JPP JOURNAL OF PROTEINS AND PROTEOMICS
3(2), July-December 2012, pp. 79-83

COMMENTARY

# PROTEIN DESIGN - A VAST UNEXPLOITED RESOURCE

**Liam M. Longo and Michael Blaber***

*Department of Biomedical Sciences, College of Medicine, Florida State University, Tallahassee, FL 32306-4300, USA*

***Abstract:*** Proteins are vastly more complex compared to typical organic molecules produced by synthetic organic chemistry, and it stands to reason that the functional capacity of proteins might correspondingly exceed that of smaller organic molecules. However, while synthetic organic chemistry has had a major economic impact, the economic impact of designed proteins has, to date, been comparatively miniscule. While synthetic organic chemistry is a relatively mature science, protein design remains in its infancy. Efficient synthesis of polypeptides with high yield and purity, and low cost, is not the issue; rather, design challenges include creating an efficient folding pathway, rapid folding kinetics, correct target conformation, appropriate thermodynamics, useful folded molecular dynamics, solubility, functional specificity, and other issues. While computational approaches ultimately will yield solutions to these problems, current protein design still benefits substantially from experimental input in the design cycle to yield success. Recent efforts in "top-down" design of symmetric protein folds have successfully yielded short peptide building blocks (30-50 amino acids) with remarkable folding properties that highlight the usefulness of symmetry in protein design; such building blocks have potential broad utility in *de novo* protein design strategies.

***Keywords:*** Protein design; top-down symmetric deconstruction; protein evolution; peptide building block.

## Lessons from Synthetic Organic Chemistry

Advancements in synthetic organic chemistry (SOC) have revolutionized a multitude of fields, ranging from aerospace, agriculture, automotive, construction, electronics, energy, food, forestry and paper, health care and pharmaceuticals, packaging, personal goods, petroleum products, printing, textiles and dyes, and water purification, among others; essentially, nearly every facet of modern-day life. At present, the corresponding economic contribution to GDP of SOC is substantial; in the UK alone the contribution is estimated at >$250 billion annually, and is associated with over 6,000,000 jobs (Delpy and Pike, 2010). Conversely, attempts to leverage the great promise of protein design (PD)—that is, engineering proteins with desirable structural and biophysical properties for human benefit—have yet to realize even a fraction of the commercial impact of SOC, despite intense scientific interest. Although SOC and PD are fundamentally different (SOC is concerned predominantly with designing the covalent connectivity between atoms; PD considers the design of specific conformational states of proteins) both disciplines seek to generate novel chemical species with precise structural and functional characteristics, and both suffer from the "combinatorial explosion" problem in design complexity. Thus, while these fields differ in details, considering the general features that have contributed to the scientific and economic successes of SOC may help focus efforts within the PD community. We suggest that the achievement of SOC can be roughly attributed to three main factors, described below.

Firstly, organic molecules are capable of great complexity, with an associated vast diversity of biological, chemical, and physical functionality.

Corresponding Author: **Michael Blaber**
*E-mail: michael.blaber@med.fsu.edu*

Received: November 23, 2012
Accepted: November 26, 2012
Published: November 30, 2012

Consistent with this view, successful targets of organic synthesis have become workhorse molecules in a range of business sectors and economic areas. Because organic molecules and polymers possess nearly limitless functional potential (and, by extension, economic potential) corporate and scientific interest continue to drive developments in SOC.

Secondly, the modern-day synthetic organic chemist has a detailed understanding of chemical reactions informed by over a century of directed research (Nicolaou and Snyder, 2003; Nicolaou and Chen, 2011; Nicolaou and Sorensen, 1996). The current theoretical description of SOC is sophisticated and draws from quantum mechanics computations, models of chemical reactivity (e.g., transition state theory), and a rich literature on reaction mechanisms. Furthermore, many of the practical challenges faced by SOC have been overcome, and the field now benefits from a host of techniques that are adept at structural characterization of organic molecules (FTIR, NMR, and crystallography are used routinely) and mature chemical separations protocols. Indeed, even chemically labile groups can be engineered into large organic compounds using protection-group approaches. Importantly, such knowledge enables the utilization of cheap and simple (i.e., commodity) precursor molecules in complex syntheses.

