

Protein engineers turned evolutionists

Sergio G Peisajovich & Dan S Tawfik

When generating novel tailor-made proteins, protein engineers routinely apply the principles of 'Darwinian' evolution. However, laboratory evolution of proteins also has the potential to test evolutionary theories and reproduce evolutionary scenarios, thus reconstructing putative protein intermediates and providing a glimpse of 'protein fossils'. This commentary describes research at the interface of applied and fundamental molecular evolution, and provides a personal view of how synergy between fundamental and applied experiments indicates novel and more efficient ways of generating new proteins in the laboratory.

Site-directed mutagenesis and attendant molecular biological tools gave scientists the power to alter the sequence, structure and function of proteins¹. This revolution led to the expectation that amino acid exchanges could be rationalized to alter protein structure and function 'to order', giving birth to the term 'protein engineering'. Time and experience, however, have tempered this assumption. It is clear now that creation of novel, functional proteins by rational design is an intricately difficult task¹. Many protein engineers have therefore turned to harnessing nature's capacity to 'engineer' proteins, primarily by directed, or *in vitro*, evolution.

Protein engineering by trial and error

Directed evolution experiments have the two characteristics of Darwinian evolution. First, a genetically diverse 'population' is created (a gene library). Second, a selection is applied to isolate the subpopulation of most active gene variants (the 'fittest'). The latter, the selection or screen, represents the most significant hurdle. The more far-fetched the target function is, the more scarce is the combination of mutations that encodes it. What is more, most mutations, including mutations that may endow a new function², compromise protein stability³. The frequency of 'viable' genes in libraries, let alone viable genes with

a new function, is therefore extremely low. In addition, whereas creation of gene libraries is based on generic protocols, selection must be individually tailored for each target function, and many targets impose severe technical limitations. Much of the research efforts by protein engineers and *in vitro* evolutionists have therefore focused on expanding the scope and throughput of screening and selection techniques.

The arsenal of methodologies for generating libraries and then selecting and screening them that became known as directed, or *in vitro*, evolution have been used to generate a plethora of new protein molecules. These successes have led to directed evolution becoming the 'go-to' technique in protein engineering; its dominance is being 'threatened' only by the growing success of computational design (although eventually these approaches are likely to be combined). It seems then that protein engineering (by directed evolution at least) is well named in that it comprises a set of techniques that can be applied to the construction of novel proteins. The irony is, of course, that evolution is all but an engineering approach. (With few notable exceptions, most functional buildings or bridges were not selected from random assemblies of beams, bolts and nuts.)

Not just an engineering endeavor

The prospect of engineering proteins has been the major driving force behind the advent of directed, *in vitro* evolution. But it has not been the only driving force. In the first studies of directed evolution, almost 40 years ago, molecular evolutionists were

driven by a desire to understand how mutations in existing genes could yield new metabolic capabilities, namely, how new enzymes diverge from existing ones. Such pioneers as Clarke and Hall used a classical procedure by which mutations permit growth of an otherwise deficient organism on a particular source of carbon or nitrogen, and could thus (given the techniques available at the time, one would say admirably) follow the evolution of new enzyme functions^{4,5}. This theme was similar to that applied today by protein engineers who are primarily interested in the power of directed evolution to generate new, tailor-made proteins. However, protein engineers did not completely ignore the possibility of addressing fundamental aspects of molecular evolution.

One of the first examples illustrating how directed evolution aimed primarily at protein engineering could also shed light on natural protein evolution came from the practical need to increase the thermal stability of enzymes. Enzymatic functions depend on thermal motions that, at the temperature at which the organism lives, allow the right balance between the 'rigidity' required for stability and the 'flexibility' required for activity. This may explain why enzymes from thermophiles (organisms that can thrive at well above 37 °C) become inactive at low temperatures, and enzymes from psychrophiles and mesophiles (organisms that live at low or medium temperatures, respectively) are unstable at high temperatures. Is this tradeoff unavoidable? Apparently not. Arnold and coworkers, for example, have evolved enzymes from both psychrophilic and mesophilic bacteria

