DOI: 10.1002/cbic.200800147

Searching Combinatorial Libraries for Native Proteins with Novel Folds

Jennifer L. Watkins and John C. Chaput*[a]

Nature uses a large but finite number of protein folds to achieve and sustain life. Whether additional folds exist beyond the set found in nature remains an interesting question with important fundamental and practical implications. By studying how proteins evolve in synthetic systems, we can begin to understand certain underlying aspects of the chemical basis of biological evolution, which in turn will allow us to constrain models that describe the evolution of biological proteins. One such question involves the long-standing debate over contingency versus determinism in natural selection.[1] Is life the result of a series of unanticipated events or the inevitable consequence of antecedents? Since it is impossible to rewind the evolutionary clock of time, methods that simulate this process in the laboratory have the ability to shed new light on questions that are otherwise difficult to address. Evolutionary strategies can also lead to new routes for creating novel synthetic proteins with tailor-made properties. The emerging field of synthetic biology promises to create living systems that synthesize chemicals, fabricate materials, produce energy, and improve the human condition and our environment. The extent to which we are successful in these areas will depend on our ability to look beyond the set of proteins found in nature and ask not what exists, but what is physically possible?

At present, all of the structures found in the protein structure data bank (PDB) can be organized into one of about 1100 different protein-fold families.[2,3] These are small proteins or domains of larger proteins that fold independently and descend from a common evolutionary ancestor. Some protein families contain many members and these are appropriately termed "superfamilies". Most families contain just a few representative members, and in some cases just one. Since these folds represent nature's set of combinatorial building blocks, understanding their distribution in biological systems allows us to hypothesize about the origins of the protein repertoire. Given that one-half of the sequences in known genomes are homologous to proteins of known structure,[4] it seems likely that the total number of protein folds will remain small. One interpretation of this observation is that biological proteins arose from a small set of primitive domains, which over time recombined to form larger structures of increasing complexity.[5,6] In this scenario, structures not easily derived from the initial set of protein folds would be absent from the set of proteins we see today in the PDB.[7]

Learning how to search large combinatorial libraries for new protein folds is a problem analogous to finding a needle in a haystack. This is because the number of amino acid sequences that are capable of folding into physically realistic structures is extremely small relative to the total number of protein sequences possible. Even modestly sized proteins have so many sequence combinations that it would be impossible to synthesize one molecule of each. For instance, a library of all possible 100-amino-acid proteins would contain 20^{100} different sequences, which far exceeds the largest library sizes that are currently possible with modern molecular biology techniques ($\sim 10^{13}$–10^{14}). To solve this problem, scientists have created combinatorial strategies that enable them to search large regions of sequence space more efficiently.[8] The goal of these studies is to design amino acid libraries with the highest probability of yielding folded proteins.

Early efforts in this area focused on the development of synthetic libraries that encoded only subsets of amino acid residues. In a classic study, Davidson and Sauer reported that proteins with native-like properties occur frequently in random libraries composed of mainly glutamine (Q), leucine (L), and arginine (R).[9] Unlike natural proteins, the QLR-derived proteins were highly insoluble and hyperstable. It was later learned that QLR libraries with lower hydrophobic content could be used to isolate synthetic proteins that fold cooperatively in the presence of chaotropic agents.[10] Hecht and co-workers developed an alternative approach to generating folded proteins.[11] Using a technique called binary patterning (Figure 1 A), they created libraries of amino acid sequences in which groups of polar and nonpolar residues are positioned in a pattern that mimics the natural periodicities of α-helical and β-sheet secondary structures. The basic premise of this strategy is that protein folds are highly degenerate, meaning that many different sequences can adopt the same shape, and therefore the ability of any given sequence to fold into a compact globular structure depends on how well individual residues will fit together to form a collapsed protein core. This strategy was used to make a library of amino acid sequences that was designed to fold into a four-helix bundle.[11] Their first attempt at these structures resulted in four-helix bundles that were molten globules, which are proteins that have discrete secondary structures, but no tertiary structure. A second-generation library

[a] J. L. Watkins, Prof. J. C. Chaput
Center for BioOptical Nanotechnology
The Biodesign Institute
Department of Chemistry and Biochemistry
Arizona State University
Tempe, AZ 85287-5201 (USA)
Fax: (+1) 480-727-0396
E-mail: john.chaput@asu.edu