Finally, the logic of chemical synthesis, exploiting cheap and simple precursor molecules, has been refined and streamlined, principally by Corey's introduction of retrosynthetic analysis (Corey and Cheng, 1989). The idea of retrosynthetic analysis—that synthetic strategies should be conceptualized staring from the target molecule and proceed via a series of back reactions ("transforms") without regard for the molecular precursors—is a distinct type of contribution compared to the work of theoreticians and experimentalists: Retrosynthetic analysis provides a logical framework for establishing a synthetic strategy, and in doing so it facilitates comparisons between competing synthetic strategies (i.e., retrosynthetic trees) thereby making optimization of cost, environmental impact, and yield far more tractable.

**Protein Engineering — the Potential**

Proteins are several orders of magnitude more complex than typical organic molecules; thus, like small molecules and organic polymers, proteins possess vast structural (and therefore functional) potential, as demonstrated by the numerous biological roles that proteins satisfy, including structural support (actin, tubulin), molecular motors (kinesin), ligands, enzymes (oxidoreductases, transferases, hydrolases, lyases, isomerases, and ligases), and materials (lens proteins, silk proteins, etc.). In addition to being functionally diverse, protein molecules can respond to external conditions (e.g., changes in pH (Wang and Xu, 2011), electric potential (DeCoursey, 2008), temperature) and several peptide systems have become targets for engineering "self-healing" materials (Gelain et al., 2011). Also, proteins and protein assemblies are capable of allosteric response and molecular communication, providing for vastly more complex functional roles.

Given the above, the economic potential inherent in protein design is vast, at least on par with SOC. Vast, but virtually unexploited: Biopharmaceuticals (i.e., protein-based therapeutics) make up only a tiny fraction of FDA-approved treatments and the use of enzymes in industrial synthetic processes lags far behind SOC enzymatic approaches. At present, only a few commercial products leverage the power of PD, notable exceptions being the presence of enzymes in detergents (Kumar *et al.*, 1998), ice structuring proteins in ice cream (Regand and Goff, 2006), as well as the expanding field of engineered human antibodies. Considering that many protein-based technologies represent patentable intellectual property—a major incentive for companies to employ PD strategies—what is holding PD back?

PD is orders of magnitude more complex than SOC and the theoretical sophistication of the PD field is not yet up to the challenge of computation-based design strategies (Snow *et al.*, 2005). Protein molecules encompass orders of magnitude more atoms than the targets of SOC, especially if solvent atoms are considered, making simulations computationally challenging and expensive. In addition, the goal of PD is to prepare a molecule with a given *conformation, dynamics, efficient folding*

*pathway, and favorable thermodynamics*, not just given atomic arrangement, thus, prediction of accurate folding kinetics as well as thermodynamics are essential for efficient PD. Forces that govern protein conformation and dynamics (e.g., hydrogen bonding, van der Waal's, electrostatic interactions) tend to be weaker, subtler than the covalent bonding that dominates SOC (Dill, 1990) and enumerating the entropic considerations that accompany protein folding is computationally formidable, making evaluations of protein structure, stability, and folding exceptionally difficult. Assuming the above problems can be solved, protein engineers still face a "combinatorial explosion" of potential amino acid conformations and sequences during the design process (Levitt *et al.*, 1997). In addition, because protein function is tied closely to conformation and dynamics, and not merely chemical structure (i.e., amino acid sequence), and a wide range of different sequences encode the same global topology, albeit with potentially critical differences in biophysical properties, the best solutions for PD project are difficult to accurately predict. Thus, while the future of PD surely belongs to the computational chemist, current limitations in computing power and programming approaches are, as of yet, insufficient to be reliably applied to solve many PD problems (Snow *et al.*, 2005).