Sergio G. Peisajovich and Dan S. Tawfik are at the Department of Biological Chemistry, Weizmann Institute of Science, Rehovot, 76100, Israel. Sergio G. Peisajovich is presently at the Department of Cellular and Molecular Pharmacology, University of California San Francisco, San Francisco, California 94158, USA.
e-mail: dan.tawfik@weizmann.ac.il

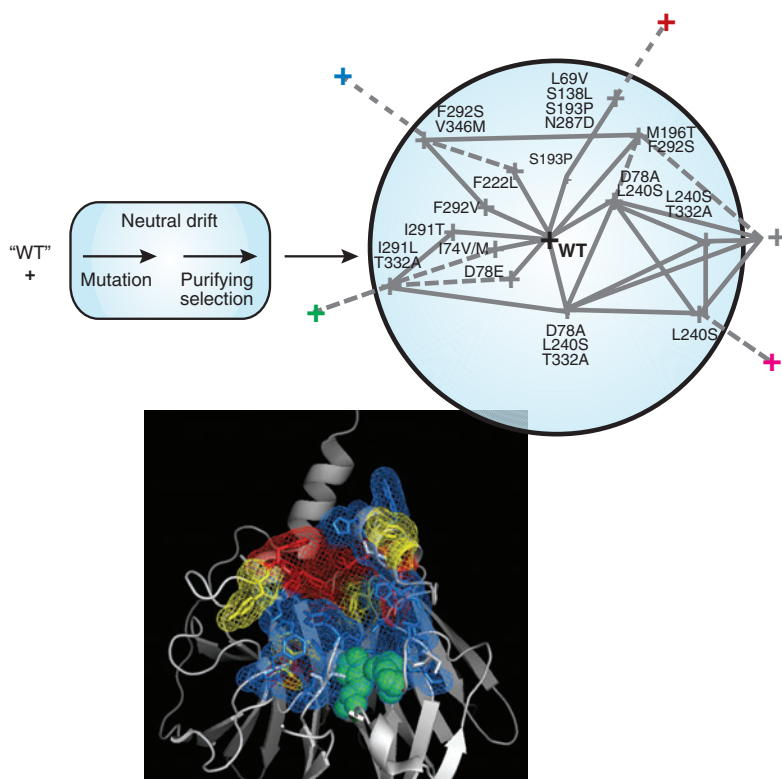


Figure 1 | Neutral drifts boost enzyme evolvability. An accelerated neutral drift performed with a lactonase dubbed PON1 introduced a wide range of apparently neutral mutations in the periphery and within the active site (blue mesh), and thus created an expanded neutral network (gray circle) from which various new specificities can stem (designated by peripheral crosses in various colors)²⁸. Mutated residues in red and yellow maintained lactonase activity at a level identical to and lower than wild-type, respectively. The His115,134 dyad that mediates PON1's lactonase activity (green spheres), and its immediate vicinity, remained intact in contrast to the extensive changes seen in other parts of the active site.

to obtain stabilities similar to those of their thermophilic homologs, yet without compromising their catalytic efficiency at lower temperatures. This suggests that the scarcity of natural enzymes displaying both properties is not the result of inherent physicochemical limitations, but is due to the rarity of mutations that increase stability while maintaining activity and to the absence of selection pressures for activity under such a broad range of temperatures⁶.

Protein promiscuity and evolution of new functions

As the methodologies of mutagenesis and selection have been developed, a somewhat painful realization emerged. A directed evolution experiment is unlikely to be successful (even if multiple selection rounds are performed) unless the starting point already possesses, albeit at a very low level, the activity one is selecting for. In other words, you cannot evolve what is not already there. So although the first rule of directed evolution is 'you get what you select for', this may be

followed by the second rule, 'you should select for what is already there'. Indeed, computational design may become a common source of starting points for optimization by directed evolution.

