

## MINIREVIEW SERIES

## Library approaches to biophysical problems

Thomas J. Magliery<sup>1</sup> and Lynne Regan<sup>1,2</sup><sup>1</sup>Department of Molecular Biophysics & Biochemistry and <sup>2</sup>Department of Chemistry, Yale University, New Haven, CT, USA

The sequence of a protein specifies its three-dimensional structure, but different sequences can specify the same fold, and the stabilities of these variants can differ vastly. What is the basis of protein stability? What are the sequence determinants of native-like thermodynamic properties? Nearly every biological process, from metabolism to signal transduction, to the structural integrity of the cell, involves proteins. Yet our understanding of the fundamental forces that stabilize structurally diverse proteins still amounts mainly to rules-of-thumb, derived from a relatively small number of mutants of a relatively small number of well-studied proteins. The answers to these questions are critical for a panoply of reasons: many common diseases are the result of protein mutations that lead to instability; many therapeutically or industrially interesting proteins are insufficiently stable for administration or process conditions; the identification of ever more sequences from genomics efforts makes the prediction of structure from sequence increasingly important; and, despite a few recent successes, protein design is still in its infancy, and even fairly successful computational approaches generally neglect protein backbone motion and employ simple potential functions of which the validity is difficult to assess. In fact, there is no generally valid way to calculate the stability of a protein, or even to quantitatively predict the effects of point mutation.

The problem of protein stability has been of interest at least since the first three-dimensional structures of proteins were determined. Why, then, after all this time, do we not have a rigorous understanding of how sequence leads to structure and stability? The answer is, at root, that protein sequence space is unimaginably vast, and that our exploration of that space is vanishingly small. Even the smallest model proteins, such as ubiquitin or rop, with about 50 residues, are single points in a  $20^{50}$ -point sequence space matrix – a collection that would have the mass of a trillion suns if each sequence were represented only once. Even if we restrict the problem considerably (for example, to all the hydrophobic core variants of a given small model protein) we are still faced with the biophysical analysis of billions of variants, and methods such as CD, NMR, calorimetry and X-ray crystallography are not suited to handle these kinds of numbers.

Of course, protein biophysicists are not the only scientists that face nature's bewildering diversity. Screens and selections applied to large collections ('libraries') of variants of organisms, proteins, nucleic acids, small molecules and even inorganic materials have revolutionized genetics, biochemistry, drug discovery and materials science. But these methods have yet to transform analogous studies of biophysical problems, fundamentally because it is difficult to screen or select for structure or stability in a high throughput fashion. In essence, two different approaches

are being applied to the problem: (a) sorting proteins based on screenable physical properties that are a result of stability; and (b) sorting proteins based on their function, where stability of the protein is required for function.

The basis of the first approach is the observation that proteins that express well, that do not aggregate, and that are resistant to proteolysis tend to be stable and native-like. Of course, it is not hard to think of examples of bona fide proteins that express poorly, aggregate upon expression, or have loops or other unstructured regions that are easily cleaved. Conversely, non-native-like proteins sometimes express well or resist proteolysis for reasons that have nothing to do with stability. This immediately raises the question: to what degree do these sorts of screens identify genuine native-like proteins? What is the actual nature of the selective pressure that is being applied?

The second approach, that of using the function of a protein as a read-out of its structural integrity, is perhaps even more laden with questions about the nature of the selective pressure that is being applied. Clearly there are convincing examples of proteins whose function depends upon their being structured and stable. However, it is probably not true that every protein in the cell is so well-behaved, and our bias in this regard may be due in part to the fact that we have chosen extremely well-behaved proteins as our models for understanding biophysical properties. More problematic is the construction of a library to screen for structural properties based on function. If the function is ligand binding, then at minimum all of the residues in direct contact with the ligand must be held constant, in case the proteins fail the screen for trivial reasons. When that binding event is intracellular and results in some downstream observable, or if the event is catalysis instead of binding, then it is even more difficult to construct rational libraries, because it is often difficult to know which residues are critical for function.

What makes these difficult problems worthy of solving is that virtually any large scale exploration of sequence space is likely to expand what is known about the relationship between sequence, structure and stability. In this series are five minireviews that describe a variety of library-based approaches to protein biophysics. Magliery & Regan present a brief overview of the methods that are currently available to screen for structured proteins, and then describe the application of some of those methods to the fundamental problem of the sequence determinants of a native-like hydrophobic core, and how that relates to overall protein stability. Special attention is given to recent work adapting the exceedingly simple and well-studied four-helix bundle protein, rop, for combinatorial experiments.

Bai & Feng describe an approach to sorting protein variants based directly on a physical property, resistance

to proteolysis. The authors consider several related approaches, their uses in improving protein stability, and a number of important technical considerations that affect the nature of the selective pressure applied.

Watters & Baker discuss an approach based indirectly on ligand-binding wherein the protein variants differ in the composition of a long loop insertion. The idea of the screen was that the conformational entropy of unfolded insertions would prevent the folding of the ligand-binding host protein, but the screen was largely confounded by both technical issues of display and by the surprisingly compact nature of even poorly structured insertions. This sort of cautionary tale calls special attention to our need to understand the true nature of the underlying selection pressure applied by screening methods.

Kotz *et al.* describe a phage-display approach wherein ligand binding is the direct readout of the stability of the protein being mutated. In addition to examination of sequence determinants of stability in the core and across the strands of  $\beta$ -sheet proteins, Cochran's lab has pioneered an important statistical approach that extends the use of phage display by quantitatively correlating phage recovery with

stability. The physical basis of this exciting method is considered in detail.

Woycechowsky & Hilvert present a review of work on the enzyme chorismate mutase, using the activity of that protein (and the necessity of that activity for cellular survival) as a way to sort libraries of protein variants to investigate determinants of tertiary packing and turns between secondary elements, among other things. This work is compared to other function-based selections that probe the stability of diverse enzymes.

This collection reflects the state-of-the-art in the application of combinatorial methods to protein biophysics, but it is also intended to be a critical look at the current limitations and opportunities. These methods highlight the need for increasing the throughput of traditional biophysical methods to examine the meaning of the results from combinatorial approaches, much as proteomics is revolutionizing the scale of biochemical investigation. Indeed, large scale approaches to biophysics are poised to revolutionize the field and finally provide satisfying answers to a problem we have been addressing, one protein at a time, for nearly half a century.

*Keywords:* combinatorial; protein stability; protein folding.



Thomas J. Magliery was born in Oak Park, Illinois, and received his AB in chemistry from Kenyon College, Gambier, Ohio. He received his PhD from the University of California, Berkeley, under the direction of Peter G. Schultz, developing combinatorial methodology that led to the first bacteria able to site-specifically insert unnatural amino acids into proteins. He has worked with Lynne Regan at Yale University for the last two years on combinatorial methods to address biophysical problems, including the role of protein cores in stability, methods to identify interacting proteins and applications of statistical methods to protein design.

Lynne Regan is Professor of Molecular Biophysics & Biochemistry and Professor of Chemistry at Yale University. She received her BA from Oxford University in biochemistry and her PhD in biology from the Massachusetts Institute of Technology under the direction of Paul Schimmel. She did postdoctoral work as a visiting scientist with William F. DeGrado before starting her group at Yale in 1990. Her research interests encompass many aspects of protein structure, folding and design with particular emphasis on protein-protein and protein-RNA interactions.