Eur. J. Biochem. **271**, 1615–1622 (2004) © FEBS 2004 doi:10.1111/j.1432-1033.2004.04072.x

MINIREVIEW

Searching for folded proteins *in vitro* and *in silico*

Alexander L. Watters[1] and David Baker[2]

[1]*Molecular and Cellular Biology Program and* [2]*Department of Biochemistry and Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA*

Understanding the sequence determinants of protein structure, stability and folding is critical for understanding how natural proteins have evolved and how proteins can be engineered to perform novel functions. The complexity of the protein folding problem requires the ability to search large volumes of sequence space for proteins with specific structural or functional characteristics. Here we describe our efforts to identify novel proteins using a phage-display selection strategy from a 'mini-exon' shuffling library generated from the yeast genome and from completely random sequence libraries, and compare the results to recent successes in generating novel proteins using *in silico* protein design.

Keywords: loop entropy; mini-exon shuffling; phage-display; protein evolution; random sequences; simplified proteins.

Introduction

To probe the sequence determinants of protein folding and to investigate the selection pressures which have shaped protein evolution it is desirable to generate novel proteins in the laboratory and to study their biophysical characteristics. There are two powerful approaches to generating such artificial proteins: combinatorial library selections and computational protein design. In this paper, we describe our results using both applications and present the results of an investigation of protein evolution by 'mini-exon' shuffling.

Phage-display

Phage-display technology is an excellent method for selecting functional binding mutants from large peptide or protein libraries [1]. This technology utilizes the ability to express foreign proteins on the outside of phage particles as fusions to the phage coat proteins. Phage expressing fusion proteins with the desired binding characteristics can then be readily selected from a large pool of potential binders. The sequence of positive clones can easily be determined by sequencing the DNA contained in the phage particle. In the experiments discussed below, all displays used the major coat protein (gene 8) of the M13 filamentous phage [2].

Correspondence to D. Baker, Department of Biochemistry and Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA. Fax: + 1 206 685 1792, Tel.: + 1 206 543 1295, E-mail: dabaker@u.washington.edu

Abbreviations: CspA, cold shock protein A from *E. coli*.

(Received 5 January 2004, revised 1 March 2004, accepted 5 March 2004)

Selection for novel sequences of natural occurring proteins

As a first step to understanding the sequence dependence of protein folding, stability and structure, we sought to identify either randomized or simplified sequences that fold to the same structure. Correct formation of the native binding interface for a protein usually requires the precise three dimensional arrangement of specific, nonlocal amino acid positions. This requirement allows selection of functionally active mutants from large libraries, yielding proteins with an overall structure similar to the wild type. In our initial studies of sequence effects on protein folding we used the 62 residue B1 domain of protein L, an α/β protein consisting of a four stranded β-sheet with a single helix packed on one side. Protein L is ideal for this study as it has a well characterized binding affinity for the light chain of IgG, lacks disulfide bonds and does not require cofactors for folding [3–5]. The sequences of strands 1, 2, 4 and the α-helix (excluding residues responsible for binding IgG) as well as both turns, could be independently mutated and yet still yield folded and functionally active variants (strand 3 was not mutated). The largest number of amino acids changed in a single functional variant was 11 (all in the helix) [6,7].

The results of the protein L studies gave us a better understanding of the evolutionary pressures on protein stability and folding. All of the mutants characterized showed lower stability than wild type protein L, however, roughly half of the mutants folded faster than wild type [7]. These results suggest evolution has selected for stability, but not fast folding. Instead, the ability to fold seems to be a consequence of possessing a stable unique native structure; interactions stabilizing the native structure also stabilize the transition state. This provides support for computational folding models that consider only native contacts when evaluating possible folding trajectories [8–11]. This lack of selection for folding rates has been further supported by our studies on the src SH3 domain (see below).

Selection for simplified proteins

The evolution of the genetic code plays a fundamental role in the evolution of folded proteins. Hypotheses on the evolution of the genetic code generally assume the initial code used fewer amino acids [12]. For this to be true, it must be possible to encode protein structures using simplified amino acid alphabets. Studies in the early nineties supported this hypothesis by showing that partial or complete mutagenesis of proteins using a subset of the current 20 amino acids still yielded folded proteins. For example, replacing all 10 core residues of T4 lysozyme with methionine yielded a slightly destabilized, but active protein [13]. Regan and coworkers generated folded Rop dimers where all core positions were replaced by either alanine or leucine [14,15]. Regan & DeGrado explicitly designed four helix bundles using only glycine, glutamate, leucine and lysine; arginine and proline were needed in the loops [16,17]. Hecht's lab showed that generation of four helix bundles was possible using only 11 of the amino acids, where the only constraint on the library was the hydrophobic–polar patterning of the sequences [18–20]. Finally, Davidson & Sauer showed folded helical proteins could be generated from random libraries of only three amino acids, where the only constraint was the relative proportion of each amino acid [21,22].