The 'starting point' paradox had not escaped the thought of scientists who previously addressed the question of how new protein functions evolve. The study of antibodies, for example, indicated that cross-reactivity, or multispecificity, is a crucial factor in any system in which a limited sequence diversity (for example, a naive antibody repertoire) gives rise to a virtually unlimited functional diversity (encountering any presented antigen). Indeed, as initially proposed by Pauling and later demonstrated experimentally, the existence of multiple conformers of the same antibody can mediate multispecificity⁷. The principle of multispecificity also has been applied to the evolution of enzymes. A hypothesis prompted by Jensen⁸, and later by O'Brien and Herschlag⁹, suggested that the earliest enzyme forms exhibited broad substrate specificity, or 'substrate ambiguity', and

later diverged to give the specific enzymes we know today.

Research into enzyme superfamilies has also revealed overlapping promiscuous activities between members of the same superfamily. It was found that the primary activity of one member can be the promiscuous activity of another, and that promiscuous activities can serve as starting points for the evolution of new family members^{10,11}. Thus, both natural and laboratory evolution appear to diverge new functions by preserving a protein's structural scaffold and key catalytic residues¹¹ while tinkering with the active-site loops, often with residues that are distal to the active site and affect loop conformations¹² (as with antibodies). Such changes can modify the substrate-binding mode and thus generate new specificities. More substantial changes in function may involve the swapping or grafting of entire loops¹³, possibly by homologous recombination (shuffling) of different family members, where the conserved scaffold and catalytic residues serve as anchoring points for recombination.

Additional understanding of the role of promiscuity in evolution came from directed evolution experiments that systematically explored how promiscuous activities evolve under selection. As it turned out, mutations selected for their large positive effects on the evolving promiscuous activity appear to have a much milder effect and sometimes almost no effect on the primary (native) function of the protein, despite the latter not being under selection¹². This weak tradeoff between old and new function suggested that such mutations could comprise the first steps in the divergence of a new function. The wider implications of this tradeoff are that a new function can develop, at least to some extent, before gene duplication, much in the spirit of the 'gene sharing' model¹⁴.

These observations are also in line with findings indicating that enzymes evolving toward a new activity can catalyze substrates for which there was no selection¹⁵. Such 'generalist' intermediates may resemble Jensen's early enzyme forms⁸ and ancestors, or evolutionary nodes¹⁶, from which a family of specialized enzymes could diverge. How do such generalists turn into specialists? In the later steps of divergence, the new function might trade off strongly with the old one. In addition, specific selection forces may remove variants still possessing the old activity¹⁷. Indeed, protein engineers are realizing that introducing a new function into an existing protein through either

computational design or directed evolution might be straightforward but erasing the old function is not. Hence, a ‘dual selection’—a selection for the new function and against the old one—often needs to be applied to accelerate specialization^{18,19}.

Reconstructing molecular fossils

Our understanding of how biomolecules evolved in nature is derived primarily from examining the endpoints of the evolutionary processes that yielded them. The snag is that our sampling is extremely sporadic. Mammals, for example are overrepresented, whereas beetles (the inspiration of Haldane’s famous theology-biology insight), of which hundreds of thousands species exist, are overwhelmingly underrepresented. As Darwin phrased it, our records provide “a history of the world, imperfectly kept and written in changing dialect. Of this history, we possess the last volume alone. Of this volume, only here and there a short chapter has been preserved; and of each page only here and there few lines.”²⁰ This is why, alongside other approaches such as the reconstruction of ancestor proteins and adaptive paths²¹, laboratory evolution can enhance our understanding of evolutionary pathways in a unique way.

As first demonstrated by Clarke and Hall^{4,5}, laboratory evolution enables studies of key intermediates along mutational paths leading from one function, or structure, into another. This variety of work fills substantial gaps in our understanding of evolutionary processes, much in the same way that ‘missing link’ fossils clarify organismal evolution.

