

MINIREVIEW

Combinatorial approaches to protein stability and structure

Thomas J. Magliery¹ and Lynne Regan^{1,2}¹Department of Molecular Biophysics & Biochemistry and ²Department of Chemistry, Yale University, New Haven, CT, USA

Why do proteins adopt the conformations that they do, and what determines their stabilities? While we have come to some understanding of the forces that underlie protein architecture, a precise, predictive, physicochemical explanation is still elusive. Two obstacles to addressing these questions are the unfathomable vastness of protein sequence space, and the difficulty in making direct physical measurements on large numbers of protein variants. Here, we review combinatorial methods that have been applied to problems in protein biophysics over the last

15 years. The effects of hydrophobic core composition, the most important determinant of structure and stability, are still poorly understood. Particular attention is given to core composition as addressed by library methods. Increasingly useful screens and selections, in combination with modern high-throughput approaches borrowed from genomics and proteomics efforts, are making the empirical, statistical correlation between sequence and structure a tractable problem for the coming years.

Introduction

Understanding the basis of protein stability and structure is a problem of fundamental chemical and physical significance. In addition, such knowledge is critical for numerous biomedical applications, including but not limited to the preparation of stable protein-based therapeutics and the treatment of pathologies related to mutated, unstable proteins [1–4]. The importance of this issue has led to considerable study, at least since the first protein crystal structures were determined [5–7]. In spite of such attention, a satisfactory understanding of how proteins adopt the conformations that they do is still far from complete.

Why has it been so difficult to develop a precise physicochemical model of protein structure? To the extent that it is true that the *in vivo* conformation of proteins is encoded entirely by the primary structure, a sufficiently broad survey of protein variants must contain, in the limit, all that we need to know to understand the basis of protein stability. The problem is that the number of possible protein variants is incomprehensibly large, the biophysical characterization of proteins is slow, and the resulting paucity of data makes it difficult to parameterize potential functions correlating structure and sequence. Sequence space for even a very small protein (e.g. 50 amino acids or 6 kDa) is mind-bogglingly large (one molecule each of the 10^{65} variants would weigh in at 10^{39} tonnes; approximately the mass of the Milky Way galaxy). We currently lack the theoretical framework to quantitatively predict the effects of even a single point mutation, even for the simplest protein-like

structures, such as coiled-coils. Remarkable computational successes, such as the *in silico* redesigns of a zinc-free 'zinc finger' [8] and a right-handed coiled-coil [9], belie the fact that we cannot reliably predict the effects of hydrophobic core mutations (even if we can distinguish some destabilized variants from some stable ones) [10,11]. Indeed, there is still widespread debate about the restrictiveness of stereochemical constraints of the amino acids on the ability to achieve stable protein structures, with extreme views favoring the dominance of hydrophobic surface burial (like an oil droplet) [12] or the difficulty of achieving intimate van der Waals packing (like a jigsaw puzzle) [13].

The problem can therefore be framed simply: we need a way to (a) make large numbers of variants of proteins and (b) to analyze them rapidly for structure and stability. Practically speaking, if we are going to analyze a large number of protein variants en masse, then we must also (c) have a way to rapidly identify which proteins were sorted into a particular category.

It is now possible, using a combination of chemical DNA oligonucleotide synthesis and PCR-based methods, to create genes encoding virtually any protein or library of protein variants that is desired. Using clever synthetic strategies, the mix of amino acids encoded at a given position can be biased by judicious mixing of phosphoramidites [14] or even specified precisely using mixtures of trinucleotide phosphoramidites [15,16] in DNA synthesis. It is possible to use the genetic code to specify mixes of amino acids with a desired property (e.g. NTN, where N is an equimolar mix of all four nucleotides, encodes a hydrophobic position with a mix of Phe, Leu, Ile, Met and Val) and at the same time reduce undesirable properties of the genetic code (e.g. NNK, where K is an equimolar mix of G and T, is less biased than NNN toward Leu, Ser and Arg, and includes only one stop codon). However, the natural repertoire of amino acids is highly restrictive compared to the useful alterations that can be made to small molecules by physical organic chemists, and methods to incorporate unnatural amino acids are only just becoming broadly practical [17].

Correspondence to L. Regan, Department of Molecular Biophysics & Biochemistry, Yale University, New Haven, CT, USA.

Fax: + 1 203 432 5767, Tel.: + 1 203 432 9843,

E-mail: lynne.regan@yale.edu

Abbreviations: TIM, triosephosphate isomerase.

(Received 5 January 2004, revised 27 February 2004, accepted 5 March 2004)

Sorting libraries of proteins for structural properties is especially challenging. The ability to make libraries of protein variants has been widely exploited to understand and alter the function of proteins, because methods like metabolic selections and phage-display make it possible to tie the function of a protein variant to a phenotype (survival or binding, for example) allowing rapid sorting of the protein variants [18]. It is much less straightforward to screen or select for protein structure and stability; X-ray crystallography, NMR spectroscopy or even CD spectroscopy are not amenable to especially high throughput approaches. However, the behavior of stable, native-like proteins differs from unstructured polypeptides, and the consequences of this can be used to sort polypeptide libraries for native-like proteins. We will discuss the methods for this in some depth.

Even so, once one has sorted proteins for physical properties, one must identify those proteins. The most straightforward way to do this is to link genotype to phenotype using a functional selection or screen. Unlike proteins, nucleic acids can be amplified and readily sequenced, allowing one to identify a single selected molecule, at least in principle. Thus, the first proteins studied for stability in library format were those for which *in vivo* genetic selections were available: tryptophan synthase [19–21], lac repressor [22] and lambda repressor [23,24]. More recently, display methods that do not require cellular function have been developed, such as phage-display, ribosome-display and mRNA-display. These methods have largely been limited to identification of protein variants that are competent for binding to an immobilized ligand, but they allow rapid identification due to the linkage of encoding genetic material.

A complementary approach to the large-scale analysis of protein variants is the design or redesign of a protein, either in systematic fashion or using combinatorial methods. Design or redesign is an especially exacting test of our understanding of protein architecture, because the extent to which we can design or redesign a particular fold is essentially a proof of the validity of the underlying hypothetical design principles. It is appropriate to call combinatorial studies of proteins ‘designs’ because these studies are essentially hypothesis-driven. At the end of the day (perhaps a rather long day), we want to be able both to understand what makes proteins ‘tick’ and to engineer proteins with native-like properties.

In this review we discuss combinatorial approaches toward understanding protein structure and stability. In the ideal case, such studies will allow us to answer questions like: Can we identify all possible sequences that can form a particular stable fold? Can we understand why these sequences ‘work’ and why others do not? How is the free-energy landscape of a fold affected by mutation? Can we use the data from these studies to predict the stability of a sequence that adopts a certain fold?

Systematic versus combinatorial studies

There are essentially two complementary approaches to tackling the incompatibility of the vast size of protein sequence space and our limited ability to examine large numbers of molecules directly for physical properties. One

can make a small number of rational protein variants and examine their physical properties thoroughly, or one can make a library of variants and sort them by screen or selection for those molecules that deserve further examination. The minireviews in this series are concerned with what screens and selections can be applied, and what they are actually selecting for.

Much of what we know rigorously about protein stability has been derived from systematic studies of small model proteins like the T4 lysozyme, the B1 domain of protein G, lambda repressor, staphylococcal nuclease, barnase and rop, as well as the *de novo* design of even smaller coiled-coils. These studies have highlighted some guiding principles for the design of native-like proteins and have provided quantitative measures of the energies associated with different types of interactions. These guiding principles, such as the necessity of defining water-soluble solvent-exposed regions and buried hydrophobic regions, the destabilizing effects of overpacking or underpacking the core, the role of buried hydrogen bonds and charge–charge interactions in specifying stability and structural uniqueness and the presence of ‘negative elements’ that disfavor other energetically near conformations, help us construct combinatorial experiments to test the generality of the underlying ideas. Systematic and *de novo* methods of protein design and redesign have been excellently reviewed elsewhere, and we will focus here on combinatorial methods [25–31].

Selecting for folded proteins

Combinatorial methods essentially require three elements: construction of a library of molecular variants, selection or screening of the library for molecules with desired properties and identification of selected variants (Fig. 1).

Constructing the library

For the purposes of the studies we will discuss, library construction is not usually a limiting step. PCR-based methods using synthetic DNA oligonucleotides, made with mixes of phosphoramidites at specific positions, make it possible to create virtually any set of desired protein variants in library sizes that vastly exceed what can be screened practically. In principle, recombinant methods like DNA shuffling can be used to rapidly create second generation libraries enriched in desirable properties [32]. There are still limitations, to be sure. DNA oligonucleotides are limited to about 100 nucleotides in conventional synthesis, requiring that longer genes must be pieced together with PCR-based methods. Using mixed phosphoramidites to create ‘degenerate’ codons, it is not possible to specify every mix of amino acids, due to the limitations of the genetic code; neither is it possible to simultaneously synthesize oligonucleotides of different integral lengths. (An EXCEL worksheet for planning degenerate codons from equimolar mixes of phosphoramidites is available from the Regan Group webpage at <http://www.csb.yale.edu/people/regan/publications.html> [T. J. Magliery, unpublished].) Achieving a specific mix of codons at a given position (for example, using trinucleotide phosphoramidites [16]), or generating a library with insertions or deletions [33], are sufficiently challenging or expensive that they are not yet widely useful. In addition,

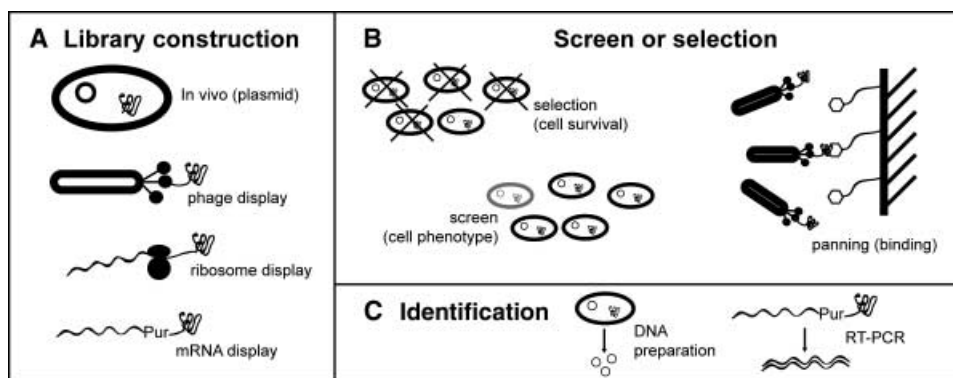


Fig. 1. Scheme of a combinatorial experiment. Protein libraries must be constructed so that screening or selection is possible, and identification of selectants is facile. (A) Proteins can be expressed in cells (usually bacteria, usually from a plasmid), displayed on the surface of filamentous phage, displayed on stalled ribosomes or covalently linked to coding RNA through puromycin. (B) Cells expressing proteins of interest are then distinguished by cellular survival (selection) or phenotype (screen); displayed proteins are typically sorted by binding to an immobilized ligand. (C) The selected proteins are then identified by isolation of DNA from cells or phage, or RT-PCR of RNA linked to protein in other *in vitro* display methods.

it will eventually be useful to make protein alterations less blunt than the exchange of the 20 members of the natural repertoire, but technology to do this is not yet widely practical [17,34,35]. For our purposes, we shall assume that useful libraries can be created in a fairly straightforward manner, and we will focus instead on the issue of screening those libraries.