While these studies show it is possible to simplify proteins, they are generally restricted to partial sequence simplifications or all helical proteins. To determine whether this is indicative of a general characteristic of all protein topologies, or whether it simply suggests formation of helical bundles are more common in sequence space than β-sheet containing proteins, we sought to simplify the sequence of the src SH3 domain in a phage-display system. A library containing amino acids I, K, E, A and G produced two SH3 variants in which the positions not involved in binding are comprised mainly of these residues (89% and 90%, respectively) [23]. Structural studies on the 90% simplified protein have shown that it folds into a structure very similar to the wild type structure [24]. The simplified proteins fold faster than the wild type protein, supporting the idea that natural selection has not operated on protein folding rates.

The hypothesis of a simplified amino acid alphabet was further enhanced by recent studies on triosephosphate isomerase. This protein is larger and structurally more complex (β/α barrel) than the previously simplified proteins. Silverman *et al.* found variants of triosephosphate isomerase could be encoded by sequences where 142 of 182 structural positions were simplified to a seven amino acid library (FVLAKEQ), while still maintaining wild type catalytic activity, and biophysical characteristics similar to naturally occurring proteins [25,26].

These studies on the sequence determinants of protein folding helped to clarify the role that sequences play in determining the structure and folding rates of these specific proteins. The experiments, however, were limited to a relatively small subset of sequence/structure space. A broader understanding of structure-sequence relationships from an evolutionary and engineering perspective requires more complex searches.

Phage-display selection of completely novel proteins

Protein structures can be classified into a finite number of protein folds, based on the connectivity and three dimensional arrangements of secondary structure elements [27–29]. While it is unlikely that all proteins within a given fold are evolutionarily related, it is assumed that one member of a fold could evolve into another without going through an unfolded intermediate. Experimental observation suggests the generation of a new fold *de novo* is difficult (see below) and mutating from one fold to another, one residue at a time, without going through an unfolded intermediate may be impossible [30]. How the current structural diversity of individual protein domains arose is thus not clear. These difficulties pose interesting questions relating to the understanding of both protein evolution and protein engineering. From an evolutionary standpoint, if finding a new fold is so difficult, how did nature manage to find the large number currently seen, many of them possibly more than once? From an engineering perspective, is it possible to find folds not seen in nature?

Within the last 10 years several groups have begun to explore the distribution of native-like features in sequence space. Studies of randomly synthesized proteins of 120–140 amino acids showed 10–50% could be expressed and of those 20% were soluble. The number of proteins examined, however, is too low to draw any general conclusions [31–34]. In a more complete study of 80–100 residue random proteins, Davidson & Sauer explored characteristics of a simplified library containing only glutamine (Q), leucine (L) and arginine (R). They estimated that > 5% of the proteins were expressible in *Escherichia coli* and 1–2% showed cooperative unfolding (a characteristic of small folded proteins) and helical secondary structure, but lacked good tertiary packing [21,22]. Few, if any, of the proteins in these screens were truly native-like, suggesting *de novo* formation of proteins may be very difficult. While building proteins from libraries with sequences biased towards the formation of secondary structure has yielded polypeptides with some native characteristics [18–20,35,36], the majority of the successes in these experiments appear to be helical proteins. Notably, the solution structure of a binary patterned, four helix bundle showed an ordered and well packed structure [37].

It is plausible that, during evolution, after an initial set of proteins had formed, new protein architectures could have been generated by recombining super-secondary structural elements. Over the last 20 years, many authors have proposed the idea of new proteins evolving by recombining pieces of nonhomologous proteins [38–40].

These theories have mainly focused on the role introns may have played in the process and not necessarily whether such shuffling has occurred [41]. Experimental and theoretical work on homologous recombination suggests that functionally viable proteins are more likely to be produced when recombination occurs between compact substructures of the protein [38,42,43]. Is it possible to recombine domain substructures of unrelated proteins to yield novel folded proteins? As fragments of naturally occurring proteins have evolved to fold in one context, could this adaptation also make them more likely than a random sequence to fold in

a new context? Appropriately sized fragments of already existing proteins would produce polypeptide fragments containing super-secondary structure motifs. Generating a library containing concatenations of these fragments would allow for the exploration of structure space from both an engineering and evolutionary point of view. Riechmann & Winter examined this possibility by screening a large library of random, 40–50 amino acid segments from the *E. coli* genome fused to the 36 N-terminal residues of the *E. coli* cold shock protein A (CspA) for folded structures. They were able to select fusions that showed native-like characteristics, suggesting that this is a viable method for producing new proteins [44,45]. Large-scale screens using more complex combinations of DNA fragments will more thoroughly explore the possibility of building proteins from nonhomologous protein pieces. To do this we needed to adapt the phage-display system to differentiate between folded and unfolded proteins.