In contrast to the still-developing theoretical aspects of PD, protein production and characterization technologies are highly evolved. The chemical synthesis of small to medium polypeptides is becoming more and more inexpensive while simultaneously increasing efficiency. Heterologous protein expression now permits cost-effective large-scale preparation of highly homogenous protein samples, and expression systems have been developed that tackle issues of post-translational modification (e.g., glycosylation, correct disulfide pairing, etc.) (Nielsen, 2012). Once purified, high-resolution crystallography and solution-state nuclear magnetic resonance can elucidate atomic details of protein conformation, while calorimetry and chaotropic-based unfolding protocols can quantify thermodynamics ($\Delta G_{unfolding}$) and folding kinetics. Methods such as phi-value analysis can identify key residues contributing to an efficient folding pathway. Thus, given the relative maturity of experimental protein characterization, the most practical model of PD is still driven principally by experimental feedback in the design process, typically most successful with design strategies involving a high degree of granularity (i.e., the construction of intermediate mutants to permit step-by-step confirmation of design principles). In this context, predictive computational studies have greater success evaluating limited rather than extensive mutational changes. Recent experimental methods, such as "top-down symmetric deconstruction" (TDSD) have achieved remarkable success in PD and understanding the evolution of protein structure (Lee *et al.*, 2011).

**Top Down Symmetric Deconstruction — a Way Forward (by Looking Backward)**

TDSD is an experimental methodology for PD useful for identifying simple polypeptide "building blocks" for symmetric protein folds (Lee and Blaber, 2011; Lee *et al.*, 2011). The method begins with a foldable polypeptide sharing the target architecture (the "proxy"), and proceeds by introducing an increasing symmetric constraint upon the primary structure. The successful end result is a purely symmetric, foldable, and thermostable polypeptide, comprised of a repeating building block sequence, that can subsequently be fragmented to study oligomerization properties (thereby supporting specific evolutionary pathways involving gene duplication and fusion events) or used as a scaffold for *de novo* protein design of novel functionality utilizing the target architecture. Since the process begins with a foldable protein, as long as intermediate mutations do not deviate from foldable sequence space, a solution is ensured. Key aspects of the TDSD design principle include the degree of granularity (i.e., the number of required intermediate mutations) as well as the logical "transforms" pursued in designing the symmetric constraint. With high granularity, movement of an intermediate mutation out of foldable sequence space unambiguously pinpoints design flaws and requires limited backtracking to correct the problem. Successful transforms to introduce a symmetric constraint have included initial focus

upon symmetric core design, followed by specific secondary structure (e.g., turn, then β-strand) symmetric design. Tertiary structure mutations to eliminate regions of asymmetry (i.e., "bulges" or insertions) can simultaneously enable symmetric primary structure solutions. Although derived from a specific proxy, the result of TDSD is a polypeptide building block potentially devoid of specific function (i.e., functionally neutral) and with hyperthermophile stability; thus, having tremendous utility in subsequent *de novo* design of a novel protein sharing the target architecture.

The recent successes of TDSD (Alsenaidy *et al.*, 2012; Lee and Blaber, 2011; Lee *et al.*, 2011; Richter et al., 2010; Yadid and Tawfik, 2011) have provided clear answers to several open questions in the protein folding and evolution field. TDSD was used to prepare a purely symmetry protein scaffold that was more stable and folded faster than the proxy, demonstrating explicitly that sequence symmetry does not *de facto* impede foldability, a controversial result (Wolynes, 1996). In addition, folding pathway studies hint that sequence symmetry may be compatible with folding pathway redundancy (thus, increasing the utility of such designed building blocks) (Longo *et al.*, 2012). Also, TDSD studies have successfully identified simple peptide motifs that can spontaneously assemble into complex architecture—in the process elucidating specific evolutionary pathways involving gene duplication and fusion in the emergence of extant protein folds—highlighting nature's own design principle in creating novel architectures and functions, and potentially elucidating molecular events in evolution that predate the last universal common ancestor. At present, the strategy of TDSD is being applied to novel problems, such as pre-biotic protein design, in an effort to understand the properties of the first proteins in abiogenesis (Longo and Blaber, 2012).

## *Abbreviations*

SOC, synthetic organic chemistry; PD, protein design; FTIR, fourier transform infrared spectroscoty; NMR, nuclear magnetic resonance; TDSD, top-down symmetric deconstruction; GDP, gross domestic product.