Screens and selections

The earliest applications of selections and screens for protein structure and stability were derived from genetic studies. Therefore, the proteins studied in this fashion were those for which a convenient genetic screen was available. For example, tryptophan synthase function is required for survival on tryptophan-free medium; lambda repressor prevents superinfection with lytic phage; and lac repressor prevents transcription of β -galactosidase, which can be assayed by survival on lactose minimal medium or hydrolysis of a chromogenic galactoside. The latter case illustrates the fundamental difference between selections and screens. In a selection, such as survival on a particular medium, only those cells with functional protein survive. This allows the examination of a large number of variants (10^9 or more), but it also prevents one from examining the nonfunctional variants (which were in dead cells). Screens, such as turnover of a chromogenic substrate, allow access to nonfunctional variants, but are not useful if only a tiny fraction of the library is active, and generally limit the number of clones that can be examined (10^3 – 10^6 , typically).

These genetic studies posited the idea that passing the screen or selection required that the protein of interest be functional, and that a functional protein must be a structured protein. However, the range of conditions that can be applied to living cells is small, and the exact nature of the selective pressure is not always easy to deduce. But the biggest limitation to these sorts of genetic approaches is that not every protein's function can be tied to the survival of a cell or some easy-to-observe phenotypic property. Ultimately, one would like to be able to study proteins whose

functions are not necessarily critical to the survival of the cell, and one would like to be able to apply selective pressures that are not compatible with cellular survival (such as high temperature or denaturant). The problem is that there is another limitation to library approaches: one must be able to identify the functional proteins at the end of the experiment.

Identification of selectants

There is no straightforward way to identify a protein sequence, particularly if only a small number of protein molecules are selected. The best possible direct solution, mass spectrometry, is typically insufficient for identification of the vanishingly small amounts of selected proteins from a library. The best practical solution conceived to date is the linkage of nucleic acid encoding the protein to the protein itself (i.e. linkage of genotype to phenotype), because even single molecules of nucleic acid can be amplified and then sequenced. The two most popular methods for achieving this linkage are by expressing the protein in a cell (usually from a plasmid) as in genetic methods, or displaying it on the surface of filamentous phage. As phage-display does not require that the protein be functional, nearly any protein can be examined by this method. In both of these cases, library size is limited by the essential step of transformation of DNA, and transformation efficiency and reaction size place this limit at about 10^{10} at the extreme in *Escherichia coli*, more often 10^6 – 10^9 . (The situation is worse in other hosts.) Two recently developed methods overcome this limitation by performing the translation reaction *in vitro*: ribosome-display [36], where the protein and mRNA are bound to the ribosome after translation, and mRNA- or puromycin-display [37], where the mRNA is covalently linked to the translated protein, allowing libraries of 10^{13} or larger. However, as with phage-display, the library members are not separately compartmentalized as they are in cells, which places some limits on the kinds of screens and selections that are applicable. Specifically, display methods are most suitable for binding studies.

Table 1. Screens and selections for folded proteins. GB1, B1 domain of protein G.

Basis	Methods	Comments	References
Cellular Expression	SDS/PAGE and crude NMR or MS screening (¹⁵ N HSQC, ¹ H 1D, amide exchange)	Low throughput but direct; requires libraries rich in interesting proteins	[12, 42–45]
	Fusion of reporter protein (green fluorescent protein, chloramphenicol acetyltransferase, lacZ α , Gal11P-AD, RNase-A S-peptide) to C-terminus of analyte	Screens for lack of aggregation or proteolysis; all but green fluorescent protein can be used in selection	[46–52]
	β -galactosidase under the control of promoters for genes that respond to 'translational stress'	Determined from microarray analysis of transcription; the specific basis of what is monitored is not well understood	[53]
	Secretion in yeast	Secondary screen is required as some unfolded proteins are secreted	[54]
Resistance to Proteolysis	Filamentous phage-display between the phage with a binding domain (like His ₆); <i>in vitro</i> treatment with protease	On beads or chips (using surface plasmon resonance); incorporation of a specific protease site is often helpful	[57–60]
	<i>In vitro</i> proteolytic treatment of ribosome-displayed proteins	Can be combined with hydrophobic interaction chromatography	[61]
Ligand Binding	Phage-displayed proteins (GB1, protein L, SH2/SH3) panned against immobilized ligand	Has been combined with 'loop-entropy screen'	[62–68]
	mRNA or ribosome-displayed proteins panned against an immobilized ligand	Allows access to very large libraries (> 10 ¹⁰) but lacks compartmentalization (like phage)	[69]
	<i>In vivo</i> binding to DNA (λ repressor) or RNA (rop) monitored by cellular function (resistance to lytic phage or plasmid copy number change)	Requires knowledge of which residues are required for binding; screens and selections are possible	[70–72]
Catalytic Activity	<i>In vivo</i> activity of proteins (barnase, chorismate mutase, triosephosphate isomerase), usually linked to cellular survival	Requires knowledge of which residues are required for catalysis; screens and selections are possible	[73–77]

Selecting for native-like proteins

Combinatorial approaches to protein biophysics require that one makes a library of polypeptides and then sorts the library for stable, structured, native-like proteins. The question is: what makes a protein 'native-like'? In essence, a native-like protein is one with thermodynamic and structural properties that are exhibited by 'normal' cellular proteins (i.e. native proteins). Presumably, these properties arise from the precise balance of interactions that native proteins possess, especially in the core. There are a number of measurable physical properties that reflect nativity. Ideally, native-like proteins will have highly cooperative denaturation transitions with high per-residue ΔH° and ΔC_p , will possess a subset of slowly exchanging amide protons, will be resistant to binding hydrophobic dyes, and will have well-resolved NMR spectra [28]. Obviously, none of these criteria is especially easy to screen in high throughput format (although the throughput of X-ray crystallography [38,39] and calorimetry [40] is increasing rapidly for drug discovery and proteomics). However, as a consequence of a native-like protein's stability and structural specificity, it is typically highly soluble, resistant to proteolysis and able to be expressed at high levels. Moreover, with few possible exceptions (so-called natively unfolded proteins [41]), functional proteins are necessarily structured proteins (probably in part due to the fact that

cellular function demands expression and proteolysis resistance). Thus, in general, proteins that bind ligands or catalyze reactions *in vivo* can be expected to be relatively native-like. (Table 1 shows a summary of screens and selection for protein stability and structure.)

Cellular expression

One straightforward strategy of screening for structured proteins is to make limited or highly biased libraries and then screen them in a relatively low throughput format for expression. Proteins that are found in high levels in the soluble cellular fraction generally do not aggregate and are resistant to proteolysis. Gronenborn *et al.* for example, have randomized the seven-residue hydrophobic core of the B1 domain of IgG-binding protein G [42]. Individual clones were examined for expression and grown in the presence of a ¹⁵N source, allowing ¹H-¹⁵N HSQC NMR analysis of crude lysate for well-dispersed amide backbone spectra. However, a number of the structured variants possessed remarkably different tertiary and quaternary structures (through 'domain swapping'). The Hecht group has engineered several generations of four-helix bundles in which each individual position is encoded by a degenerate codon that specifies hydrophobic, hydrophilic or turn residues [12,43]. The resulting polypeptides were then examined for expression and later for well-dispersed ¹H NMR spectra from a

rapid, crude preparation of protein [44]. Rosenbaum *et al.* also used hydrogen-deuterium exchange of fairly crude preparations from binary pattern libraries to screen for proteins with subsets of slowly exchanging amide protons [45]. These methods rely on the generation of libraries wherein sequence space is relatively rich in native-like proteins.

Waldo and colleagues fused green fluorescent protein to the C-terminus of analyte proteins [46,47]. Cellular fluorescence was found to correspond to the solubility of the analyte protein (implying correct folding), presumably due to aggregation or degradation of misfolded analyte fusions. This idea has been employed with other protein fusions as well [48], including chloramphenicol acetyltransferase [49], *lacZ α* [50], Gal11P-activation domain [51] and RNase-A S-peptide [52], all of which allow selection (as opposed to screening), opening the door to larger library sizes.

Lesley *et al.* examined the differential expression of genes in *E. coli* during the overexpression of proteins of varying solubility using DNA microarrays [53]. A set of 'translation stress' proteins was upregulated, including some heat-shock genes and some ribosome-associated genes. The promoter regions of a number of the up-regulated genes were cloned into a plasmid to control the expression of β -galactosidase, resulting in strains in which protein misfolding is reported by β -galactosidase activity, for example using Gal-ONp chromogenic substrate. Another approach based on physiological response to protein misfolding was introduced by Hagihara & Kim, who exploited the fact that the yeast secretory pathway prevents the release of misfolded polypeptides [54]. Robust correspondence to the degree of protein folding required secondary screens for secretion into liquid culture after screening on agar plates, as well as nonreducing SDS/PAGE to identify proteins that migrate in a single, tight band.