Reconstructing the evolution of protein folds

Whereas the evolution of protein function has been the focus of many studies, the evolution of protein structures and folds has been addressed less frequently. Examples include the selection of stable, folded proteins by complementing a fragment of an existing protein that in itself does not form a stable structure, with short, randomly truncated segments of the *E. coli* genome. These experiments indicated that novel protein domains are easily created by non-homologous recombination of segments from unrelated proteins. Notably, these novel domains are maintained by oligomerization²², thus supporting the hypothesized role of oligomerization in primordial protein folds. In fact, the distinctively modular nature of protein folds such as TIM (β/α) barrels and β -propellers suggests that these proteins also emerged

from smaller elements that assembled non-covalently. We reconstructed this scenario by selecting 100-amino-acid segments from tachylecin-2 (a 5-bladed β -propeller with five sugar-binding sites) that spontaneously assemble into stable homo-pentamers and exhibit the same function as the intact monomeric 236-amino-acid wild-type protein²³. This experiment substantiates the hypothesis that proteins with high internal symmetry evolved from short, oligomerizing segments that, at later stages, duplicated, fused and rearranged—ultimately yielding the folds we recognize today.

The latter mechanisms (duplication, fusion and rearrangement of whole gene segments) comprise the routes by which new protein folds and topologies emerged through evolution²⁴. One such postulated route—‘permutation by duplication’—is thought to be responsible for extensive divergence within several protein superfamilies. This process is based on duplication and in-frame fusion of existing genes followed by partial degeneration of the 5’ and 3’ coding regions of the covalently fused dimer, ultimately leading to the topological change known as circular permutation. However, failure to identify existing partially truncated intermediates has questioned the validity of the permutation by duplication route.

With the aim of examining such complex gene rearrangements, we reproduced this route by directed evolution²⁵ (related topological rearrangements were recently explored in the chelatase family²⁶). We selected functional truncated intermediates and thus revealed inherent modularities within single domains of the DNA methyltransferase family. These intermediates led to circular permutants that either resembled three known families or belonged to a new family. The latter was subsequently identified in bacterial genomes. Thus, directed evolution has the power not only to reconstruct evolutionary pathways, but also to predict the existence of natural forms that have not yet been identified.

Protein engineers have also capitalized on these rearrangements and applied them to evolve novel enzyme variants with improved specificity and rates²⁷.

Neutral drifts—a novel tool for protein engineering

A recent development at the interface of protein engineering and molecular evolution regards laboratory reproductions of ‘neutral drifts’—that is, the gradual accumulation of

mutations under selection to maintain a protein’s original function and structure (purifying or nonadaptive selection; see **Box 1** for definitions). Such drifts have now been reproduced in two laboratories—Arnold’s and ours^{3,28–31}. For a protein engineer, evolving a protein toward something it already does well (namely its native function) makes little sense. So why do it?

Most of the genetic diversity of this planet—including the vast majority of sequence differences between members of the same protein family, such as polymorphs, orthologs and even paralogs, is the outcome of nonadaptive changes. It is assumed that many of these amino acid exchanges are not entirely neutral, although their fixation might be coincidental rather than the result of adaptive forces³². Despite the ubiquity of such neutral changes, we know little about them. For example, theoreticians have predicted fascinating properties of ‘neutral networks’ or quasi-species (such as higher stability, mutational robustness and evolvability) that develop when proteins drift to create a ‘cloud’ of sequences around the wild-type sequence³³.

Neutral drifts of two different enzymes, P450-BM3 and serum paraoxonase (PON1), demonstrated that the potential for adaptation develops dramatically when the neutral network of a protein expands (**Fig. 1**). In the PON1 drift almost half of the 311 neutral variants characterized exhibited substantial changes in promiscuous activities, specificities or inhibition (simulating the evolution of drug resistance)²⁸. Similar fluctuations in promiscuous activities were observed in the P450 neutral variants³⁰. Several of the PON1 neutral variants isolated were one or even two mutations closer to a potential new phenotype (aryl esterase, thiolactonase, phosphotriesterase or ‘drug resistance’). Indeed, a subsequent screen of these 311 variants with a new substrate (an analog of the nerve agent cyclosarin) yielded two variants with up to 72-fold higher activity relative to wild-type PON1²⁸.

These data throw up an intriguing dichotomy. In the field of protein engineering the accepted dogma on library size is that ‘big is beautiful’. Why, then, is it that such a small library of only 311 members can yield such a range of improved variants? The reason is that standard libraries used for directed evolution comprise a majority of ‘dead’ variants due to deleterious mutations. In contrast, the neutral variants were all folded and active. Nevertheless, these variants carry numerous mutations in, and around, the active site.