The requirement for a specific binding characteristic in phage-display is a serious restriction, because folded proteins do not have general binding characteristics distinguishing them from unfolded proteins. Two methods to distinguish between folded and unfolded proteins have recently been incorporated into phage-display systems. Multiple groups have developed a technique in which folded polypeptides are selected on the basis of their resistance to proteolysis [46–48]. The second technique, used in our lab and described below, utilizes the difference in conformational-backbone entropy between folded (low) and unfolded (high) proteins [49]. In this system the queried protein is inserted into a loop of another protein (host protein). The basis for the selection is the ability of the host protein to bind its natural ligand. For the host protein to fold it must be able to bring the N- and C-termini of the loop together. Thus, folding of the insert protein in such a way as to bring its N- and C-termini close together allows the host protein to fold. However, if the insert is unfolded, the loss of conformational entropy needs to be compensated by the free energy gained in folding the host protein. Theoretical studies of simple polymers [50] suggest the loss in entropy due to loop closure is approximately:

$$\Delta S = -3/2\ R \ln N \qquad \text{Eqn (1)}$$

where N is the length of the loop in amino acids and R is the gas constant. Experiments on protein stability where short sequences are inserted into existing turns suggest a loss of 0.1–0.26 kcal·mol^{-1} per inserted residue, depending on the amino acid identity [51–54]. Based on Eqn (1) and these studies, we estimated that the loss of free energy due to the incorporation of an 80–100 residue insert into the loop of a folded protein would be 4–6 kcal·mol^{-1}. Therefore, a folded insert allowing the host protein to fold can potentially be distinguished from an unfolded insert by the ability of the host protein to bind its natural ligand.

Using this idea we developed a loop entropy selection technique based on a mutant of the lck SH2 domain, a 110 residue protein that binds a phosphorylated tyrosine-containing peptide [55] with a stability of 2.5 kcal·mol^{-1} [56], i.e. not enough to overcome the insertion of an unfolded protein. Phage-display experiments demonstrate that the phage containing the insertion of the folded src SH3 domain into the SH2 loop can be recovered at levels similar to the engineered SH2. Phage containing inserts of either an unfolded mutant of SH3 (L32E) or a long, mostly polar sequence, are recovered at levels equal to background [49].

Using this selection method, we sought to probe early events in protein evolution. It is plausible that protein evolution occurred in two stages: the initial generation of folded, functional polymers from random amino acid sequences, and a subsequent diversification of protein architectures through recombination between substructures present in the initial protein population. To probe the second stage, a 'mini-exon' shuffling library was made by recombining fragments of already existing proteins, and to probe the first stage a library of random sequences was used.

'Mini-exon' shuffling library

For the first library we selected the genome of *Saccharomyces cerevisiae* as the source for our fragments. Isolating large amounts of genomic DNA is relatively easy and most of the yeast genome is comprised of protein coding sequences with few introns [57]. To generate the initial peptide fragments needed for the 'mini-exon' shuffling library purified yeast nuclei were treated with a nonspecific DNase leaving only the DNA bound by nucleosomes (Fig. 1). The protected DNA fragments, estimated to be $\approx$ 130–150 base pairs on a 3% (w/v) agarose gel, were isolated and cloned into a phage-display vector to select for in-frame fragments lacking stop codons. Linkers were added to the fragments and cloned into a phagemid vector between the protein L gene and gene 8. To select for in-frame fragments, phage were panned against IgG, to which protein L specifically binds. In-frame fragments were then concatenated into dimers, cloned into the SH2 loop entropy selection phagemid vector and transformed into XL1-Blue cells to generate a library with 10^8 clones. To select for fusions capable of binding the SH2 ligand, phage were panned under low stringency binding conditions (4 °C, overnight, three washes) to maximize the recovery of positive clones. The recovered phage were then subjected to successive rounds of either low or high stringency selections (25 °C, 2 h, five to seven washes). After three rounds of selection the percentage recovery was at or above the recovery of the wild type SH2 domain. Sequences of clones from the unselected phage through two rounds of selection were examined to characterize the members of each stage of selection.

Examining the amino acid composition of the sequences recovered from the loop entropy selection and of the yeast proteome reveals significant differences between the two groups. Along with an increase in proline content (Fig. 2) there was a large enrichment in the percentage of small amino acids (A, G, S, T) and decreases in aliphatic, aromatic and charged amino acids. The loss of amino acids characteristic of hydrophobic cores, combined with an increase in residues found in loops and less well ordered structures, suggests the lack of independent folding domains in the insert sequences. Increases in proline and cysteine could counter the loop entropy selection; proline because it has a significantly lower backbone conformational entropy than the other amino acids and cysteine because disulphide bond formation could close the loop without introducing