A caveat to this approach is that it is not difficult to think of bona fide native proteins that express poorly, aggregate or are susceptible to degradation. Conversely, some selections for cellular expression have resulted in surprising escape variants. Revertants of a defective mutant of Arc repressor, for example, were found to express at high levels despite poor thermodynamic stability. These revertants acquired C-terminal extensions through frame-shift mutation which were shown to protect these and other proteins from intracellular proteolysis [55]. This is a clear case of 'getting what you select for'. Screens for cellular expression will yield folded proteins only to the extent to which folding is required for cellular expression. The underlying assumptions of all screens and selections must be carefully scrutinized for the true nature of the selective pressure being applied.

Resistance to *in vitro* proteolysis

Clearly, any protein that can be purified must be sufficiently resistant to proteolysis that its production exceeds its degradation. However, a number of researchers have shown that proteolysis resistance can be used directly as a marker of foldedness [56]. Woolfson and colleagues fused ubiquitin core variants between phage coat protein pIII and a hexahistidine tag [57]. After binding to a Ni-nitrilotriacetic

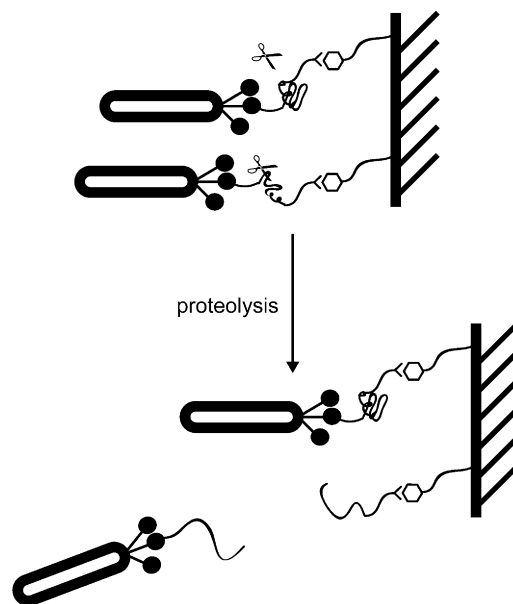


Fig. 2. Scheme of phage-display/proteolysis. Analyte proteins are displayed in the surface of phage, typically between a coat protein and a binding domain, such as a hexahistidine tag. The phage are then immobilized (for example, on Ni-nitrilotriacetic acid agarose) and treated with protease. Unfolded proteins are more rapidly cleaved and released from the solid support. After washing these phage away, those displaying folded proteins can be released by elution (for example, with imidazole), and can be used to reinfect cells or directly analyzed for DNA sequence.

acid surface, the phage fusions are treated with chymotrypsin and then eluted after washing (Fig. 2). The resulting selected phage can be used to reinfect bacteria and selection can be repeated to enrich in phage encoding the most resistant proteins. Similar methods were developed by Kristensen & Winter [58], Sieber *et al.* [59] and Bai and coworkers [60]. Bai's method includes the engineering of a specific protease site near the site of redesign, which Bai demonstrates will sometimes be critical [60a].

Matsuura & Plückthun have also used proteolysis resistance with ribosome-displayed proteins [61]. In combination with hydrophobic interaction chromatography, which removes (presumably unfolded) polypeptides with large hydrophobic patches exposed, this represents a selection based almost entirely on physical parameters of the polypeptide. (It still demands efficient ribosomal display, however.)

Ligand binding

In contrast to methods that screen or select for physical properties (more or less) directly, another way to look for native-like proteins is to infer native-like properties from function. This, however, presents a problem for library design: if one wants functional selectants to differ only structurally, then one must not mutate residues that directly affect function. Of course, some residues will have both functional and structural roles. The simplest function is arguably ligand binding. If, for example, one makes libraries of protein variants that differ in hydrophobic core compo-

sition but maintain all the surface residues necessary for binding, it is probable that most of the variation in ligand affinity will be due to the structural integrity of the protein. Thus, one must choose to make libraries of systematically well-studied proteins, or one must first delineate the 'functional residues' oneself.

Two examples of this approach are discussed in accompanying reviews [61a,61b]. Cochran and coworkers have examined the effect of cross-strand pairs in β -sheets by displaying variants of the B1 domain of IgG-binding protein G on filamentous phage [62]. Baker and coworkers have interrogated structural variant libraries of phage-displayed IgG-binding protein L and SH2/SH3 domains for binding to their ligands (IgG and a phosphotyrosyl peptide, respectively) [63–66]. A variation on this idea was to combinatorially design proteins *de novo* by inserting random sequences into a loop of the SH2 domain and screening for binding to the phosphotyrosyl ligand peptide [67]. In principle, folded insertions should reduce the entropic penalty for inserting a long loop, however, surprisingly, Baker and colleagues found that the free-energy penalty for long loops was generally small even for unfolded insertions (probably due to enthalpic effects and the entropic contribution of hydrophobic collapse) [68].

Both Plückthun and Szostak's groups have used ligand binding as a selection *in vitro* (using ribosome-display [36] and mRNA-display [37], respectively). For example, Keefe & Szostak isolated several ATP-binding protein aptamers from fully randomized 80-mers that bear no sequence resemblance to each other or to proteins known in nature [69]. However, these proteins were not sufficiently soluble to be examined *in vitro* except as fusions to maltose binding protein (presumably covalent linkage to the highly charged RNA template aids solubility in the selection).

Lim & Sauer carried out among the first and probably best-known combinatorial experiments in protein structure based on the binding of N-terminal variants of the λ repressor to lytic λ phage DNA, conferring resistance to phage infection and lysis to those cells with functional repressors [70,71]. Using lytic phages of differing virulence, the 'activity' of a repressor variant could be estimated (i.e. the stringency of the selection could be roughly controlled). These experiments are explained in more detail below.

Magliery & Regan have recently developed both positive and negative screens for the function of rop, a four-helix bundle protein that regulates the copy number of ColE1 plasmids [72]. Rop facilitates the binding of an inhibitory RNA to the RNA that primes plasmid replication (by binding to hairpin loops in both of those RNAs). By expressing green fluorescent protein from a ColE1 plasmid, cellular fluorescence reports the copy number of the plasmid and therefore rop functionality. This screen has been applied to libraries of hydrophobic core variants of rop (see below).

Catalytic activity

One can also infer native-like protein properties from catalytic activity of a protein variant, but the library design is even more complex than in the case of ligand binding, because the requirements for catalysis are more precise and less well understood. This is the basis of early 'genetic'

approaches to understanding the functional requirements of proteins like tryptophan synthase [20]. One such selection developed by Fersht and coworkers is based on the well-studied ribonuclease barnase [73]. As this is a negative selection (barnase activity is lethal to *E. coli*), barnase variants were encoded using two 'amber' stop codons (UAG) and transformed into both sup⁻ and supD *E. coli*, where death in the latter 'amber'-suppressing strain implies barnase activity. The use of the selection is described below. Hilvert and coworkers have extensively randomized chorismate mutase, which is required for the biosynthesis of phenylalanine and tyrosine, and therefore amenable to selection on media lacking these amino acids [74–76]. Chorismate mutase is thought to catalyze a Claisen condensation principally by binding chorismate in a conformation that favors the pericyclic reaction and allows transition-state stabilization by a cationic group, and the simplicity of this mechanism makes it possible to generate 'structural' variants without perturbing the function [76a].

Harbury and coworkers recently employed a selection based on triosephosphate isomerase (TIM) activity [77]. Although TIM barrels are more complex than simple structures like four-helix bundles (they possess two concentric hydrophobic cores, for example), they represent about 10% of known enzyme structures and are therefore tremendously important to understand structurally. This selection exploited the DNA shuffling method, wherein variants with a large number of randomized residues were shuffled with wild type TIM. The frequency of reversion of the randomized residues to the wild type residue is related to its necessity for activity.

Application of selections to protein design

Hydrophobic core redesign

Protein folding is driven in a large part by the formation of a hydrophobic core; it is clear from systematic studies that, at minimum, a protein has an 'inside' and an 'outside.' However, it is much less clear how specific the composition of the core must be for stability and overall structural uniqueness. Two limiting views of the basis of protein structure model the core of a protein as an oil droplet that separates from water, in which achieving intimate van der Waals contacts is relatively easy [12], or as a jigsaw puzzle, in which the complementary sizes, shapes and stereochemistries of residues are critical and restrictive [13]. Systematic studies offer support for both views. For example, a mutant of T4 lysozyme with 10 mutations of core residues to methionine retains substantial activity (20%) despite being much less stable ($\Delta\Delta G = 7.3 \text{ kcal}\cdot\text{mol}^{-1}$) [78]. In general, cavity-filling and cavity-creating mutations in T4 lysozyme are tolerated with small losses in activity and stability. However, these mutations result in proteins with similar backbone conformations as well as similar rotameric forms of interior sidechains; indeed, small backbone compensations seem to dominate over changes in sidechain positions [26]. (It is worth noting that this is the opposite paradigm to that employed in computational design programs like ROC [79], ORBIT [80] or the Hellinga group's dead-end elimination algorithm [81,82], wherein the backbone is fixed and residues are substituted and rotated to the lowest energy

solution. Harbury *et al.* have created a computational approach with backbone freedom [9].)

