, *39*, 10790.
(21) Jung, S.; Honegger, A.; Plueckthun, A. *J. Mol. Biol.* **1999**, *294*, 163.
(22) Wang, L.; Kong, X.; Zhang, H.; Wang, X.; Zhang, J. *Biochem. Biophys. Res. Commun.* **2000**, *276*, 346.
(23) Reidhaar-Olson, J. F.; Sauer, R. T. *Science* **1988**, *241*, 53.
(24) Giver, L.; Arnold, F. H. *Curr. Opin. Chem. Biol.* **1998**, *2*, 335.
(25) Cesareni, G.; Castagnoli, L.; Cestra, G. *Comb. Chem. High Throughput Screening* **1999**, *2*, 1.
(26) Dell, A.; Imperiali, B.; McLaughlin, L. *Curr. Opin. Chem. Biol.* **1997**, *1*, 523.
(27) Imperiali, B.; Ottesen, J. J. *J. Pept. Res.* **1999**, *54*, 177.
(28) Venkatesh, N.; Im, S. H.; Balass, M.; Fuchs, S.; Katchalski-Kazir, E. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 761.
(29) Wrighton, N. C.; Farrell, F. X.; Chang, R.; Kashyap, A. K.; Barbone, F. P.; Mulcahy, L. S.; Johnson, D. L.; Barrett, R. W.; Jolliffe, L. K.; Dower, W. J. *Science* **1996**, *273*, 458.
(30) Maeji, N. J.; Valerio, R. M.; Bray, A. M.; Campbell, R. A.; Geysen, H. M. *React. Polym.* **1994**, *22*, 203.
(31) Barbas III, C. F.; Burton, D. R.; Scott, J. K.; Silverman, G. J. *Phage Display: A Laboratory Manual*; Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY, 2001.
(32) Regan, L.; Degrado, W. F. *Science* **1988**, *241*, 976.
(33) Hecht, M. H.; Richardson, J. S.; Richardson, D. C.; Ogden, R. C. *Science* **1990**, *249*, 884.
(34) Kamtekar, S.; Schiffer, J. M.; Xiong, H.; Babik, J. M.; Hecht, M. H. *Science* **1993**, *262*, 1680.
(35) Roy, S.; Helmer, K. J.; Hecht, M. H. *Folding Des.* **1997**, *2*, 89.
(36) Roy, S.; Ratnaswamy, G.; Boice, J. A.; Fairman, F.; McLendon, G.; Hecht, M. H.; *J. Am. Chem. Soc.* **1997**, *119*, 5302.
(37) Roy, S.; Hecht, M. H. *Biochemistry*. **2000**, *39*, 4603.
(38) Rosenbaum, D. M.; Roy, S.; Hecht, M. H. *J. Am. Chem. Soc.* **1999**, *121*, 9509.
(39) Axe, D. D.; Foster, N. W.; Fersht, A. R. *Proc. Natl. Acad. Sci.* **1996**, *93*, 5590.
(40) Gassner, N. C.; Baase, W. A.; Matthews, B. W. *Proc. Natl. Acad. Sci.* **1996**, *93*, 12155.
(41) Lim, W. A.; Sauer, R. T. *Nature* **1989**, *339*, 31.
(42) Riddle, D. S.; Santiago, J. V.; Bray-Hall, S. T.; Doshi, N.; Grantcharova, V. P.; Yi, Q.; Baker, D. *Nat. Struct. Biol.* **1997**, *4*, 805.
(43) Silverman, J. A.; Balakrishnan, R.; Harbury, P. B. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 3092.
(44) Lau, K. F.; Dill, K. A. *Proc. Natl. Acad. Sci. U.S.A.* **1990**, *87*, 638.
(45) Wei, Y.; Hecht, M. H. Unpublished results.
(46) Wei, Y.; Liu, T.; Pelczer, I.; Sazinsky, S. L.; Moffet, D. A.; Hecht, M. H. Submitted for publication.
(47) West, M. W.; Wang, W.; Patterson, J.; Mancias, J. D.; Beasley, J. R., Hecht, M. H. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 11211.
(48) Broome, B. M.; Hecht, M. H. *J. Mol. Biol.* **2000**, *296*, 961.