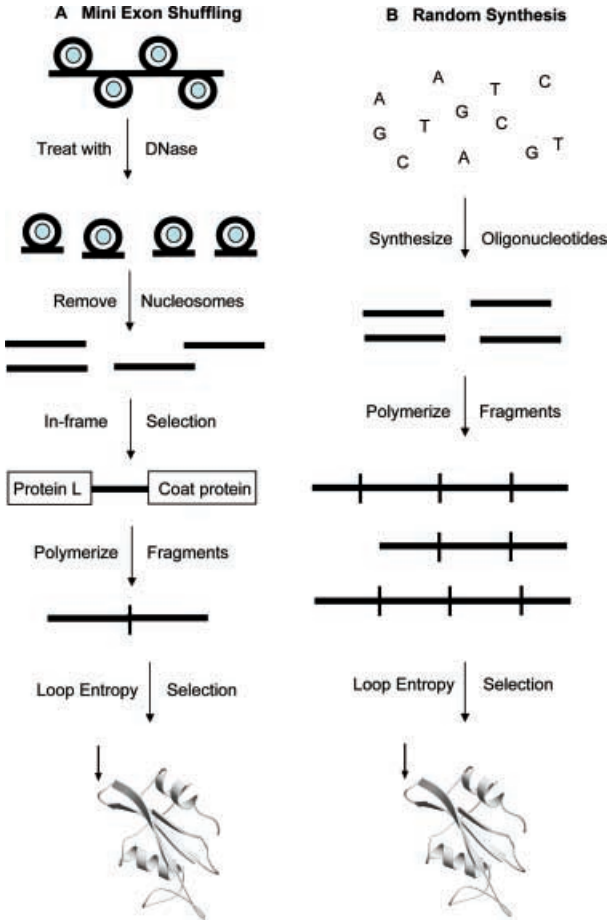

Fig. 1. Schematic diagram depicting the generation of the 'mini-exon' shuffling library and the random synthesis library for the loop entropy reduction screens. Short in-frame fragments were generated for both libraries (either by nucleosome protection or oligonucleotide synthesis). These fragments were polymerized and cloned into a loop in the lck SH2 domain (insertion point marked by arrow).

structure into the backbone of the insert. The lack of an increase in cysteines suggests that the latter possibility is not a problem. One or more of three distinct steps in the generation and recovery of the loop entropy library could be the source of the amino acid bias: (a) fragment generation, (b) in-frame selection and (c) loop entropy selection.

Although bias occurs during fragment generation and in-frame selection, comparisons of the amino acid composition between various stages of the library (Fig. 2) show the largest bias occurs during loop entropy selection. There appears to be a continuing bias towards proline and the appearance of a significant bias towards other small amino acids (A, G, S, T) and against amino acids needed for a hydrophobic core (F, I, L, M, V, W, Y). A likely explanation for the skew in amino acid composition is that aggregation-prone sequences are strongly selected against. Even though the selection appears to accept sequences not expected to be ordered, our previous SH3 controls suggest folded proteins should also be recovered. It is not clear if the initial library was devoid of folded proteins or whether these more simplified proteins out-competed the folded inserts.

A bias towards shorter inserts is also evident in both the loop entropy and in-frame selections. During the in-frame selection the average fragment size drops from 90 to 80 base pairs; in the loop entropy reduction selection the preselected inserts average 66 amino acids in length, and after one round this number drops to 51. The reduction in length of the loop entropy selection is not surprising as this should reduce the conformational entropy of an insert without forming a stable structure. The reduction in length during the in-frame selection is a problem because shorter inserts are less likely to be capable of forming a hydrophobic core in the loop entropy selection stage, thereby making folded proteins less likely in the library. In addition, selection of shorter in-frame sequences would reduce the selection against noncoding fragments relative to fragments from yeast gene coding frames due to the larger average length of in-frame fragments from a protein coding frame (33 amino acids) than noncoding frames (25 amino acids). The percentage of individual fragments originating from the coding frame of a gene does not change significantly when proceeding from in-frame fragments to the first round of the loop entropy selection (29% after the in-frame selection to 26% in the phage recovered from the loop entropy library). In the loop entropy library, however, the coding frame fragments are often from low complexity portions of proteins, such that after the first round of selection the amino acid composition of the inserts containing at least one actual protein coding fragment is not significantly different from the inserts comprised of two fragments from noncoding protein frames. This suggests the bias in the fragment generation stages (i.e. shorter fragments with certain amino acid biases) severely limits the number of inserts comprised of two fragments with protein-like sequences.

Random sequence library

The random library consists of 60 base-pair fragments ligated together to produce 180–300 base pairs of full length sequences (Fig. 1). The individual fragments were synthesized with a nucleotide bias to recreate the amino acid distribution of natural proteins, while eliminating cysteines and stop codons. Comparing the sequences recovered from the random libraries to the expected library design shows the random library has similar qualitative problems as the 'mini-exon' library, i.e. a decrease in amino acids needed for the formation of hydrophobic cores and the preferential selection of shorter sequences.

To understand the nature of the sequences selected in the loop entropy screen, the biophysical properties of sequences recovered from the random sequence library were investigated. Seven SH2-loop insert clones were chosen for characterization [56]. Four of these clones were chosen for their high representation in the selected phage libraries (clones 283, 290, 425 and 344). Another (clone 333), while not seen after the first round of panning, was chosen because of its length (100 amino acids) and level of hydrophobicity (29.5%; F, I, L, M, V, W, Y residues). The final two were selected from the phage pool as negative controls, prior to any rounds of selection (217, 227). All seven SH2-insert fusion proteins and three autonomous inserts (283, 290 and 425) were purified. Circular dichroism