However, the only way to rigorously examine how core sequence corresponds to stability and structure is to make many core variants and examine them for biophysical parameters. A number of excellent reviews have been written on this subject [31,83–87]. The seminal studies of Lim & Sauer, and further work with Richards, are among the first and best-known attempts to address this issue. Seven buried residues in the N-terminal domain of λ repressor were completely randomized in groups of three residues [70]. Between 0.2% and 2% of mutants were active, depending upon the library and level of function demanded. The residues in active clones were dominated by Ala, Cys, Thr, Val, Ile, Leu, Met and Phe, a list that is interesting in that it includes a subset of the polar amino acids (no carboxamides or charged groups) and excludes Trp and Tyr while accepting Phe, perhaps due to conformational and hydrogen bonding requirements. The core volumes of active variants differed by only about 10%, or about +2 to –3 methylene groups relative to wild type, with slightly less variation among those with wild type-like activity. However, fewer proteins are active than would be predicted from these sequence and volume constraints alone, suggesting that factors such as stereochemical constraints on packing complementarity (jigsaw puzzle-like behavior) are prevalent.

A library in which the amino acids at three core positions were restricted to the hydrophobics (Val, Leu, Ile, Met and Phe, encoded by the mixed codon $DTS = \{AGT\}T\{CG\}$) was further analyzed [71]. About 70% of the 78 isolated variants were active (out of 125 possible combinations), but only two retained wild type-like stability and activity. Proteins with full activity at low temperatures or reduced but temperature-independent activity (implying similarity of structure and/or stability to the wild type) varied in volume over a very narrow range (two methylene groups), but those with any activity varied almost as much as all possible variants in the library (including inactive variants). This suggests that the overall structure is very tolerant of steric changes, but that precise structure and high stability are specified by a much smaller range of sequences.

One of these variants, the overpacked V36L M40L V47I mutant which has reduced activity (10-fold lower affinity for operator DNA) but high stability ($T_m = 59.6$ °C, as opposed to 55.7 °C for wild type), was crystallized for X-ray analysis (Fig. 3) [88,89]. The overpacking was accommodated primarily by a main-chain shift of the C-terminal helix away from the helices that contain the mutations, with the largest movements on the scale of 1 Å. The motion is rigid-body, in the sense that the helices themselves were not perturbed. The rotameric states of the internal side chains were all near ideal and essentially unchanged from wild type, and the packing was improved compared to wild type. This seems to highlight the importance of packing complementarity and the stereochemical nature of the constraints on that packing. However, the fact that the architecture of the repressor is fairly complex makes it difficult to extrapolate these results, except in general terms.

Barnase is a small (110 residue) protein that is structurally well-characterized, but is fairly complicated in architecture (Fig. 4) [90]. There are three fairly discrete core regions. The main core is composed of 13 amino acids that allow the

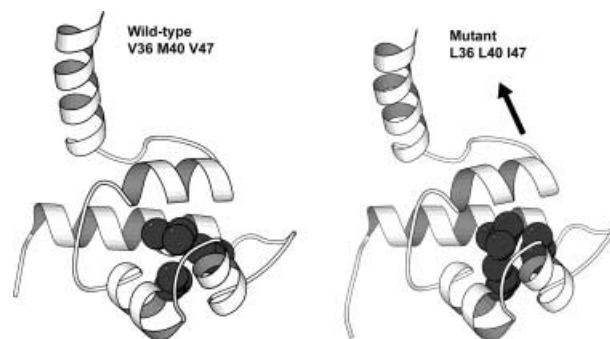


Fig. 3. Repacking λ repressor. An overpacked λ repressor, V36L M40L V47I, clearly has the same overall architecture as wild type repressor, but the C-terminus of helix 4 has shifted away from the core to accommodate the overpacking (as indicated by the arrow). Interestingly, most of the core residues retained near-ideal rotameric conformations in the mutant protein, meaning that subtle backbone rearrangement was preferred over stereochemical rearrangement of core residues. These three residues were altered using a combinatorial strategy described in the text. Rendered using MOLSCRIPT [113] from PDB entries 1LMB (wild type) and 1LLI (mutant).

packing of an α -helix against a five-strand antiparallel β -sheet. Axe *et al.* set out to explore a much larger sequence space than that addressed in the Lim & Sauer studies [73]. When the main core was mutated to all-hydrophobic amino acids in three stages, 57% of clones were active upon randomization of the six helix residues, and 23% were active upon additional randomization of six sheet-side residues. The frequency of active catalysts with random hydrophobic cores is strikingly large, as the oil-droplet model would suggest; nevertheless, four out of five cores with all hydrophobic amino acids are not functional (less than 0.2% wild type activity), implying jigsaw puzzle-like limits, as well. Moreover, the authors estimate that wild type-like activity is at least 1000-fold less common than the lower activity required to pass the selection. But even this must be put into perspective: half a billion different combinations of hydrophobic residues would be expected to be functionally equivalent to the wild type sequence. The core volumes of the active mutants varied by about 10%, which is striking considering that the largest (Phe13) and smallest (Val13)

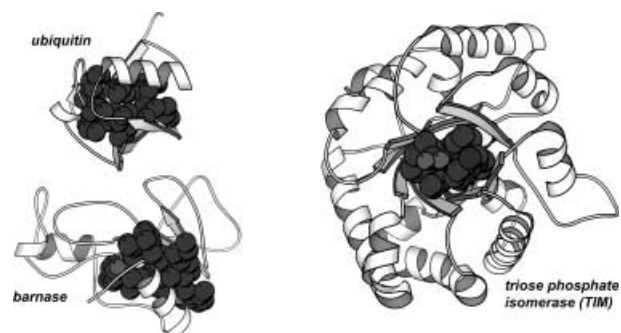


Fig. 4. Ubiquitin, barnase and triosephosphate isomerase (TIM). Side-chains of hydrophobic core residues randomized in work discussed in the text are rendered as spheres. For TIM, only those residues in the interior β -core are highlighted. Rendered using MOLSCRIPT from PDB entries 1UBI (ubiquitin), 1A2P (barnase) and 1YPI (yeast TIM).

random cores that could be produced in this experiment only differ by about 30%.

Recently, Silverman *et al.* employed an ambitious combinatorial approach to understanding the sequence requirements of the ubiquitous enzymatic fold called the $(\beta/\alpha)_8$ barrel, whose archetype is TIM [77]. Despite its importance, TIM is not an especially good model protein; it is fairly large, difficult to purify and has a complex double hydrophobic core (Fig. 4). The authors first sought to directly randomize the structural residues in TIM to estimate the overall tolerance to mutation. The library strategy was not only to avoid mutation of functionally important residues but to maintain the polarity of residues based on phylogenetic analysis (that is, multiple sequence alignment). Hence, hydrophilic residues were randomized to Lys, Glu and Gln; hydrophobic residues were mutated to Phe, Ile, Leu and Val; charged residues were mutated to Lys or Glu for basic or acidic positions, respectively; and variable positions were mutated to Ala. Only about one in 10^{10} variants in this library was active, in stark contrast to the high frequency of active core variants from barnase and λ repressor. Moreover, the identities of a handful of conserved hydrophobic residues and one conserved hydrophilic residue in selectants were biased significantly from the amino acid distribution in the naïve library, indicating an apparent violation of mere oil-droplet-like behavior.

Considering the low frequency of active variants, Silverman *et al.* needed another approach to examine the mutability of individual positions in library format. The approach was to mutate structural residues conservatively (e.g. V→L or D→N) in groups and then shuffle the resultant multiply mutated genes with wild type TIM. This procedure is known as ‘back-crossing’ in molecular breeding, and it is used to eliminate neutral mutations acquired during a molecular evolution experiment [32]. Here, the authors hypothesized that the frequency of reversion to wild type, which could occur in a variety of mutagenic backgrounds, is essentially a measure of the independent importance of the residue to structure (because only ‘structural’ residues were randomized). At 52 out of 105 positions, reversion to wild type occurred more frequently than expected by chance. Only four of these mutations were alone (i.e. in a wild type background) sufficient to reduce TIM activity below selectable levels, demonstrating the power of this approach in detecting important but less dramatic effects. The central core of the protein was surprisingly sensitive to mutation; 13 of 18 residues reverted frequently to wild type from the all-Val starting state, which is only a single methylene group larger than the wild type core. Other than these central core residues and glycines that act as β -stop signals, nearly every other kind of structural residue was highly mutable, including α/β interfaces, turns and α -helical capping and stop signals.

Finucane *et al.* found that the core of ubiquitin (Fig. 4) is also highly sensitive to mutation [91]. A library of ubiquitin variants in which eight core residues were randomized with hydrophobic amino acids was screened using phage-display and proteolysis, as described above. The selectants all have fewer than five mutations (by random chance, one would expect 6% to have fewer than five mutations), their consensus differs from wild type in only one position, and none of them are as stable as wild type. Lazar *et al.* used a

computational approach to redesign nine residues of the hydrophobic core of ubiquitin with Val, Leu, Ile and Phe [92]. Nine designed variants were evaluated *in vitro* and found to possess the overall ubiquitin fold, but all were less stable than wild type. This is in contrast to the Handel group’s computational redesign of 434 cro [79], and the authors suggest that β -sheet cores may be more sensitive to mutation than helical cores. While this trend appears to be true for T4 lysozyme, λ repressor, TIM and ubiquitin, the highly mutable barnase core is formed by the packing of a helix against β -strands, like the core of ubiquitin.

