# Evolution and evolvability of proteins in the laboratory

**Michael W. Deem***

*Departments of Bioengineering and Physics & Astronomy, Rice University, 6100 Main Street, MS 142, Houston, TX 77005-1892*

The introduction of DNA shuffling in 1994 substantially increased the power of laboratory evolution protocols to optimize protein function and led to a significant increase in the number of scientists pursuing directed protein evolution (1). The introduction of family shuffling in 1998 allowed for the homologous recombination of natural diversity among families of proteins and further increased the range of protein sequence and function space that could be accessed in the laboratory (2). The realization that evolution of truly novel protein function likely requires nonhomologous exchange of genetic material led to the introduction of the THIO-ITCHY technique in 2001 (3). In this issue of PNAS, Saraf *et al.* (4) develop the theory behind nonhomologous evolution of protein function by using the information contained within the natural diversity found in a protein fold family (Fig. 1).

The theory developed by Saraf *et al.* (4) quantifies by how much a test protein sequence differs in its structural and chemical properties from the consensus sequence of a protein fold. The test sequences are those that arise by the mutation and recombination dynamics of the THIO-ITCHY technique (3). The first step of the procedure is to determine which pairs of residues in the protein are evolutionary conserved, or correlated (5), in the family of proteins with the chosen fold. It is implicitly assumed that many of the dominant interactions are pairwise additive. A measure of the strictness of this conservation is calculated. The second step of the procedure is to determine whether the test protein sequence also conserves the structural and chemical properties at the pairs of residues identified to be conserved. The amino acid properties for which conservation is examined are charge, volume, and hydrophobicity. Through quantifying by how much the test sequence fluctuates away from the average properties of the protein fold, relative to the natural fluctuations within the fold ensemble, the authors are able to rank the probable activity of the test sequences. Although the ranking was of enzymatic activity in this case, the theory is generalizable to any measurable figure of merit that is correlated with protein structure.

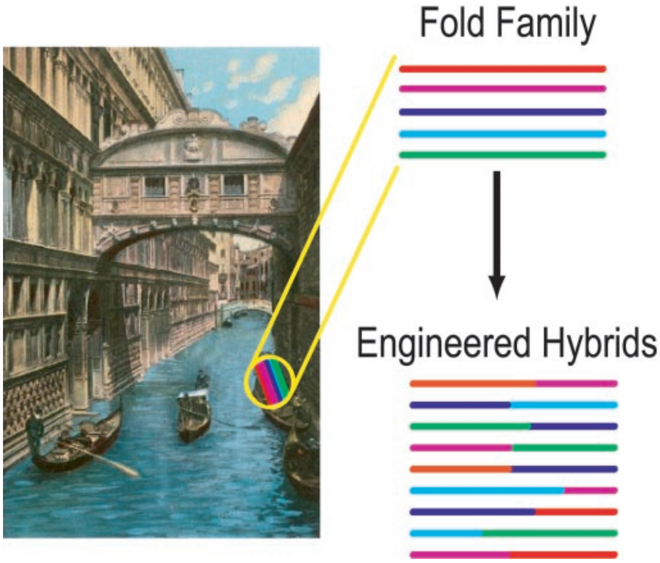

**Fig. 1.** In the canal of functional proteins (25, 26) lie members of a protein fold. The theory of Saraf *et al.* (4) predicts the probable activities of the nonhomologously engineered protein hybrids of members of this fold.

Evolution of better protein-based therapeutics is one example of how this technology can be used. An interesting application comes from the Stemmer group (6) where a tetravalent vaccine for dengue fever has been developed. Dengue fever comes in four strains, and vaccines vainly struggle for dominance over all strains. Most vaccines provide protection against only one or two strains, and there is currently no Food and Drug Administration-approved vaccine against all four strains. The Maxygen group (6) used a combination of nonhomologous exon shuffling and bioinformatics theory that identified and enhanced crossover sites for optimal shuffling to evolve a single protein antigen that provokes an antibody response against all four dengue strains. This vaccine has enormous potential importance to the 2.5 billion people who live in dengue-infected regions, of whom 100 million are stricken with dengue each year.

Predictive models for analyzing the outcome of experimental rounds of selection and mutation also may be useful in the analysis of disease evolution. At the level of point mutation, for example, diversity within disease protein folds is now a factor in protein inhibitor design (7, 8). Larger-scale genetic changes such as recombination (9, 10) and transposition (11), however, also play an important role in creating pathogen diversity. Quantitative theories can answer the following questions. How useful is recombination to evolution? Can we predict such usefulness? Can we predict likely recombinations (12)? Finally, can we suggest optimal treatment strategies in light of the recombination predictions? Perhaps the most immediate application of such a program would be to HIV dynamics and treatment (13).

An interesting question that the theory of Saraf *et al.* (4) might be able to address is why recombination and C+G content are correlated. This correlation, which has been observed for over a decade (14), still is unexplained. Various mechanisms that might lead to this correlation have been posited. On the one hand are the theories that suggest the C+G bias is a result of selective pressures that only become apparent with the increased sequence-searching ability that recombination provides. On the other hand are theories that suggest C+G content might bias recombination rates, through chromatic associations that cause physical exposure of the DNA, CG-biased mismatch repair, or an underlying bias of the associated biochemical machinery. Most of these theories rely on statistical analysis of DNA sequence data for their support. The

See companion article on page 4142.