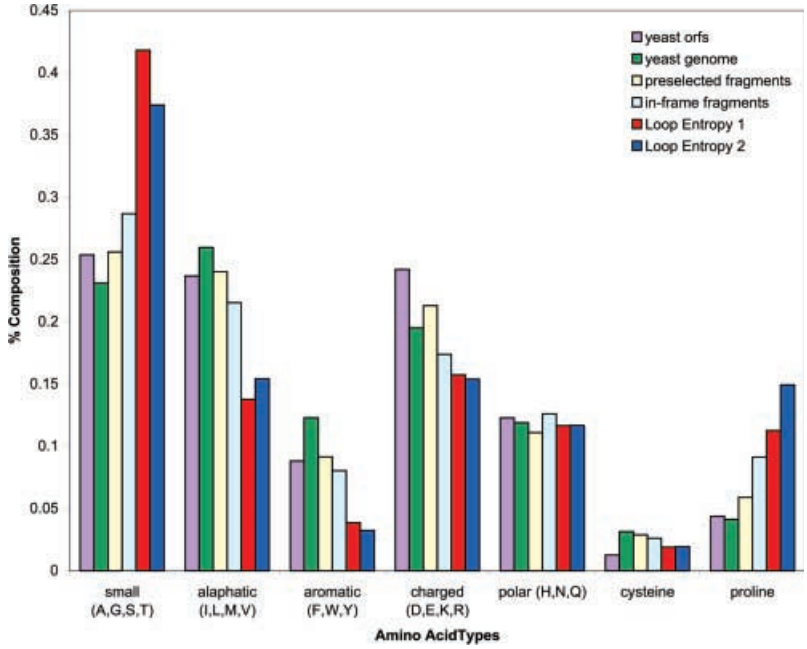

Fig. 2. Amino acid frequencies at different stages in the generation of the 'mini-exon' shuffling library. Yeast ORFs (purple) is the amino acid distribution for the suspected protein coding genes in *Saccharomyces cerevisiae* [57]. The yeast genome (green) is the hypothetical amino acid distribution one would see if the genome of *S. cerevisiae* was translated in all three frames. The preselected fragments (yellow) are fragments from nucleosome protection that have not undergone in-frame selection. The in-frame fragments (light blue) show the amino acid distribution in the fragments after the in-frame selection. Round 1 (red) and Round 2 (dark blue) of the loop entropy library are the sequences recovered from the first and second rounds of panning the 'mini-exon' shuffling library.

wavelength scans of the inserts in isolation were similar to peptides with minimal helical content, but primarily random coil structure. CD spectra of the SH2-insert fusions showed little difference from the sum of the spectra for SH2 and the isolated insert, suggesting that the inserts did not acquire structure through insertion into the SH2 loop. Equilibrium denaturation studies of the SH2-insert fusions showed that most of the inserts (six out of seven) had little or no effect on the stability of the SH2 domain ($\Delta\Delta G \approx -0.8$ to 0.2 kcal·mol^{-1}). Surface plasmon resonance studies showed five out of six of the tested SH2-insert fusions are capable of binding the SH2 peptide ligand independently of the phage context. The only apparent differences between recovered and unselected proteins were lower than expected hydrophobic amino acid contents and increased partitioning into the soluble fraction during protein expression in *E. coli* [56]. This suggests the limiting factor in the loop entropy selection is incorporation of the fusion proteins into a large number of phage. Deleterious effects on either incorporation of the fusion protein into the capsid or on phage production, possibly due to solubility of the fusion protein, may be enhanced by the presence of exposed, large hydrophobic residues.

Based on the QLR random libraries of Davidson & Sauer [21,22], compact proteins with high secondary, but low tertiary structure content relative to native proteins, should have been present in a random library of this size ($\approx 10^8$). The lack of these 'molten globule' proteins in recovered sequences suggests that the selection of 'folded' proteins in this screen is fairly strict. The inability of these proteins to form a well defined, compact, hydrophobic core may interfere with the folding or activity of the SH2 domain. The lack of 'native-like' proteins suggests the complexity ($\approx 10^8$) of the library was too small to produce folded proteins capable of being recovered by this screen. In contrast, Hecht and coworkers found folded four helix bundles by examining less than 100 binary patterned sequences [18–20]. The restriction of library contents from random sequences to specifically patterned sequences appears to enrich the content of folded proteins to greater than one in 100.

The search for genomic sequences capable of complementing the N-terminal portion of CspA [44] suggests that the 'mini-exon' library should contain folded proteins. While the apparent strictness of the selection may have limited the recovery of some folded proteins, bias against fragments from coding frames with classical protein-like sequences and towards low complexity coding frames and noncoding frames in both the initial fragment generation step and the in-frame selection may have limited the number of folded inserts in the 'mini-exon' library.

These observations do not answer the obvious question as to why the apparently unstructured inserts of 50–100 amino acids do not disrupt the folding and function of the SH2 domain. The SH2 domain's ability to retain its native structure suggests that other forces important to protein stability [58] are compensating for these entropic costs; in our experiments the free energy loss due to the entropic cost of loop insertion may have been compensated by partial collapse of the insert and/or interactions between the insert and the host SH2 domain. Alternatively, these differences between expected and observed changes in the free-energy of folding could result from an overestimation of the entropic cost of loop closure (Eqn 1).