BOX 1 DEFINITIONS

Neutral drift: sequence changes occurring under nonadaptive regimes (see purifying selection below) while maintaining the original structure and function of an organism, and/or a protein.

Accelerated neutral drifts: comprise iterative rounds of *in vitro* random mutagenesis at relatively high rates (~2 mutations per gene per round) and selection of protein variants that maintain the protein's native activity and expression level.

Purifying selection: the removal of deleterious or harmful mutations (or alleles) in a population or a gene library under nonadaptive evolution, namely while maintaining the original structure and function of an organism and/or a protein. A synonymous term is **negative selection**.

Positive (or adaptive) selection: the fixation of advantageous mutations or alleles in a population under changing circumstances, for example, selection of mutations that endow an enzyme with a new function.

Some of these mutations have the ability to mediate new specificities by dramatically increasing existing promiscuous activities, while exerting little effect on native activity which was selected for^{28,30}.

The TEM-1 drift we performed also measured how the fitness (activity, stability and others) of a protein changes as mutations accumulate. It provided experimental proof and biophysical explanation of a phenomenon fundamental to evolutionary dynamics and previously seen only in computer simulations: a tight correlation between negative epistasis and robustness³. The P450 drift directly demonstrated that increased stability and mutational tolerance underlies the large, polymorphic populations that a neutral drift can yield³¹. The TEM-1 experiment indicated that mutations that act as 'global suppressors' are enriched during the neutral drift. These mutations increase TEM-1's stability, suppress the effect of a broad range of destabilizing mutations and thereby substantially increase the probability of a new function emerging²⁹.

Notably, all identified global suppressors emerged in positions where the sequence of TEM-1 deviates from its family (and predicted ancestor) consensus, and comprise back-to-consensus/ancestor mutations. Such changes, readily predicted from sequence alignments, have been shown to stabilize numerous proteins³⁴. Incorporating these stabilizing mutations into proteins should enable protein engineers to generate variants that comprise highly evolvable starting points.

These seemingly theoretical studies^{3,28–31} prompt completely new ways of performing protein evolution *in vitro*. To date, all directed

evolution experiments have been performed by random mutation of the starting gene and selection for the desired new function from the resulting libraries. However, alternative approaches may yield better results with much smaller libraries. A neutral drift can be first performed, resulting in a library that has increased stability and a large although nondeleterious variation. As such, these libraries—although small in size—can possess a vastly increased potential to evolve new functions. In addition, enriching libraries with consensus³⁴ or ancestor mutations³⁵ can also boost the protein's tolerance to mutations and compensate for the deleterious effects of function-altering mutations^{2,36}, thus dramatically increasing the frequency of variants conferring new functions.

In summary, we hope to have shown how complementary applied and fundamental aspects of molecular evolution can be. Evolutionary biology provides crucial theoretical insights from which one can design experiments that examine further these theories and models. By doing so, the veracity of various hypotheses regarding the evolution of new protein structures and functions can be tested. The 'pot of gold' at the end of this scientific rainbow contains a better and more profound understanding of natural evolution and new, more efficient ways of engineering proteins in the laboratory.

ACKNOWLEDGMENTS

This brief and very personal commentary addresses the interface between protein engineering and fundamental molecular evolution. The reference list is therefore rather subjective, and we apologize for not citing a myriad of interesting and relevant work. Research Grants by the Israel Science Foundation and the Estate of Fannie Sherr are gratefully acknowledged.

S.G.P. was the recipient of a D. Stone Postdoctoral Fellowship from the Feinberg Graduate School.