An even more sobering fact for the protein designer to confront is that comparatively conservative mutations of the monomeric hydrophobic core of the B1 domain of IgG-binding protein G resulted in radical ‘domain-swapped’ quaternary interactions leading to oligomericity of variants [93,94] (Fig. 5). A switch to an intertwined tetramer occurred with mutation of five out of nine core positions that were randomized with hydrophobic amino acids. This type of swapping may be at the root of the amyloidogenicity of some GB1 mutants [95]. This is also reminiscent of radical rearrangements of the four-helix bundle rop, which is an antiparallel homodimer (Fig. 5). Mutation of the six central four-residue ‘layers’ of the core to contain Ala₂Leu₂ results in a molecule that binds RNA *in vitro* (which is rop’s function), which was presumed to imply structural similarity to the wild type [96]. However, Ala₂Ile₂-6 is inactive, and the crystal structure reveals that the orientation of the

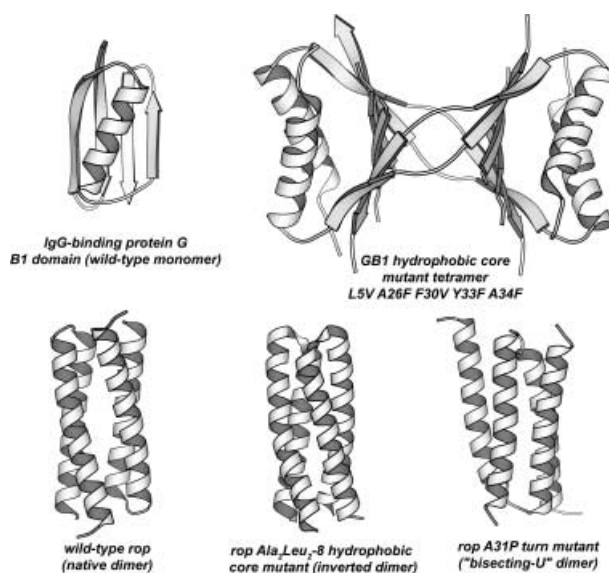


Fig. 5. Domain swapping and other quaternary rearrangements in protein G B1 domain and rop. Top: Mutagenesis of five residues in the core of the IgG-binding protein G B1 domain (left) results in a ‘domain swapped’ (right) tetramer, generally preserving but rearranging the secondary structural elements. Rendered using MOLSCRIPT from PDB entries 1PGA (GB1) and 1MVK (B1 core mutant). Bottom: Three different quaternary topologies are observed for wild type rop (native dimer, left), a rop mutant with a repacked hydrophobic core (inverted dimer, center) and a rop mutant that differs only in a single residue of the interhelical turn (bisecting-U dimer, right). Rendered using MOLSCRIPT from PDB entries 1ROP (wild type), 1F4M (rop Ala₂Leu₂-8), and 1B6Q (rop A31P).

monomers is inverted, splitting the binding site [97]. A mutation of the turn residue Ala31 to Pro results in another surprise in rop: the monomers remain antiparallel but interdigitate [98] in what has been dubbed a bisecting-U motif [99]. Although this is not a core mutation, it is perhaps more strange in that the core contacts are completely rearranged as a result of a turn-residue mutation. These sorts of results contrast with the view that the core provides stability but does not define the structure itself, a view that emerges from redesigns like that of ubiquitin in which even destabilized variants with multiple core mutations have the overall ubiquitin fold.

De novo four-helix bundles

A great deal of attention has been given to the design of coiled-coils and four-helix bundle proteins over about the last 15 years [28]. We will shortly discuss two efforts in the combinatorial design and redesign of four-helix bundles, but it is worth noting some of the lessons from the *de novo* design of these types of proteins, which have shed considerable light on the problem of protein stability and conformational specificity [30]. The α_2 series of peptides were designed to form dimeric four-helix bundles, like the protein rop. The early α_2 B peptide, composed of two identical helices consisting of Leu, Glu and Lys, formed a very stable, helical dimer, but was topologically dynamic and molten globule-like [100]. In the next generation design, α_2 C, the degeneracy of the helices was broken by replacing half of the Leu with aromatic and β -branched side chains that have considerable stereochemical preferences, resulting in a molecule that exhibits cooperative thermal denaturation [101]. However, it was not until the α_2 D design that a truly native-like protein was achieved, exhibiting sharp, disperse NMR spectra and resistance to hydrophobic dye binding, by changing two apolar residues to polar residues and adding an interfacial His residue [102]. This apoprotein (it can also bind Zn^{2+}) showed considerable conforma-

tional specificity despite being of lower overall stability than α_2 B, illustrating the importance of specific polar interactions and 'negative' elements to discourage the population of energetically near conformations or topologies. However, like the rop(A31P) mutant described above, this protein was found upon crystallization to be in the 'bisecting-U' conformation [99]. The DeGrado group's design paradigm is 'hierarchical', in that it first considers gross effects such as binary patterning (i.e. defining an 'inside' and an 'outside') and secondary-structural propensity of residues, and then fine-tunes packing complementarity, specific polar interactions and negative elements.

The Hecht group has taken a combinatorial approach to the problem of four-helix bundles by designing single-chain proteins in which nearly every position is encoded by a degenerate codon that results in hydrophobic, hydrophilic or turn residues. The first-generation library (Fig. 6A) consisted of 74 amino acids with four 14 residue randomized amphipathic helices, three turns of defined sequence (GPDSG, GPSGG and GPRSG), an initial Met-Gly and terminal Arg. Remarkably, 29 of 48 randomly selected clones expressed soluble protein (the only screening step applied here); most that were analyzed were found to be helical, globular and monomeric. Several possessed some native-like characteristics, such as cooperative denaturation, resistance to hydrophobic dye binding, and reasonable NMR spectra, although most were molten globule-like [103,104].

Hecht speculated that the helices might not be long enough for native-like behavior because most natural helical bundles are composed of helices with more than 20 residues. Therefore, a second-generation library was created by modifying and extending one of the molten globules from the initial library (Fig. 6A). A tyrosine was inserted at position 2 for quantitation and to prevent demethionylation; prolines were removed from the turns to prevent problems with *cis/trans* isomerization; and the N-cap, C-cap and half the turn residues were encoded with polar degenerate codons (N-caps were restricted to Asn, Thr

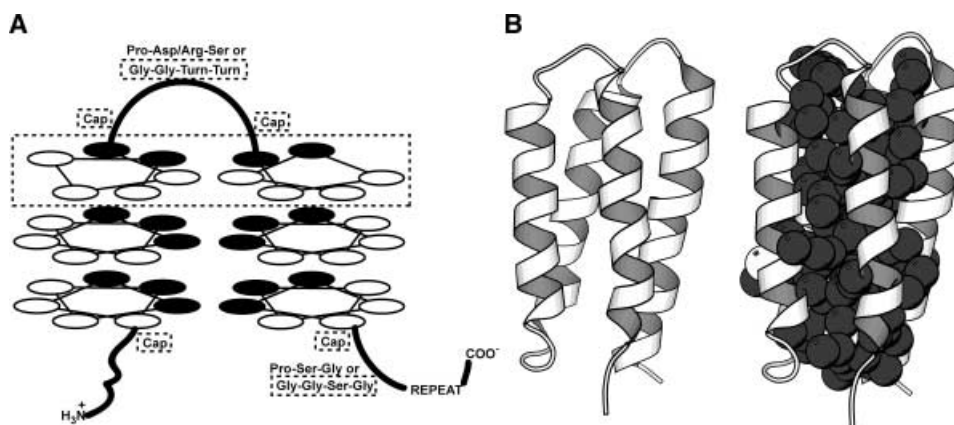


Fig. 6. Four-helix bundles from binary-patterned combinatorial libraries. (A) Schematic representation of the Hecht group's first and second generation libraries (dashed boxes indicate new or altered features in the second generation library). The original library consisted of four 14 residue helices connected by glycine N- and C-caps with Pro-X-X linkers (X varied with the position of the turn; see diagram). Hydrophobic positions are indicated by filled circles; hydrophilic residues are indicated by empty circles. The second generation library extended the helices to 20 residues each with an additional polar position in the extensions; added more reasonable N- and C-capping residues (polar residues); and replaced the Pro-X-X turns with flexible Gly-Gly-X-X sequences. Only half of the sequence is diagrammed, as it repeats to form the four-helix bundle. (B) Structure of one of the second generation variants. On the right, the nonpolar residues are rendered as spheres. Rendered with MOLSCRIPT from PDB entry 1P68.

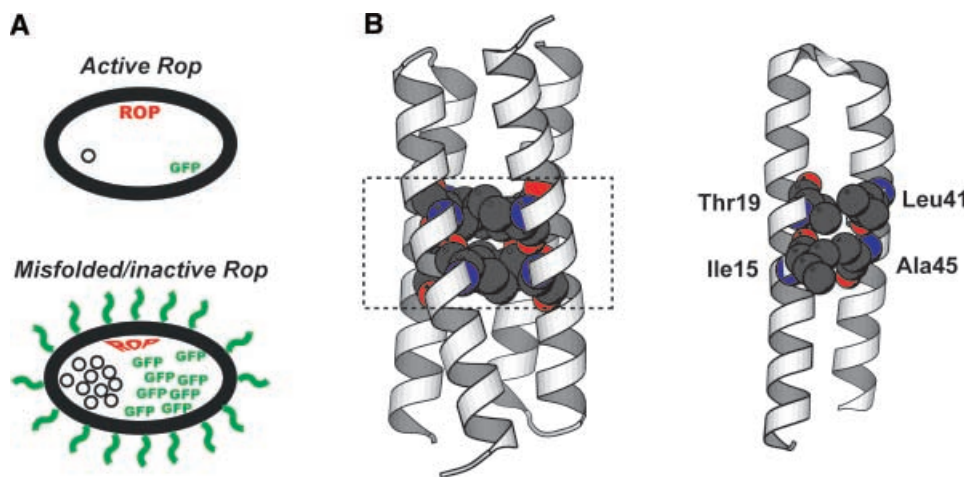


Fig. 7. Screening for structured rop variants. (A) Rop modulates the copy number of ColE1 plasmids. A cell-based screen for rop activity was created by expressing green fluorescent protein from a ColE1 plasmid, wherein rop activity is reported by cellular fluorescence. By expressing green fluorescent protein from the *araBAD* promoter, the phenotype of the screen can be reversed, such that cells with active rop are fluorescent (not shown). (B) The Nnk₄-2 rop library was created by randomization of the two central 'layers' of the rop core. On the right, the four residues randomized in the monomer are highlighted. Rendered with MOLSCRIPT from PDB entry 1ROP.

and Ser). Most significantly, the resultant proteins were extended to 102 residues by adding six randomized residues to each helix in the binary pattern. Five arbitrary library members were characterized, and all were helical, monomeric and stable. NOESY, ¹⁵N-¹H HSQC and ¹³C-¹H HSQC NMR spectra indicated that four of the five proteins had well-ordered and persistent main-chain and sidechain structure. The best of these was shown to have a substantial enthalpic contribution to its thermal denaturation, and the solution structure has subsequently been solved (Fig. 6B) [105]. This lends considerable credence to the view that proteins can achieve native-like properties without specifying jigsaw-puzzle like interactions, but it is less clear if anything was special about the arbitrary scaffold for the second-generation library or if it was typical. It would be interesting to repeat the experiment, randomizing all the appropriate positions in the second-generation library. Likewise, it would be interesting to know the importance of the turn and capping residues that were additionally randomized here. The Hecht group is pursuing experiments to probe both of these questions (M.H. Hecht, Princeton University, Princeton, NJ, personal communication).