Even though these experiments failed to provide us with the insights we sought, they do inform our perspective on protein evolution. While most multidomain proteins were probably formed by linear concatenation of individual domains, surveys of the Protein Data Bank suggest that almost 30% of domains are noncontiguous due to the insertion of one domain into another [59]. Such connections are more likely than linear connections to couple the state of one domain to the state of the other in such a way as to increase rigidity in the connections and promote allosteric interactions across the two domains. Recent experiments

have shown that such cross domain communication can arise by inserting one domain into another without selecting for positive interdomain contacts [60–62]. Our findings suggest these types of connections can arise readily during evolution because the structural constraints on the insertion of a long polypeptide into the loop of a folded domain are not as strict as previously believed. The results suggest that such long insertions would not be under strong negative selection, but instead would be nearly evolutionarily neutral, allowing the inserted sequence to slowly evolve structure and function. Interestingly the termini in naturally occurring proteins are closer to each other than would be expected by chance [63], consistent with an evolutionary model in which complex multidomain proteins can arise from the insertion of peptide modules into the loops of other folded modules.

Library screening *in silico*

Recent advances in computational protein folding and design have improved screens for folded proteins *in silico*. Computational screens have the advantage of screening larger volumes of sequence space, and directly select for stability. For a protein of 100 amino acids, dead end elimination algorithms can effectively search all 20^{100} possible sequences [64–66]. In contrast, phage-display diversity is less than 10^{10} and requires functional activity to indirectly select for stability.

To explore areas of structure space not known to be sampled in nature we computationally designed a protein sequence, named Top7, that adopts a novel topology. The topology of Top7 was chosen because it had not been observed in the Protein Data Bank. A sequence predicted to fold into this topology was identified after repeated iterative rounds of computational structure and sequence optimization. The crystal structure of Top7, determined to 2.5 Å, has a C-alpha rmsd to the designed structure of 1.2 Å [67].

We have also developed and used computational design algorithms to redesign protein folding pathways [68–70], redesign the sequences of small proteins [71,72], design novel domain swapped dimers [73], generate a novel homing endonuclease by engineering a binding interface between two domains that do not naturally interact [74] and redesign natural interfaces to generate new cognate pairs [75].

Conclusions

Due to the relative scarcity of native-like folded proteins in sequence space, methods that are capable of screening large numbers of possible sequences ($> 10^7$) are needed. We have used a phage-display system to exploit the strong correlation between structure and function in proteins to select for structurally related proteins. These experiments shed light on the sequence determinants of folding and the evolutionary pressures on protein folding and stability. Our more recent loop entropy selections for folded proteins in a random sequence library and a 'mini-exon' shuffling library illustrated the scarcity of well folded sequences in sequence space and suggested that the effects of inserting apparently disordered loops into proteins are less than previously thought, but did not allow us to further explore the limits of evolution in sequence space. In contrast, using computational design methodologies, which can search much larger volumes of sequence space, we have been able to produce a protein with a fold not previously seen in nature. A powerful combination of experimental and *in silico* selection strategies would be to use experimental molecular evolution methods such as phage-display to optimize the properties of computationally designed sequences or to search through focused libraries generated using computational design methods.

Acknowledgements

We wish to thank Michelle Scalley-Kim, Karen Butner, Philippe Minard, Charlotte Berkes and Ingo Ruczinski for their suggestions. This work was supported by a grant from the NIH (D. B.) and a Molecular Biophysics Training Grant (A. W.) from the NIH.

References

1. Scott, J.K. & Smith, G.P. (1990) Searching for peptide ligands with an epitope library. *Science* **249**, 386–390.
2. 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.
3. Wikstrom, M., Sjobring, U., Kastern, W., Bjorck, L., Drakenberg, T. & Forsen, S. (1993) Proton nuclear magnetic resonance sequential assignments and secondary structure of an immunoglobulin light chain-binding domain of protein L. *Biochemistry* **32**, 3381–3386.
4. Wikstrom, M., Drakenberg, T., Forsen, S., Sjobring, U. & Bjorck, L. (1994) Three-dimensional solution structure of an immunoglobulin light chain-binding domain of protein L. Comparison with the IgG-binding domains of protein G. *Biochemistry* **33**, 14011–14017.
5. Kastern, W., Sjobring, U. & Bjorck, L. (1992) Structure of peptostreptococcal protein L and identification of a repeated immunoglobulin light chain-binding domain. *J. Biol. Chem.* **267**, 12820–12825.
6. Gu, H., Kim, D. & Baker, D. (1997) Contrasting roles for symmetrically disposed beta-turns in the folding of a small protein. *J. Mol. Biol.* **274**, 588–596.
7. 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.
8. Alm, E., Morozov, A.V., Kortemme, T. & Baker, D. (2002) Simple physical models connect theory and experiment in protein folding kinetics. *J. Mol. Biol.* **322**, 463–476.
9. Alm, E. & Baker, D. (1999) Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. *Proc. Natl Acad. Sci. USA* **96**, 11305–11310.
10. Munoz, V. & Eaton, W.A. (1999) A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc. Natl Acad. Sci. USA* **96**, 11311–11316.
11. Galzitskaya, O.V. & Finkelstein, A.V. (1999) A theoretical search for folding/unfolding nuclei in three-dimensional protein structures. *Proc. Natl Acad. Sci. USA* **96**, 11299–11304.
12. Kuhn, H. & Waser, J. (1994) On the origin of the genetic code. *FEBS Lett.* **352**, 259–264.
13. 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.
14. 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.