*E-mail: mwdeem@rice.edu.

www.pnas.org/cgi/doi/10.1073/pnas.0400475101

sequence-level theory of Saraf *et al.* might help to settle this question directly, at least in the *in vitro* setting.

The theoretical models of Saraf *et al.* (4) also may be used to examine such basic questions as how diversity and evolvability arise in natural systems. Within the evolutionary biology community, there is rigid resistance to the concept that evolvability is a selectable trait: because evolvability is a characteristic of the future, causality would seem to prevent its selection. Selection, however, operates at the group level. Indeed, the mechanism for the evolution and maintenance of adaptability traits is population-based and requires a dynamic environment. The general framework relating evolvability and environmental dynamics recently has been presented (15). Simulations (16, 17) and theories (18) suggest in a general way how evolvability arises. Detailed studies at the sequence level, which the work of Saraf *et al.* enables, could further clarify the mechanisms by which evolvability arises in a population.

Recent bioinformatics studies show that the frequency of alternative splicing is much lower within annotated domains than random chance would dictate (19). That is, the process of alternative splicing tends to insert or delete entire protein domains more frequently, and disrupt protein domains less frequently, than expected by chance. One might ask whether this positive selection for evolvability also has left a mark on the within-domain splice sites. That is, do the splice sites that occur within domains tend to occur in positions that tend not to be "clashing," in the language of Saraf *et al.* (4)? On a related note, do species with high crossover rates tend to have domain families that lead to fewer clashes upon recombination than would be expected by chance?

**Sequence-level theory may elucidate how diversity and evolvability arise in protein families.**

One interesting feature of crossover-type protein evolution experiments, which the theory of Saraf *et al.* (4) reproduces, is a characteristic V shape in the activity as a function of crossover position. This characteristic shape is a reflection of the typical reduction in activity as the evolved sequences become more distinct from the parent sequences: a crossover in the middle creates maximal distinction on average. This result points toward the need to develop methods to search the protein sequence space in a nonrandom way. By biasing the search of protein sequence space with what we know about protein structure, it is possible to make large jumps in sequence space between functional regions (20). Pathway evolution (21) and module recombination (22) are macromolecular examples of this finding. The combination of predictive ability to create new structural folds (23) and predictive ability to rationally (24) or combinatorially (4) optimize fold function should yield intriguing and possibly profound results in the coming years.

1. Stemmer, W. P. C. (1994) *Nature* **370,** 389–391.
2. Crameri, A., Raillard, S. A., Bermudez, E. & Stemmer, W. P. C. (1998) *Nature* **391,** 288–291.
3. Lutz, S., Ostermeier, M. & Benkovic, S. J. (2001) *Nucleic Acids Res.* **29,** E16.
4. Saraf, M. C., Horswill, A. R., Benkovic, S. J. & Maranas, C. D. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 4142–4147.
5. Rod, T. H., Radkiewicz, J. L. & Brooks, C. L. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 6980–6985.
6. Stemmer, W. & Holland, B. (2003) *Am. Sci.* **91,** 526–533.
7. Freire, E. (2002) *Nat. Biotech.* **20,** 15–16.
8. Leslie, M. (2002) *Science* **297,** 1615.
9. Colegrave, N. (2002) *Nature* **420,** 664–666.
10. Worobey, M., Rambaut, A. & Holmes, E. C. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 7352–7357.
11. Shapiro, J. A. (1997) *Trends Genet.* **13,** 98–104.
12. Moore, G. L., Maranas, C. D., Lutz, S. & Benkovic, S. J. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 3226–3231.
13. Lathrop, R. H. & Pazzani, M. J. (1999) *J. Comb. Optim.* **3,** 301–320.
14. Eyre-Walker, A. (1993) *Proc. R. Soc. London B* **252,** 237–243.
15. Sato, K., Ito, Y., Yomo, T. & Kaneko, K. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 14086–14090.
16. Travis, J. M. J. & Travis, E. R. (2002) *Proc. R. Soc. London B* **269,** 591–597.
17. Kepler, T. B. & Perelson, A. S. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 11514–11519.
18. Blasio, F. V. D. (1999) *Phys. Rev. E* **60,** 5912–5917.
19. Kriventseva, E. V., Koch, I., Apweiler, R., Vingron, M., Bork, P., Gelfand, M. S. & Sunyaev, S. (2003) *Trends Genet.* **19,** 124–128.
20. Bogarad, L. D. & Deem, M. W. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 2591–2595.
21. Stemmer, W. P. C. (2002) *J. Mol. Catal. B* **19,** 3–12.
22. Dueber, J. E., Yeh, B. J., Chak, K. & Lim, W. A. (2003) *Science* **301,** 1904–1908.
23. Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L. & Baker, D. (2003) *Science* **302,** 1364–1368.
24. Looger, L. L., Dwyer, M. A., Smith, J. J. & Hellinga, H. W. (2003) *Nature* **423**, 185–190.
25. Waddington, C. H. (1942) *Nature* **150,** 563–565.
26. Kauffman, S. A. (1993) *The Origins of Order* (Oxford Univ. Press, New York).