Rop

For the last decade, the Regan lab has studied the structure, function, stability and folding of the four-helix bundle protein rop. Rop is an excellent model system for understanding protein structure and stability: it can be expressed in large quantities, it is highly soluble, its crystal [106] and solution [107] structures have been solved, and the residues required for function (RNA binding) have been identified [108]. Moreover, it is an exceedingly simple, regular structure, which permits a rational understanding of the effects of mutation [96,109] in a way that is less straight forward in other more structurally complex model proteins like λ repressor or barnase. This, in turn, permits the rational construction of variant libraries.

Until recently, however, one of the most significant drawbacks of the rop system was that it was difficult to assay for its activity with individual protein variants, and it was much more difficult to screen large numbers of rop variants for activity. As mentioned above, we have developed a robust screen for rop activity, which now permits us to interrogate large libraries of rop variants (Fig. 7A) [72]. (Three other screens for rop function have been reported, but not widely used, including one quite recently [110–112].) We are interested in screening libraries of rop variants that will permit a statistical analysis of sequences that are compatible with rop structure and stability, making it possible to rigorously examine the design principles that have evolved from *de novo* and systematic studies.

The first application of this screen was to assess the *in vivo* activity of systematically designed core mutants [72]. Surprisingly, there was not a one-to-one correspondence of the stability of the proteins or their ability to bind small hairpin RNAs *in vitro* to *in vivo* activity. While unstable variants that did not bind RNA *in vitro* were inactive, only one stable, RNA binding variant was active, that with the central two 'layers' of the core composed of Ala₂Leu₂. Even a variant with Ala₂Leu₂ in the four central layers was just slightly active *in vivo*. Rop cellular function requires the binding of much larger ColE1 origin-derived RNAs than those used *in vitro*, and the redesigned rop variants are known to have considerably faster kinetics of association and dissociation. This suggests that the screen is an exquisite assay for the functional and structural constraints on a protein *in vivo*.

We have subsequently applied this screen to a library of rop variants in which the two central layers (four residues in the monomer) of the core were completely randomized using the codon NNK to encode all 20 amino acids (Fig. 7B; T. J. Magliery & L. Regan, unpublished observation). The amino acids elicited at these positions in active variants were not especially influenced by helical propensity, and the observed residues were nearly the same as those seen

in the first Lim & Sauer experiment with the entirely different architecture, λ repressor (the hydrophobics except for Trp were observed, Ser, Thr, Cys and His but not charged residues or carboxamides were seen). Surprisingly, the sum of the van der Waals volumes of the sidechains in each layer varied substantially, from 160 Å³ to 320 Å³. This represents over eight methylene groups of variation (for example, both LAAL and LMLL are active, where these represent the residues at positions 15, 19, 41 and 45). Wild type rop contains a Thr at position 19, and a large number of the variants had Ser or Thr at positions 19 or 45. On the other hand, there was virtually no pattern to the sizes or identities of the residues at individual positions in variants that contained all hydrophobic amino acids at these positions. Some of the selected proteins have relatively high T_m s and cooperative melting transitions, but some have flatter thermal transitions and less well-dispersed ¹H-¹⁵N HSQC NMR spectra, suggesting more molten-globule like molecules.

We are in the process of analyzing a larger number of these variants in more depth, including crystallographically. However, we believe that the large variation in core size is probably related to the fact that this is a protein dimerization interface, wherein the monomers can translate with respect to each other to accommodate different core volumes. We are also intrigued that the all-hydrophobic and alcohol-containing variants might represent two different regimes of protein stability, wherein geometry becomes more important for hydrogen bonding (jigsaw-puzzle behavior) but is swamped out by hydrophobic partitioning in the absence of polar sidechains (oil-droplet behavior). Further libraries have been created to explore these issues, and we will also expand the scope of these studies to larger portions of the core (T. J. Magliery & L. Regan, unpublished observation). Due to the simplicity of the rop structure, we hope that statistical analyses of such libraries will inform both *de novo* design of helical bundles and provide rigorous data on principles that apply more generally.

Conclusion

We find on survey of the literature that we are limited in our ability to analyze large collections of protein variants in two distinct ways: direct analysis of biophysical properties is difficult to carry out on large numbers of proteins, and inferential methods of screening are complicated by the assumptions on which they are predicated. Every screen has trivial positives and negatives associated with it. *In vivo*, certain proteins will overexpress or fail to express (or display) for reasons that are often difficult to identify (e.g. transcription or degradation) but not directly related to stability. Selections based on function, even with a function as simple as binding, may enforce sequence constraints that are not evident from crystallographic structures or alanine scanning for functional residues, such as distantly coupled residues. Even *in vitro* display methods, where it is possible to directly address biophysical properties, depend upon translation and are complicated by nonobvious effects of the molecule on which they are displayed, such as solubility. The lesson of biological selection is that 'you get what you select for', not what

you hope you are selecting for, and the degree to which these correspond is always up for debate.

On the other hand, our exploration of protein sequence space is desperately meagre at the moment, and methods that allow us to triage large libraries and characterize a reasonable number of interesting molecules are required. The extent to which translation and solubility impact combinatorial experiments can be addressed by in part by judicious library design, and restraints imposed by function can be minimized by choosing simple functions, like ligand binding. But the chief merit of these approaches, despite their difficulties, is that modern techniques of library construction, DNA preparation, sequencing and, increasingly, protein purification and *in vitro* analysis, give us the ability to examine many more variants than was possible even a decade ago when Lim & Sauer carried out their first experiments in this field. Innovations from genomic and proteomic approaches, including robotics and high throughput instrumentation, make this an exciting time to explore protein sequence space, because it will be possible to generate statistically significant results for use in improving parameterization of computational methods. Right now, even the most straightforward questions about protein stability and structure have only rules-of-thumb as answers, but combinatorial approaches will make it possible to add quantitative weight to trends derived from systematic studies. These issues are critical for making better protein-based therapeutics and treating diseases that result from protein mutation; they lie at the center of our understanding of biophysical phenomena; and they are increasingly accessible with the library-scale methods presented here.

Acknowledgements

T. J. M. is an NIH Postdoctoral Fellow (GM065750-02). Work on rop was supported by a grant to L. R. from the NIH (GM49146-09).

References

1. Bishop, B., Koay, D.C., Sartorelli, A.C. & Regan, L. (2001) Reengineering granulocyte colony-stimulating factor for enhanced stability. *J. Biol. Chem.* **276**, 33465–33470.
2. Bullock, A.N. & Fersht, A.R. (2001) Rescuing the function of mutant p53. *Nat. Rev. Cancer* **1**, 68–76.
3. Graddis, T.J., Remmele, R.L. Jr & McGrew, J.T. (2002) Designing proteins that work using recombinant technologies. *Curr. Pharm. Biotechnol.* **3**, 285–297.
4. Buxbaum, J.N. (2003) Diseases of protein conformation: what do *in vitro* experiments tell us about *in vivo* diseases? *Trends Biochem. Sci.* **28**, 585–592.
5. Anfinsen, C.B. (1972) The formation and stabilization of protein structure. *Biochem. J.* **128**, 737–749.
6. Tanford, C. (1978) The hydrophobic effect and the organization of living matter. *Science* **200**, 1012–1018.
7. Richards, F.M. (1997) Protein stability: still an unsolved problem. *Cell. Mol. Life Sci.* **53**, 790–802.
8. Dahiyat, B.I. & Mayo, S.L. (1997) *De novo* protein design: fully automated sequence selection. *Science* **278**, 82–87.
9. Harbury, P.B., Plecs, J.J., Tidor, B., Alber, T. & Kim, P.S. (1998) High-resolution protein design with backbone freedom. *Science* **282**, 1462–1467.