(49) Johnson, B. H.; Hecht, M. H. *Bio/Technology* **1994**, *12*, 1357.
(50) Robertson, D. E.; Farid, R. S.; Moser, C. C.; Urbauer, J. L.; Mulholland, S. E.; Pidikiti, R.; Lear, J. D.; Wand, A. J.; DeGrado, W. F.; Dutton, P. L. *Nature* **1994**, *368*, 425.
(51) Gibney, B. R.; Rabanal, F.; Skalicky, J. J.; Wand, A. J.; Dutton, P. L. *J. Am. Chem. Soc.* **1999**, *121*, 4952.
(52) Skalicky, J. J.; Gibney, B. R.; Rabanal, F.; Urbauer, R. J. B.; Dutton, P. L.; Wand, A. J. *J. Am. Chem. Soc.* **1999**, *121*, 4941.
(53) Case, M. A.; McLendon, G. L. *J. Am. Chem. Soc.* **2000**, *122*, 8089.
(54) Arndt, K. M.; Pelletier, J. N.; Mueller, K. M.; Alber, T.; Michnick, S. W.; Plueckthun, A. *J. Mol. Biol.* **2000**, *295*, 627.
(55) Joslyn, G.; Richardson, D. S.; White, R.; Alber, T. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 11109.
(56) Braisted, A. C.; Wells, J. A. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 5688.
(57) Nilsson, B.; Moks, T.; Jansson, B.; Abrahamsen, L.; Elmblad, A.; Holmgren, E.; Henrichson, C.; Jones, T. A.; Uhlen, M. *Protein Eng.* **1987**, *1*, 107.
(58) Cedergren, L.; Anderssen, R.; Jansson, B.; Uhlen, M.; Nilsson, B. *Protein Eng.* **1993**, *6*, 441.
(59) Deisenhofer, J. *Biochemistry* **1981**, *20*, 2361.
(60) Gouda, H.; Torigoe, H.; Saito, A.; Arata, Y.; Shimada, I. *Biochemistry* **1992**, *31*, 9665.
(61) Hutson, J. S.; Cohen, C.; Maratea, D.; Fields, F.; Tai, M. S.; Cabral-Denison, N.; Juffras, R.; Rueger, D. C.; Ridge, R. J.; Oppermann, H.; Keck, P.; Baird, L. G. *Biophys J.* **1992**, *62*, 87.
(62) Starovasnik, M. A.; Braisted, A. C.; Wells, J. A. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 10080.
(63) *Enzyme Nomenclature*; Academic Press: San Diego, 1992; p 862.
(64) Hellinga, H. W. *Folding Des.* **1998**, *3*, R1.
(65) Dieckmann, G. R.; McRorie, D. K.; Tierney, D. L.; Utschig, L. M. Singer, C. P. O'Halloran, T. V.; Penner-Hahn, J. E.; DeGrado, W. F., Pecoraro, V. L. *J. Am. Chem. Soc.* **1997**, *119*, 6195.
(66) Klemba, M.; Gardner, K. H.; Marino, S.; Clarke, N. D.; Regan, L. *Nat. Struct. Biol.* **1995**, *2*, 368.
(67) Lu, Y.; Valentine, J. S. *Curr. Opin. Struct. Biol.* **1997**, *7*, 495.
(68) Wilcox, S. K.; Putnam, C. D.; Sastry, M.; Blankenship, J.; Chazin, W. J.; McRee, D. E.; Goodin, D. B. *Biochemistry* **1998**, *37*, 16853.
(69) Benson, D. E.; Wisz, M. S.; Hellinga, H. W. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 6292.
(70) Benson, D. R.; Hart, B. R.; Zhu, X.; Doughty, M. B. *J. Am. Chem. Soc.* **1995**, *117*, 8502.
(71) Wilcox, S. K.; Putnam, C. D.; Sastry, M.; Blankenship, J.; Chazin, W. J.; McRee, D. E.; Goodin, D. B. *Biochemistry* **1998**, *37*, 16853.
(72) Sigman, J. A.; Kwok, B. C.; Lu, Y. *J. Am. Chem. Soc.* **2000**, *122*, 8192.
(73) Hellinga, H. W. *J. Am. Chem. Soc.* **1998**, *120*, 10055.
(74) Gibney, B. R.; Dutton, P. L. *Protein Sci.* **1999**, *8*, 1888.