15. 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.
16. DeGrado, W.F., Wasserman, Z.R. & Lear, J.D. (1989) Protein design, a minimalist approach. *Science* **243**, 622–628.
17. Regan, L. & DeGrado, W.F. (1988) Characterization of a helical protein designed from first principles. *Science* **241**, 976–978.
18. 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.
19. 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.
20. 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.
21. Davidson, A.R. & Sauer, R.T. (1994) Folded proteins occur frequently in libraries of random amino acid sequences. *Proc. Natl Acad. Sci. USA* **91**, 2146–2150.
22. Davidson, A.R., Lumb, K.J. & Sauer, R.T. (1995) Cooperatively folded proteins in random sequence libraries. *Nat. Struct. Biol.* **2**, 856–864.
23. 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.
24. Yi, Q., Rajagopal, P., Klevit, R.E. & Baker, D. (2003) Structural and kinetic characterization of the simplified SH3 domain FP1. *Protein Sci.* **12**, 776–783.
25. 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.
26. Silverman, J.A. & Harbury, P.B. (2002) The equilibrium unfolding pathway of a (beta/alpha) 8 barrel. *J. Mol. Biol.* **324**, 1031–1040.
27. Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B. & Thornton, J.M. (1997) CATH – a hierarchic classification of protein domain structures. *Structure* **5**, 1093–1108.
28. Murzin, A.G., Brenner, S.E., Hubbard, T. & Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540.
29. Holm, L. & Sander, C. (1998) Touring protein fold space with Dali/FSSP. *Nucleic Acids Res.* **26**, 316–319.
30. Blanco, F.J., Angrand, I. & Serrano, L. (1999) Exploring the conformational properties of the sequence space between two proteins with different folds: an experimental study. *J. Mol. Biol.* **285**, 741–753.
31. Prijambada, I.D., Yomo, T., Tanaka, F., Kawama, T., Yamamoto, K., Hasegawa, A., Shima, Y., Negoro, S. & Urabe, I. (1996) Solubility of artificial proteins with random sequences. *FEBS Lett.* **382**, 21–25.
32. Yamauchi, A., Yomo, T., Tanaka, F., Prijambada, I.D., Ohhashi, S., Yamamoto, K., Shima, Y., Ogasahara, K., Yutani, K., Kataoka, M. & Urabe, I. (1998) Characterization of soluble artificial proteins with random sequences. *FEBS Lett.* **421**, 147–151.
33. Doi, N., Yomo, T., Itaya, M. & Yanagawa, H. (1998) Characterization of random-sequence proteins displayed on the surface of *Escherichia coli* RNase HI. *FEBS Lett.* **427**, 51–54.
34. Doi, N., Itaya, M., Yomo, T., Tokura, S. & Yanagawa, H. (1997) Insertion of foreign random sequences of 120 amino acid residues into an active enzyme. *FEBS Lett.* **402**, 177–180.
35. Matsuura, T., Ernst, A. & Pluckthun, A. (2002) Construction and characterization of protein libraries composed of secondary structure modules. *Protein Sci.* **11**, 2631–2643.
36. Wang, W. & Hecht, M.H. (2002) Rationally designed mutations convert *de novo* amyloid-like fibrils into monomeric beta-sheet proteins. *Proc. Natl Acad. Sci. USA* **99**, 2760–2765.
37. 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.
38. Gilbert, W., de Souza, S.J. & Long, M. (1997) Origin of genes. *Proc. Natl Acad. Sci. USA* **94**, 7698–7703.
39. Doolittle, R.F. (1995) The multiplicity of domains in proteins. *Annu. Rev. Biochem.* **64**, 287–314.
40. Doolittle, R.F. (1995) The origins and evolution of eukaryotic proteins. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* **349**, 235–240.
41. Stoltzfus, A., Spencer, D.F., Zuker, M., Logsdon, J.M. Jr & Doolittle, W.F. (1994) Testing the exon theory of genes: the evidence from protein structure. *Science* **265**, 202–207.
42. Voigt, C.A., Martinez, C., Wang, Z.G., Mayo, S.L. & Arnold, F.H. (2002) Protein building blocks preserved by recombination. *Nat. Struct. Biol.* **9**, 553–558.
43. Meyer, M.M., Silberg, J.J., Voigt, C.A., Endelman, J.B., Mayo, S.L., Wang, Z.G. & Arnold, F.H. (2003) Library analysis of SCHEMA-guided protein recombination. *Protein Sci.* **12**, 1686–1693.
44. Riechmann, L. & Winter, G. (2000) Novel folded protein domains generated by combinatorial shuffling of polypeptide segments. *Proc. Natl Acad. Sci. USA* **97**, 10068–10073.
45. Fischer, N., Riechmann, L. & Winter, G. (2004) A native-like artificial protein from antisense DNA. *Protein Eng.* **17**, 13–20.
46. 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.
47. Kristensen, P. & Winter, G. (1998) Proteolytic selection for protein folding using filamentous bacteriophages. *Fold. Des.* **3**, 321–328.
48. Sieber, V., Pluckthun, A. & Schmid, F.X. (1998) Selecting proteins with improved stability by a phage-based method. *Nat. Biotechnol.* **16**, 955–960.
49. 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.
50. Chan, H. & Dill, K. (1988) Intrachain loops in polymers. *J. Chem. Phys.* **90**, 492–509.
51. Ladurner, A.G. & Fersht, A.R. (1997) Glutamine, alanine or glycine repeats inserted into the loop of a protein have minimal effects on stability and folding rates. *J. Mol. Biol.* **273**, 330–337.
52. Nagi, A.D. & Regan, L. (1997) An inverse correlation between loop length and stability in a four-helix-bundle protein. *Fold. Des.* **2**, 67–75.
53. Viguera, A.R. & Serrano, L. (1997) Loop length, intramolecular diffusion and protein folding. *Nat. Struct. Biol.* **4**, 939–946.
54. Grantcharova, V.P., Riddle, D.S. & Baker, D. (2000) Long-range order in the src SH3 folding transition state. *Proc. Natl Acad. Sci. USA* **97**, 7084–7089.
55. Eck, M.J., Shoelson, S.E. & Harrison, S.C. (1993) Recognition of a high-affinity phosphotyrosyl peptide by the Src homology-2 domain of p56lck. *Nature* **362**, 87–91.
56. Scalley-Kim, M., Minard, P. & Baker, D. (2003) Low free energy cost of very long loop insertions in proteins. *Protein Sci.* **12**, 197–206.
57. Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H. & Oliver, S.G. (1996) Life with 6000 genes. *Science* **274** (546), 563–567.
58. Dill, K.A. (1990) Dominant forces in protein folding. *Biochemistry* **29**, 7133–7155.