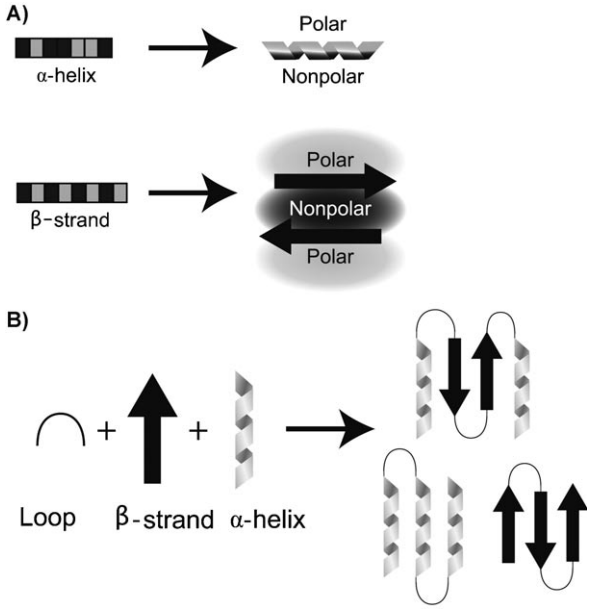

Figure 1. Strategies for building combinatorial libraries based on secondary structural elements. A) Binary patterning is a technique that positions polar and nonpolar residues in arrangements that favor α-helix or β-sheet formation. This strategy can be used to identify novel sequences that fold into predetermined structures.[11] B) Schultz and co-workers developed an alternative approach for creating designed libraries.[20] With their technique, structural elements of natural proteins recombine to make libraries of DNA that code for novel arrangements of α-helices, β-sheets, and loops. These libraries can be used to identify amino acid sequences with the potential to fold into stable structures of any topology accessible to nature's set of secondary structures.

was constructed in which the helical regions were extended to improve the stability of the fold.[12] This library gave rise to a stably folded four-helix bundle protein whose structure (Figure 2A) has now been solved by solution NMR.[13]

While focused libraries provide access to regions of sequence space with high likelihoods of finding stably folded structures, less constrained libraries allow researchers to explore new areas of the protein universe. In an attempt to understand how frequently natural selection would have produced proteins that could fold themselves into a shape with a known function, Keefe and Szostak evolved a series of ATP-binding proteins from an unbiased pool of 4×10^{12} random sequences.[14] After many iterative rounds of in vitro selection and directed evolution, several proteins emerged that bound ATP with high affinity and specificity.[15,16] The three-dimensional structure (Figure 2B) of one of these proteins has now been solved by solution NMR and X-ray crystallography, and reveals a novel zinc-nucleated α/β fold with a unique topology.[16–18] A fundamentally different approach to generating proteins that fold into structures with novel topologies was developed by Baker and co-workers.[19] Here, a general computational strategy was developed that iterates between protein sequence design and protein structure prediction to create a 93-residue α/β protein with a novel topology. To test the accuracy of their design strategy, the three-dimensional structure (Figure 2C) of the protein was determined by X-ray crystallography. The resulting structure had a root-mean-square deviation of 1.2 Å relative to the designed structure, thus indicating that in this particular case the experimentally determined structure closely matched the design prediction.

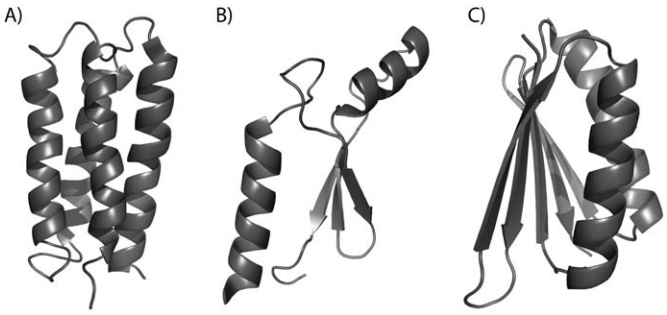

Figure 2. Examples of synthetic proteins whose structures have been solved by NMR or X-ray crystallography. A) A designed four-helix bundle protein created by binary patterning (PDB ID: 1P68).[13] B) An ATP binding protein evolved from an unconstrained library of 4×10^{12} random sequences (PDB ID: 2P09).[16–18] C) An α/β-protein obtained from a library of computational sequences (PDB ID: 1QYS).[19] The two α/β-proteins fold into novel topologies. This figure was generated by using PyMOL.[21]

In a recent paper published in the *Journal of the American Chemical Society*, Schultz and co-workers report a new strategy for generating water-soluble proteins from large pools of semirandom sequences.[20] The authors describe a protein evolution approach (Figure 1B) in which defined secondary structural elements were used to assemble a combinatorial library encoding randomly distributed regions of α-helices, β-sheets, and loops. The library was constructed from the nucleic acid sequences of 190 nonredundant *Escherichia coli* proteins of known structure. The set of parent proteins represent the four classifications of protein fold topologies, namely all-α, all-β, α/β, and α+β protein conformations. Combinatorial assembly of the different secondary structures with chain initiators and terminators resulted in a library of double-stranded DNA that coded for proteins with shuffled secondary structures. The pool was inserted into the enhanced green fluorescence protein (EGFP) fusion vector and sorted by fluorescence-activated cell sorting (FACS) to identify individual proteins that remain soluble when expressed as GFP-fusion proteins in vivo. Additional screening steps were then used to identify four clones that express in soluble form and have significant secondary structure. Although most of the clones shared no sequence homology to any known protein,

one clone did show strong homology to a domain of aspartate racemase from an unrelated bacterium. This unexpected result suggests that their starting library was large enough to explore new regions of protein shape space, but small enough to rediscover a natural protein fold.