1. Brannigan, J.A. & Wilkinson, A.J. *Nat. Rev. Mol. Cell Biol.* **3**, 964–970 (2002).
2. Wang, X., Minasov, G. & Shoichet, B.K. *J. Mol. Biol.* **320**, 85–95 (2002).
3. Bershtein, S., Segal, M., Bekerman, R., Tokuriki, N. & Tawfik, D.S. *Nature* **444**, 929–932 (2006).
4. Brown, J.E., Brown, P.R. & Clarke, P.H. *J. Gen. Microbiol.* **57**, 273–285 (1969).
5. Hall, B.G. *FEMS Microbiol. Lett.* **174**, 1–8 (1999).
6. Arnold, F.H., Wintrode, P.L., Miyazaki, K. & Gershenson, A. *Trends Biochem. Sci.* **26**, 100–106 (2001).
7. James, L.C., Roversi, P. & Tawfik, D.S. *Science* **299**, 1362–1367 (2003).
8. Jensen, R.A. *Annu. Rev. Microbiol.* **30**, 409–425 (1976).
9. O'Brien, P.J. & Herschlag, D. *Chem. Biol.* **6**, R91–R105 (1999).
10. Afriat, L., Roodveldt, C., Manco, G. & Tawfik, D.S. *Biochemistry* **45**, 13677–13686 (2006).
11. Glasner, M.E., Gert, J.A. & Babbitt, P.C. *Curr. Opin. Chem. Biol.* **10**, 492–497 (2006).
12. Aharoni, A. *et al. Nat. Genet.* **37**, 73–76 (2005).
13. Park, H.S. *et al. Science* **311**, 535–538 (2006).
14. Piatigorsky, J. *Gene sharing and evolution* (Harvard University Press, Boston, 2007).
15. Matsumura, I. & Ellington, A.D. *J. Mol. Biol.* **305**, 331–339 (2001).
16. Wouters, M.A., Liu, K., Riek, P. & Husain, A. *Mol. Cell* **12**, 343–354 (2003).
17. Tawfik, D.S. *Science* **311**, 475–476 (2006).
18. Varadarajan, N., Gam, J., Olsen, M.J., Georgiou, G. & Iverson, B.L. *Proc. Natl. Acad. Sci. USA* **102**, 6855–6860 (2005).
19. Collins, C.H., Leadbetter, J.R. & Arnold, F.H. *Nat. Biotechnol.* **24**, 708–712 (2006).
20. Graur, D. & Li, W.-H. *Fundamentals of Molecular Evolution* (University of Chicago Press, Chicago, 2000).
21. Dean, A.M. & Thornton, J.W. *Nat. Rev. Genet.* **8**, 675–688 (2007).
22. de Bono, S., Riechmann, L., Girard, E., Williams, R.L. & Winter, G. *Proc. Natl. Acad. Sci. USA* **102**, 1396–1401 (2005).
23. Yadid, I. & Tawfik, D.S. *J. Mol. Biol.* **365**, 10–17 (2007).
24. Koonin, E.V., Wolf, Y.I. & Karev, G.P. *Nature* **420**, 218–223 (2002).
25. Peisajovich, S.G., Rockak, L. & Tawfik, D.S. *Nat. Genet.* **38**, 168–174 (2006).
26. Pisarchik, A., Petri, R. & Schmidt-Dannert, C. *Protein Eng. Des. Sel.* **20**, 257–265 (2007).
27. Qian, Z. & Lutz, S. *J. Am. Chem. Soc.* **127**, 13466–13467 (2005).
28. Amitai, G., Devi-Gupta, R. & Tawfik, D.S. *HFSP J.* **1**, 67–78 (2007).
29. Bershtein, S. & Tawfik, D.S. (submitted).
30. Bloom, J.D., Romero, P.A., Lu, Z. & Arnold, F.H. *Biol. Direct* **2**, 17 (2007).
31. Bloom, J.D. *et al. BMC Biol.* **5**, 29 (2007).
32. Kimura, M. *Phil. Trans. R. Soc. Lond. B* **312**, 343–354 (1986).
33. van Nimwegen, E., Crutchfield, J.P. & Huynen, M. *Proc. Natl. Acad. Sci. USA* **96**, 9716–9720 (1999).
34. Lehmann, M. *et al. Protein Eng.* **15**, 403–411 (2002).
35. Watanabe, K., Ohkuri, T., Yokobori, S. & Yamagishi, A. *J. Mol. Biol.* **355**, 664–674 (2006).
36. Bloom, J.D., Labthavikul, S.T., Otey, C.R. & Arnold, F.H. *Proc. Natl. Acad. Sci. USA* **103**, 5869–5874 (2006).