59. Jones, S., Stewart, M., Michie, A., Swindells, M.B., Orengo, C. & Thornton, J.M. (1998) Domain assignment for protein structures using a consensus approach: characterization and analysis. *Protein Sci.* **7**, 233–242.
60. Betton, J.M., Jacob, J.P., Hofnung, M. & Broome-Smith, J.K. (1997) Creating a bifunctional protein by insertion of beta-lactamase into the maltodextrin-binding protein. *Nat. Biotechnol.* **15**, 1276–1279.
61. Collinet, B., Herve, M., Pecorari, F., Minard, P., Eder, O. & Desmadril, M. (2000) Functionally accepted insertions of proteins within protein domains. *J. Biol. Chem.* **275**, 17428–17433.
62. Tucker, C.L. & Fields, S. (2001) A yeast sensor of ligand binding. *Nat. Biotechnol.* **19**, 1042–1046.
63. Thornton, J.M. & Sibanda, B.L. (1983) Amino and carboxy-terminal regions in globular proteins. *J. Mol. Biol.* **167**, 443–460.
64. De Maeyer, M., Desmet, J. & Lasters, I. (2000) The dead-end elimination theorem: mathematical aspects, implementation, optimizations, evaluation, and performance. *Methods Mol. Biol.* **143**, 265–304.
65. Dahiyat, B.I., Sarisky, C.A. & Mayo, S.L. (1997) *De novo* protein design: towards fully automated sequence selection. *J. Mol. Biol.* **273**, 789–796.
66. Gordon, D.B. & Mayo, S.L. (1999) Branch-and-terminate: a combinatorial optimization algorithm for protein design. *Structure Fold. Des.* **7**, 1089–1098.
67. Kuhlman, B., Dantas, G., Ireton, G.C., Varani, G., Stoddard, B.L. & Baker, D. (2003) Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**, 1364–1368.
68. Kuhlman, B., O'Neill, J.W., Kim, D.E., Zhang, K.Y. & Baker, D. (2002) Accurate computer-based design of a new backbone conformation in the second turn of protein L. *J. Mol. Biol.* **315**, 471–477.
69. Nauli, S., Kuhlman, B., Le Trong, I., Stenkamp, R.E., Teller, D. & Baker, D. (2002) Crystal structures and increased stabilization of the protein G variants with switched folding pathways NuG1 and NuG2. *Protein Sci.* **11**, 2924–2931.
70. Nauli, S., Kuhlman, B. & Baker, D. (2001) Computer-based redesign of a protein folding pathway. *Nat. Struct. Biol.* **8**, 602–605.
71. Kuhlman, B. & Baker, D. (2000) Native protein sequences are close to optimal for their structures. *Proc. Natl Acad. Sci. USA* **97**, 10383–10388.
72. Dantas, G., Kuhlman, B., Callender, D., Wong, M. & Baker, D. (2003) A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins. *J. Mol. Biol.* **332**, 449–460.
73. Kuhlman, B., O'Neill, J.W., Kim, D.E., Zhang, K.Y. & Baker, D. (2001) Conversion of monomeric protein L to an obligate dimer by computational protein design. *Proc. Natl Acad. Sci. USA* **98**, 10687–10691.
74. Chevalier, B.S., Kortemme, T., Chadsey, M.S., Baker, D., Monnat, R.J. & Stoddard, B.L. (2002) Design, activity, and structure of a highly specific artificial endonuclease. *Mol. Cell* **10**, 895–905.
75. Kortemme, T., Joachimiak, L.A., Bullock, A.N., Shuler, A.D., Stoddard, B.L. & Baker, D. (2004) Computational redesign of protein–protein interaction specificity. *Nat. Struct. Mol. Biol.* **11**, 371–379.