The strategy taken by Schultz and his team of researchers is particularly exciting as it provides what appears to be a remarkably efficient method for finding proteins with the freedom to fold into any topology that is accessible to the structural elements found in nature.[20] Assuming these proteins can be evolved to adopt discrete tertiary structures, this approach would dramatically accelerate the rate at which novel protein folds are discovered. Whether their library design will be able to sample folds with more diverse arrangements of secondary structures remains to be determined. One could imagine, for example, that proteins identified from a library of known secondary structures would be biased toward solutions found in nature. If this turns out to be true, then less biased libraries will be needed to find proteins with more diverse topologies,[14] while focused libraries could be used as rapid starting points for generating new function.

Looking ahead to the future, combinatorial protein libraries will almost certainly continue to play an important role in our quest to understand the distribution of protein folds in sequence space. As progress continues, it will be interesting to see which technique or combination of techniques leads to the most efficient routes for generating new protein structures and functions. Perhaps one day enough information will be gained from these studies that we will be able to make tailor-made proteins from scratch. Until then, combinatorial libraries provide a useful tool for studying how sequence information relates to structural topology.

Acknowledgements

We wish to thank the Biodesign Institute at Arizona State University for funding.

Keywords: de novo evolution · NMR spectroscopy · novel topology · protein folding · synthetic biology · X-ray crystallography

[1] S. J. Gould, *Wonderful Life: The Burgess Shale and the Nature of History*, Norton, New York, **1989**, pp. 292–323.
[2] CATH Protein Structure Classification http://cathdb.info/latest/index.html.
[3] SCOP Protein Structure Classification http://scop.mrc-lmb.cam.ac.uk/scop.
[4] C. Chothia, J. Gough, C. Vogel, S. A. Teichmann, *Science* **2003**, *300*, 1701–1703.
[5] W. Gilbert, *Nature* **1978**, *271*, 501.
[6] R. F. Doolittle, *Annu. Rev. Biochem.* **1995**, *64*, 287–314.
[7] J. A. Gerlt, P. C. Babbitt, *Annu. Rev. Biochem.* **2001**, *70*, 209–246.
[8] M. H. Hecht, A. Das, A. Go, L. H. Bradley, Y. Wei, *Protein Sci.* **2004**, *13*, 1711–1723.
[9] A. R. Davidson, R. T. Sauer, *Proc. Natl. Acad. Sci. USA* **1994**, *91*, 2146–2150.
[10] A. R. Davidson, K. J. Lumb, R. T. Sauer, *Nat. Struct. Biol.* **1995**, *2*, 856–864.
[11] S. Kamtekar, J. M. Schiffer, H. Xiong, J. M. Babik, M. H. Hecht, *Science* **1993**, *262*, 1680–1685.
[12] Y. Wei, T. Liu, S. Sazinsky, D. A. Moffet, I. Pelczer, M. H. Hecht, *Protein Sci.* **2003**, *12*, 92–102.
[13] Y. N. Wei, S. Kim, D. Fela, J. Baum, M. H. Hecht, *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 13270–13273.
[14] A. D. Keefe, J. W. Szostak, *Nature* **2001**, *410*, 715–718.
[15] J. C. Chaput, J. W. Szostak, *Chem. Biol.* **2004**, *11*, 865–874.
[16] M. D. Smith, M. A. Rosenow, M. Wang, J. P. Allen, J. W. Szostak, J. C. Chaput, *PLoS ONE* **2007**, *2*, e467.
[17] S. S. Mansy, J. Zhang, R. Kummerle, M. Nilsson, J. J. Chou, J. W. Szostak, J. C. Chaput, *J. Mol. Biol.* **2007**, *371*, 501–513.
[18] P. Lo Surdo, M. A. Walsh, M. Sollazzo, *Nat. Struct. Mol. Biol.* **2004**, *11*, 382–383.
[19] B. Kuhlman, G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard, D. Baker, *Science* **2003**, *302*, 1364–1368.
[20] J. J. Graziano, W. Liu, R. Perera, B. H. Geierstanger, S. A. Lesley, P. G. Schultz, *J. Am. Chem. Soc.* **2008**, *130*, 176–185.
[21] W. L. DeLano, http://www.pymol.org, Delano Scientific.

Received: March 6, 2008
Published online on May 7, 2008