10. Guerois, R., Nielsen, J.E. & Serrano, L. (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.* **320**, 369–387.
11. Mendes, J., Guerois, R. & Serrano, L. (2002) Energy estimation in protein design. *Curr. Opin. Struct. Biol.* **12**, 441–446.
12. Kamtekar, S., Schiffer, J.M., Xiong, H., Babik, J.M. & Hecht, M.H. (1993) Protein design by binary patterning of polar and nonpolar amino acids. *Science* **262**, 1680–1685.
13. Ponder, J.W. & Richards, F.M. (1987) Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.* **193**, 775–791.
14. Wolf, E. & Kim, P.S. (1999) Combinatorial codons: a computer program to approximate amino acid probabilities with biased nucleotide usage. *Protein Sci.* **8**, 680–688.
15. Sonddek, J. & Shortle, D. (1992) A general strategy for random insertion and substitution mutagenesis: substoichiometric coupling of trinucleotide phosphoramidites. *Proc. Natl Acad. Sci. USA* **89**, 3581–3585.
16. Arndt, K.M., Pelletier, J.N., Muller, K.M., Alber, T., Michnick, S.W. & Pluckthun, A. (2000) A heterodimeric coiled-coil peptide pair selected *in vivo* from a designed library-versus-library ensemble. *J. Mol. Biol.* **295**, 627–639.
17. Magliery, T.J., Pastrnak, M., Anderson, J.C., Santoro, S.W., Herberich, B., Meggers, E., Wang, L. & Schultz, P.G. (2003) *In vitro* tools and *in vivo* engineering: incorporation of unnatural amino acids into proteins. In *Translation Mechanisms* (Lapointe, J. & Brakier-Gingras, L., eds), pp. 95–114. Landes Bioscience, Georgetown, TX.
18. Lin, H.N. & Cornish, V.W. (2002) Screening and selection methods for large-scale analysis of protein function. *Angew. Chem. Int. Ed.* **41**, 4403–4425.
19. Yanofsky, C., Henning, U., Helinski, D. & Carlton, B. (1963) Mutational alteration of protein structure. *Fed. Proc.* **22**, 75–79.
20. Murgola, E.J. & Yanofsky, C. (1974) Selection for new amino acids at position 211 of the tryptophan synthetase alpha chain of *Escherichia coli*. *J. Mol. Biol.* **86**, 775–784.
21. Tweedy, N.B., Hurle, M.R., Chrnyk, B.A. & Matthews, C.R. (1990) Multiple replacements at position 211 in the alpha subunit of tryptophan synthase as a probe of the folding unit association reaction. *Biochemistry* **29**, 1539–1545.
22. Kleina, L.G. & Miller, J.H. (1990) Genetic studies of the lac repressor. XIII. Extensive amino acid replacements generated by the use of natural and synthetic nonsense suppressors. *J. Mol. Biol.* **212**, 295–318.
23. Hecht, M.H., Nelson, H.C. & Sauer, R.T. (1983) Mutations in lambda repressor's amino-terminal domain: implications for protein stability and DNA binding. *Proc. Natl Acad. Sci. USA* **80**, 2676–2680.
24. Hecht, M.H., Hahir, K.M., Nelson, H.C., Sturtevant, J.M. & Sauer, R.T. (1985) Increasing and decreasing protein stability: effects of revertant substitutions on the thermal denaturation of phage lambda repressor. *J. Cell. Biochem.* **29**, 217–224.
25. Baldwin, E.P. & Matthews, B.W. (1994) Core-packing constraints, hydrophobicity and protein design. *Curr. Opin. Biotechnol.* **5**, 396–402.
26. Matthews, B.W. (1995) Studies on protein stability with T4 lysozyme. *Adv. Protein Chem.* **46**, 249–278.
27. Lazar, G.A. & Handel, T.M. (1998) Hydrophobic core packing and protein design. *Curr. Opin. Chem. Biol.* **2**, 675–679.
28. DeGrado, W.F., Summa, C.M., Pavone, V., Nastri, F. & Lombardi, A. (1999) *De novo* design and structural characterization of proteins and metalloproteins. *Annu. Rev. Biochem.* **68**, 779–819.
29. Regan, L. (1999) Protein redesign. *Curr. Opin. Struct. Biol.* **9**, 494–499.
30. Hill, R.B., Raleigh, D.P., Lombardi, A. & DeGrado, W.F. (2000) *De novo* design of helical bundles as models for understanding protein folding and function. *Acc. Chem. Res.* **33**, 745–754.
31. Woolfson, D.N. (2001) Core-directed protein design. *Curr. Opin. Struct. Biol.* **11**, 464–471.
32. Stemmer, W.P. (1994) Rapid evolution of a protein *in vitro* by DNA shuffling. *Nature* **370**, 389–391.
33. Murakami, H., Hohsaka, T. & Sisido, M. (2002) Random insertion and deletion of arbitrary number of bases for codon-based random mutation of DNAs. *Nat. Biotechnol.* **20**, 76–81.
34. Wang, L., Brock, A., Herberich, B. & Schultz, P.G. (2001) Expanding the genetic code of *Escherichia coli*. *Science* **292**, 498–500.
35. Frankel, A. & Roberts, R.W. (2003) *In vitro* selection for sense codon suppression. *RNA* **9**, 780–786.
36. Hanes, J. & Pluckthun, A. (1997) *In vitro* selection and evolution of functional proteins by using ribosome display. *Proc. Natl Acad. Sci. USA* **94**, 4937–4942.
37. Roberts, R.W. & Szostak, J.W. (1997) RNA-peptide fusions for the *in vitro* selection of peptides and proteins. *Proc. Natl Acad. Sci. USA* **94**, 12297–12302.
38. Stevens, R.C. (2000) High-throughput protein crystallization. *Curr. Opin. Struct. Biol.* **10**, 558–563.
39. Kuhn, P., Wilson, K., Patch, M.G. & Stevens, R.C. (2002) The genesis of high-throughput structure-based drug discovery using protein crystallography. *Curr. Opin. Chem. Biol.* **6**, 704–710.
40. Weber, P.C. & Salemme, F.R. (2003) Applications of calorimetric methods to drug discovery and the study of protein interactions. *Curr. Opin. Struct. Biol.* **13**, 115–121.
41. Uversky, V.N. (2002) Natively unfolded proteins: a point where biology waits for physics. *Protein Sci.* **11**, 739–756.
42. Gronenborn, A.M., Frank, M.K. & Clore, G.M. (1996) Core mutants of the immunoglobulin binding domain of streptococcal protein G: stability and structural integrity. *FEBS Lett.* **398**, 312–316.
43. Wei, Y., Liu, T., Sazinsky, S.L., Moffet, D.A., Pelczar, I. & Hecht, M.H. (2003) Stably folded *de novo* proteins from a designed combinatorial library. *Protein Sci.* **12**, 92–102.
44. Roy, S., Helmer, K.J. & Hecht, M.H. (1997) Detecting native-like properties in combinatorial libraries of *de novo* proteins. *Fold. Des.* **2**, 89–92.
45. Rosenbaum, D.M., Roy, S. & Hecht, M.H. (1999) Screening combinatorial libraries of *de novo* proteins by hydrogen-deuterium exchange and electrospray mass spectrometry. *J. Am. Chem. Soc.* **121**, 9509–9513.
46. Waldo, G.S., Standish, B.M., Berendzen, J. & Terwilliger, T.C. (1999) Rapid protein-folding assay using green fluorescent protein. *Nat. Biotechnol.* **17**, 691–695.
47. Waldo, G.S. (2003) Improving protein folding efficiency by directed evolution using the GFP folding reporter. *Methods Mol. Biol.* **230**, 343–359.
48. Waldo, G.S. (2003) Genetic screens and directed evolution for protein solubility. *Curr. Opin. Chem. Biol.* **7**, 33–38.
49. Maxwell, K.L., Mittermaier, A.K., Forman-Kay, J.D. & Davidson, A.R. (1999) A simple *in vivo* assay for increased protein solubility. *Protein Sci.* **8**, 1908–1911.
50. Wigley, W.C., Stidham, R.D., Smith, N.M., Hunt, J.F. & Thomas, P.J. (2001) Protein solubility and folding monitored *in vivo* by structural complementation of a genetic marker protein. *Nat. Biotechnol.* **19**, 131–136.
51. der Maur, A.A., Escher, D. & Barberis, A. (2001) Antigen-independent selection of stable intracellular single-chain antibodies. *FEBS Lett.* **508**, 407–412.
52. Kelemen, B.R., Klink, T.A., Behlke, M.A., Eubanks, S.R., Leland, P.A. & Raines, R.T. (1999) Hypersensitive substrate for ribonucleases. *Nucleic Acids Res.* **27**, 3696–3701.