(75) Arnold, P. A.; Shelton, W. R.; Benson, D. R. *J. Am. Chem. Soc.* **1997**, *119*, 3181.
(76) Choma, C. T.; Lear, J. D.; Nelson, M. J.; Dutton, P. L.; Robertson, D. E.; DeGrado, W. F. *J. Am. Chem. Soc.* **1994**, *116*, 856.
(77) Rau, H. K.; DeJonge, N.; Haehnel, W. *Angew. Chem., Int. Ed.* **2000**, *39*, 250.
(78) Schnepf, R.; Hoerth, P.; Bill, E.; Wieghardt, K.; Hildebrandt, P.; Haehnel, W. *J. Am. Chem. Soc.* **2001**, *123*, 2186.
(79) Rojas, N. R. L.; Kamtekar,S.; Simons, C. T.; Mclean, J. E.; Vogel, K. M.; Spiro, T. G.; Farid, R. S.; Hecht, M. H. *Protein Sci.* **1997**, *6*, 2512.
(80) Moffet, D. A.; Case, M. A.; House, J. C.; Vogel, K.; Williams, R. D.; Spiro, T. G.; McLendon, G. L.; Hecht, M. H. *J.Am. Chem. Soc.* **2001**, *123*, 2109.
(81) Moffet, D. A.; Certain, L. K.; Smith, A. J.; Kessel, A. J.; Beckwith, K. A.; Hecht, M. H. *J. Am. Chem. Soc.* **2000**, *122*, 7612.
(82) Regan, L. *Structure* **1998**, *6*, 1.
(83) Hellinga, H. W. *Nat. Struct. Biol.* **1998**, *5*, 525.
(84) Gordon, D. B.; Marshall, S. A.; Mayo, S. L. *Curr. Opin. Struct. Biol.* **1999**, *9*, 509.
(85) Street, A. G.; Mayo, S. L. *Structure* **1999**, *7*, R105.
(86) Desjarlais, J. R.; Handel, T. M. *Protein Sci.* **1995**, *4*, 2006.
(87) Shenkin, P. S.; Farid, H.; Fetrow, J. S. *Proteins: Struct., Funct., Genet.* **1996**, *26*, 323.
(88) Jiang, X.; Bishop, E. J.; Farid, R. S. *J. Am. Chem. Soc.* **1997**, *119*, 838.
(89) Jiang, X.; Farid, H.; Pistor, E.; Farid, R. S. *Protein Sci.* **2000**, *9*, 403.
(90) Dahiyat, B. I.; Mayo, S. L. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 10172.
(91) Dahiyat, B. I.; Mayo, S. L. *Science* **1997**, *278*, 82.
(92) Desmet, J.; De Meyer, M.; Hazes, B.; Lasters, I. *Nature* **1992**, *356*, 539.
(93) De Meyer, M.; Desmet, J.; Lasters, I. *Folding Des.* **1997**, *2*, 53.
(94) Hellinga, H. W.; Richards, F. M. *J. Mol. Biol.* **1991**, *222*, 763.
(95) Benson, D. E.; Wisz, M. S.; Liu, W.; Hellinga, H. W. *Biochemistry* **1998**, *37*, 7070.
(96) Wisz, M. S.; Garrett, C. Z.; Hellinga, H. W. *Biochemistry* **1998**, *37*, 8269.
(97) Coldren, C. D.; Hellinga, H. W.; Caradonna, J. P. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 6635.
(98) Pinto, A. L.; Hellinga, H. W.; Cardonna, J. P. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 5562.
(99) Harbury, P. B.; Plecs, J. J.; Tidor, B.; Alber, T.; Kim, P. S. *Science* **1998**, *282*, 1462.
(100) Harbury, P. B.; Tidor, B.; Kim, P. S. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 8408.
(101) Harbury, P. B.; Zhang. T.; Kim, P. S.; Alber, T. *Science* **1993**, *262*, 1401.
(102) Cho, G.; Keefe, A. D.; Liu, R.; Wilson, D. S.; Szostak, J. W. *J. Mol. Biol.* **2000**, *297*, 309.
(103) Keefe, A. D.; Szostak. J. W. *Nature* **2001**, *401*, 715.

CR000051E