## *References*

Alsenaidy, M.A., Wang, T., Kim, J.H., Joshi, S.B., Lee, J., Blaber, M., Volkin, D.B., and Middaugh, C.R. (2012). An empirical phase diagram approach to investigate conformational stability of "second-generation" functional mutants of acidic fibroblast growth factor (FGF-1). Protein Sci *21*, 418-432.

Corey, E.J., and Cheng, X.-M. (1989). The logic of chemical synthesis. New York: John Wiley & Sons, Inc.

DeCoursey, T.E. (2008). Voltage-gated proton channels. Cell Mol Life Sci *65*, 2554-2573.

Delpy, D., and Pike, R. (2010). The Economic Benefits of Chemistry Research to the UK. Oxford Economics monograph, 1-158.

Dill, K.A. (1990). Dominant forces in protein folding. Biochemistry *29*, 7133-7155.

Gelain, F., Silva, D., Caprini, A., Taraballi, F., Natalello, A., Villa, O., Nam, K.T., Zuckermann, R.N., Doglia, S.M., and Vescovi, A. (2011). BMHP1-derived self-assembling peptides: hierarchically assembled structures with self-healing propensity and potential for tissue engineering applications. ACS Nano *5*, 1845-1859.

Kumar, C.G., Malik, R.K., and Tiwari, M.P. (1998). Novel enzyme-based detergents: An Indian perspective. Current Sci *75*, 1312-1318.

Lee, J., and Blaber, M. (2011). Experimental support for the evolution of symmetric protein architecture from a simple peptide motif. Proc Natl Acad Sci U S A *108*, 126-130.

Lee, J., Blaber, S.I., Dubey, V.K., and Blaber, M. (2011). A polypeptide "building block" for the ß-trefoil fold identified by "top-down symmetric deconstruction". J Mol Biol *407*, 744-763.

Levitt, M., Gerstein, M., Huang, E., Subbiah, S., and Tsai, J. (1997). Protein folding: the endgame. Annu Rev Biochem *66*, 549-579.

Longo, L., Lee, J., and Blaber, M. (2012). Experimental Support for the Foldability-Function Tradeoff Hypothesis: Segregation of the Folding Nucleus and Functional Regions in FGF-1. Protein Sci *(in press)*.

Longo, L.M., and Blaber, M. (2012). Protein design at the interface of the pre-biotic and biotic worlds. Arch Biochem Biophys (*in press*).

Nicolaou, K.C., and Chen, J.S. (2011). Classics in total synthesis III : further targets, strategies, methods. Weinheim: Wiley-VCH.

Nicolaou, K.C. and Snyder, S.A. (2003). Classics in total synthesis II : more targets, strategies, methods. Weinheim: Wiley-VCH.

Nicolaou, K.C., and Sorensen, E.J. (1996). Classics in total synthesis : targets, strategies, methods. Weinheim ; New York: VCH.

Nielsen, J. (2012). Production of biopharmaceutical proteins by yeast: Advances through metabolic engineering. Bioengineered (*in press*).

Regand, A., and Goff, H.D. (2006). Ice recrystallization inhibition in ice cream as affected by ice structuring proteins from winter wheat grass. J Dairy Sci *89*, 49-57.

Richter, M., Bosnali, M., Carstensen, L., Seitz, T., Durchschlag, H., Blanquart, S., Merkl, R., and Sterner, R. (2010). Computational and experimental evidence for the evolution of a $(ba)_8$-barrel protein from an ancestral quarter-barrel stabilized by disulfide bonds. J Mol Biol *398*, 763-773.

Snow, C.D., Sorin, E.J., Rhee, Y.M., and Pande, V.S. (2005). How well can simulation predict protein folding kinetics and thermodynamics? Annu Rev Biophys Biomol Struct *34*, 43-69.

Wang, Y.Z., and Xu, T.L. (2011). Acidosis, acid-sensing ion channels, and neuronal cell death. Mol Neurobiol *44*, 350-358.

Wolynes, P.G. (1996). Symmetry and the energy landscapes of biomolecules. Proc Natl Acad Sci USA *93*, 14249-14255.

Yadid, I., and Tawfik, D.S. (2011). Functional b-propeller lectins by tandem duplications of repetitive units. Prot Eng Des Sel *24*, 185-195.