53. Lesley, S.A., Graziano, J., Cho, C.Y., Knuth, M.W. & Klock, H.E. (2002) Gene expression response to misfolded protein as a screen for soluble recombinant protein. *Protein Eng.* **15**, 153–160.
54. Hagihara, Y. & Kim, P.S. (2002) Toward development of a screen to identify randomly encoded, foldable sequences. *Proc. Natl Acad. Sci. USA* **99**, 6619–6624.
55. Bowie, J.U. & Sauer, R.T. (1989) Identification of C-terminal extensions that protect proteins from intracellular proteolysis. *J. Biol. Chem.* **264**, 7596–7602.
56. Parsell, D.A. & Sauer, R.T. (1989) The structural stability of a protein is an important determinant of its proteolytic susceptibility in *Escherichia coli*. *J. Biol. Chem.* **264**, 7590–7595.
57. Finucane, M.D., Tuna, M., Lees, J.H. & Woolfson, D.N. (1999) Core-directed protein design. I. An experimental method for selecting stable proteins from combinatorial libraries. *Biochemistry* **38**, 11604–11612.
58. Kristensen, P. & Winter, G. (1998) Proteolytic selection for protein folding using filamentous bacteriophages. *Fold. Des.* **3**, 321–328.
59. Sieber, V., Pluckthun, A. & Schmid, F.X. (1998) Selecting proteins with improved stability by a phage-based method. *Nat. Biotechnol.* **16**, 955–960.
60. Chu, R., Takei, J., Knowlton, J.R., Andrykovitch, M., Pei, W., Kajava, A.V., Steinbach, P.J., Ji, X. & Bai, Y. (2002) Redesign of a four-helix bundle protein by phage display coupled with proteolysis and structural characterization by NMR and X-ray crystallography. *J. Mol. Biol.* **323**, 253–262.
- 60a. Bai, Y. & Feng, H. (2004) Selection of stably folded proteins by phage-display with proteolysis. *Eur. J. Biochem.* **271**, 1609–1614.
61. Matsuura, T. & Pluckthun, A. (2003) Selection based on the folding properties of proteins with ribosome display. *FEBS Lett.* **539**, 24–28.
- 61a. Kotz, J.D., Bond, C.J. & Cochran, A.G. (2004) Phage-display as a tool for quantifying protein stability determinants. *Eur. J. Biochem.* **271**, 1623–1629.
- 61b. Watters, A.L. & Baker, D. (2004) Searching for folded proteins *in vitro* and *in silico*. *Eur. J. Biochem.* **271**, 1615–1622.
62. Distefano, M.D., Zhong, A. & Cochran, A.G. (2002) Quantifying beta-sheet stability by phage display. *J. Mol. Biol.* **322**, 179–188.
63. Gu, H., Yi, Q., Bray, S.T., Riddle, D.S., Shiau, A.K. & Baker, D. (1995) A phage display system for studying the sequence determinants of protein folding. *Protein Sci.* **4**, 1108–1117.
64. Riddle, D.S., Santiago, J.V., Bray-Hall, S.T., Doshi, N., Grantcharova, V.P., Yi, Q. & Baker, D. (1997) Functional rapidly folding proteins from simplified amino acid sequences. *Nat. Struct. Biol.* **4**, 805–809.
65. Kim, D.E., Gu, H. & Baker, D. (1998) The sequences of small proteins are not extensively optimized for rapid folding by natural selection. *Proc. Natl Acad. Sci. USA* **95**, 4982–4986.
66. Yi, Q., Rajagopal, P., Klevit, R.E. & Baker, D. (2003) Structural and kinetic characterization of the simplified SH3 domain FPI. *Protein Sci.* **12**, 776–783.
67. Minard, P., Scalley-Kim, M., Watters, A. & Baker, D. (2001) A 'loop entropy reduction' phage-display selection for folded amino acid sequences. *Protein Sci.* **10**, 129–134.
68. Scalley-Kim, M., Minard, P. & Baker, D. (2003) Low free energy cost of very long loop insertions in proteins. *Protein Sci.* **12**, 197–206.
69. Keefe, A.D. & Szostak, J.W. (2001) Functional proteins from a random-sequence library. *Nature* **410**, 715–718.
70. Lim, W.A. & Sauer, R.T. (1989) Alternative packing arrangements in the hydrophobic core of lambda repressor. *Nature* **339**, 31–36.
71. Lim, W.A. & Sauer, R.T. (1991) The role of internal packing interactions in determining the structure and stability of a protein. *J. Mol. Biol.* **219**, 359–376.
72. Magliery, T.J. & Regan, L. (2004) A cell-based screen for function of the four-helix bundle protein Rop: a new tool for combinatorial experiments in biophysics. *Protein Eng. Des. Select.* **17**, 77–83.
73. Axe, D.D., Foster, N.W. & Fersht, A.R. (1996) Active barnase variants with completely random hydrophobic cores. *Proc. Natl Acad. Sci. USA* **93**, 5590–5594.
74. MacBeath, G., Kast, P. & Hilvert, D. (1998) Redesigning enzyme topology by directed evolution. *Science* **279**, 1958–1961.
75. Taylor, S.V., Kast, P. & Hilvert, D. (2001) Investigating and engineering enzymes by genetic selection. *Angew. Chem. Int. Ed.* **40**, 3311–3335.
76. Taylor, S.V., Walter, K.U., Kast, P. & Hilvert, D. (2001) Searching sequence space for protein catalysts. *Proc. Natl Acad. Sci. USA* **98**, 10596–10601.
- 76a. Woycechowsky, K.J. & Hilvert, D. (2004) Deciphering enzymes. Genetic selection as a probe of structure and mechanism. *Eur. J. Biochem.* **271**, 1630–1637.
77. Silverman, J.A., Balakrishnan, R. & Harbury, P.B. (2001) Reverse engineering the (beta/alpha) 8 barrel fold. *Proc. Natl Acad. Sci. USA* **98**, 3092–3097.
78. Gassner, N.C., Baase, W.A. & Matthews, B.W. (1996) A test of the 'jigsaw puzzle' model for protein folding by multiple methionine substitutions within the core of T4 lysozyme. *Proc. Natl Acad. Sci. USA* **93**, 12155–12158.
79. Desjarlais, J.R. & Handel, T.M. (1995) *De novo* design of the hydrophobic cores of proteins. *Protein Sci.* **4**, 2006–2018.
80. Dahiyat, B.I. & Mayo, S.L. (1996) Protein design automation. *Protein Sci.* **5**, 895–903.
81. Looger, L.L. & Hellinga, H.W. (2001) Generalized dead-end elimination algorithms make large-scale protein side-chain structure prediction tractable: implications for protein design and structural genomics. *J. Mol. Biol.* **307**, 429–445.
82. Looger, L.L., Dwyer, M.A., Smith, J.J. & Hellinga, H.W. (2003) Computational design of receptor and sensor proteins with novel functions. *Nature* **423**, 185–190.
83. Bowie, J.U., Reidhaar-Olson, J.F., Lim, W.A. & Sauer, R.T. (1990) Deciphering the message in protein sequences: tolerance to amino acid substitutions. *Science* **247**, 1306–1310.
84. Richards, F.M. & Lim, W.A. (1993) An analysis of packing in the protein folding problem. *Q. Rev. Biophys.* **26**, 423–498.
85. Cordes, M.H., Davidson, A.R. & Sauer, R.T. (1996) Sequence space, folding and protein design. *Curr. Opin. Struct. Biol.* **6**, 3–10.
86. Sauer, R.T. (1996) Protein folding from a combinatorial perspective. *Fold. Des.* **1**, R27–R30.
87. Saven, J.G. (2002) Combinatorial protein design. *Curr. Opin. Struct. Biol.* **12**, 453–458.
88. Lim, W.A., Farruggio, D.C. & Sauer, R.T. (1992) Structural and energetic consequences of disruptive mutations in a protein core. *Biochemistry* **31**, 4324–4333.
89. Lim, W.A., Hodel, A., Sauer, R.T. & Richards, F.M. (1994) The crystal structure of a mutant protein with altered but improved hydrophobic core packing. *Proc. Natl Acad. Sci. USA* **91**, 423–427.
90. Buckle, A.M., Henrick, K. & Fersht, A.R. (1993) Crystal structural analysis of mutations in the hydrophobic cores of barnase. *J. Mol. Biol.* **234**, 847–860.
91. Finucane, M.D. & Woolfson, D.N. (1999) Core-directed protein design. II. Rescue of a multiply mutated and destabilized variant of ubiquitin. *Biochemistry* **38**, 11613–11623.
92. Lazar, G.A., Desjarlais, J.R. & Handel, T.M. (1997) *De novo* design of the hydrophobic core of ubiquitin. *Protein Sci.* **6**, 1167–1178.

93. Frank, M.K., Dyda, F., Dobrodumov, A. & Gronenborn, A.M. (2002) Core mutations switch monomeric protein GB1 into an intertwined tetramer. *Nat. Struct. Biol.* **9**, 877–885.
94. Byeon, I.J., Louis, J.M. & Gronenborn, A.M. (2003) A protein contortionist: core mutations of GB1 that induce dimerization and domain swapping. *J. Mol. Biol.* **333**, 141–152.
95. Ramirez-Alvarado, M. & Regan, L. (2002) Does the location of a mutation determine the ability to form amyloid fibrils? *J. Mol. Biol.* **323**, 17–22.
96. Munson, M., Balasubramanian, S., Fleming, K.G., Nagi, A.D., O'Brien, R., Sturtevant, J.M. & Regan, L. (1996) What makes a protein a protein? Hydrophobic core designs that specify stability and structural properties. *Protein Sci.* **5**, 1584–1593.
97. Willis, M.A., Bishop, B., Regan, L. & Brunger, A.T. (2000) Dramatic structural and thermodynamic consequences of repacking a protein's hydrophobic core. *Structure* **8**, 1319–1328.
98. Glykos, N.M., Cesareni, G. & Kokkinidis, M. (1999) Protein plasticity to the extreme: changing the topology of a 4-alpha-helical bundle with a single amino acid substitution. *Structure* **7**, 597–603.
99. Hill, R.B. & DeGrado, W.F. (1998) Solution structure of alpha2D, a natively like *de novo* designed protein. *J. Am. Chem. Soc.* **120**, 1138–1145.
100. Ho, S.P. & DeGrado, W.F. (1987) Design of a 4-helix bundle protein: synthesis of peptides which self-associate into a helical protein. *J. Am. Chem. Soc.* **109**, 6751–6758.
101. Raleigh, D.P. & DeGrado, W.F. (1992) A *de novo* designed protein shows a thermal induced transition from a native to a molten globule-like state. *J. Am. Chem. Soc.* **114**, 10079–10081.
102. Raleigh, D.P., Betz, S.F. & DeGrado, W.F. (1995) A *de novo* designed protein mimics the native state of natural proteins. *J. Am. Chem. Soc.* **117**, 7558–7559.
103. Roy, S., Ratnaswamy, G., Boice, J.A., Fairman, R., McLendon, G. & Hecht, M.H. (1997) A protein designed by binary patterning of polar and nonpolar amino acids displays native-like properties. *J. Am. Chem. Soc.* **119**, 5302–5306.
104. Roy, S. & Hecht, M.H. (2000) Cooperative thermal denaturation of proteins designed by binary patterning of polar and nonpolar amino acids. *Biochemistry* **39**, 4603–4607.
105. Wei, Y., Kim, S., Fela, D., Baum, J. & Hecht, M.H. (2003) Solution structure of a *de novo* protein from a designed combinatorial library. *Proc. Natl Acad. Sci. USA* **100**, 13270–13273.
106. Banner, D.W., Kokkinidis, M. & Tsernoglou, D. (1987) Structure of the ColE1 rop protein at 1.7 Å resolution. *J. Mol. Biol.* **196**, 657–675.
107. Eberle, W., Pastore, A., Sander, C. & Rosch, P. (1991) The structure of ColE1 rop in solution. *J. Biomol. NMR* **1**, 71–82.
108. Predki, P.F., Nayak, L.M., Gottlieb, M.B. & Regan, L. (1995) Dissecting RNA–protein interactions: RNA–RNA recognition by Rop. *Cell* **80**, 41–50.
109. Munson, M., O'Brien, R., Sturtevant, J.M. & Regan, L. (1994) Redesigning the hydrophobic core of a four-helix-bundle protein. *Protein Sci.* **3**, 2015–2022.
110. Cesareni, G., Muesing, M.A. & Polisky, B. (1982) Control of ColE1 DNA replication: the rop gene product negatively affects transcription from the replication primer promoter. *Proc. Natl Acad. Sci. USA* **79**, 6313–6317.
111. Castagnoli, L., Vetriani, C. & Cesareni, G. (1994) Linking an easily detectable phenotype to the folding of a common structural motif. Selection of rare turn mutations that prevent the folding of Rop. *J. Mol. Biol.* **237**, 378–387.
112. Christ, D. & Winter, G. (2003) Identification of functional similarities between proteins using directed evolution. *Proc. Natl Acad. Sci. USA* **100**, 13202–13206.
113. Kraulis, P.J. (1991) MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **24**, 946–950.