Minireview

Anticipatory evolution and DNA shuffling
Jamie M Bacher*, Brian D Reiss† and Andrew D Ellington*†‡

Addresses: *Institute for Cellular and Molecular Biology, †Center for Nano- and Molecular Science and Technology and ‡Department of Chemistry and Biochemistry, University of Texas at Austin, Austin, TX 78712, USA.

Correspondence: Andrew D Ellington. E-mail: andy.ellington@mail.utexas.edu

Abstract

DNA shuffling has proven to be a powerful technique for the directed evolution of proteins. A mix of theoretical and applied research has now provided insights into how recombination can be guided to more efficiently generate proteins and even organisms with altered functions.

Proteins are machines created by evolution, but it is unclear just how finely evolution has guided their sequence, structure, and function. It is undoubtedly true that individual mutations in a protein affect both its structure and its function and that such mutations can be fixed during evolutionary history, but it is also true that there are other elements of protein sequence that have been acted upon by evolution. For example, the genetic code appears to be laid out so that mutations and errors in translation are minimally damaging to protein structure and function [1]. Could the probability that a beneficial mutation is found and fixed in the population also have been manipulated during the course of evolution, so that the proteins we see today are more capable of change than the proteins that may have been cobbled together following the 'invention' of translation? Have proteins, in fact, evolved to evolve? There is already some evidence that bacteria are equipped to evolve phenotypes that are more capable of further adaptation (reviewed in [2-4]). For example, mutator [5] and hyper-recombinogenic [6] strains arise as a result of selection experiments. The development of DNA shuffling (reviewed in [7,8]) and the appearance of several recent papers using this technique [9-11] provide us with a surprising new opportunity to ask and answer these fundamental questions at the level of individual genes, and perhaps even genomes.

DNA shuffling, a method for *in vitro* recombination, was developed as a technique to generate mutant genes that would encode proteins with improved or unique functionality [12,13]. It consists of a three-step process that begins with the enzymatic digestion of genes, yielding smaller fragments of DNA. The small fragments are then allowed to randomly hybridize and are filled in to create longer fragments. Ultimately, any full-length, recombined genes that are re-created are amplified via the polymerase chain reaction. If a series of alleles or mutated genes is used as a starting point for DNA shuffling, the result is a library of recombined genes that can be translated into novel proteins, which can in turn be screened for novel functions. Genes with beneficial mutations can be shuffled further, both to bring together these independent, beneficial mutations in a single gene and to eliminate any deleterious mutations. Although multiple, beneficial mutations could potentially be generated just as well by serial mutagenesis and screening, DNA shuffling is much quicker: for example, the starting population of a library generated by mutagenic PCR typically contains 70-99% nonfunctional variants [14], whereas most variants formed by DNA shuffling are functional. Thus, DNA shuffling should allow a streamlined exploration of sequence space and acquisition of novel protein phenotypes easily, as has indeed proven to be the case for a number of protein targets [15,16].

Beyond biotechnology applications, DNA shuffling can potentially be used to recapitulate natural recombination and to ask whether recombination generally leads to better or novel proteins. In this regard, DNA shuffling can be carried out not only with genes that are closely related

alleles, but also with a group of phylogenetically related genes that may differ by up to 40%, a process known as family shuffling [15]. As mentioned above, it was strongly suspected that by starting with a population of genes already known to be functional, family shuffling could move the most beneficial mutations into the same gene and thus quickly optimize or alter protein function. In fact, however, this intuition should hold true only if mutant alleles can generally either act in an additive or synergistic fashion. If mutant alleles are neutral or interfere with each other, then there will be no generic benefit to recombination.

In order to address this hypothesis, Joern *et al.* [9] have developed a novel technique for mapping recombination events by probe-hybridization analysis. Shuffled libraries were generated by crossing genes for several dioxygenase: toluene dioxygenase, *tod*C1C2; tetrachlorobenzene dioxygenases, *tec*A1A2; and biphenyl dioxygenase, *bph*A1A2. Shuffled variants from the three-parent library were screened for toluene dioxygenase activity, and randomly selected variants were sequenced to determine the actual number of crossovers that had occurred to give rise to functional and nonfunctional variants. Unsurprisingly, it was found that crossovers commonly occurred in regions of high homology: although regions that contained ten or more common, identical residues made up less than 10% of the lengths of the genes, over 60% of the crossovers occurred in these regions. Interestingly, it was found that the number of crossover events did not correlate with protein function, suggesting that individual segments of a protein might act independently during evolution [9]. It is also possible that the proteins were so closely related to one another that multiple crossovers did not reduce or alter functionality.

Building on these results, Voigt *et al.* [10] hypothesized that functional genes derived by DNA shuffling (and perhaps by natural recombination) should preserve clustered sets of structural interactions (the so-called 'schemas') of the original protein (Figure 1a). In order to validate this hypothesis, the authors developed an algorithm that attempted to predict the effect of crossover events at specific sites in a gene. In particular, the algorithm assessed which amino acids were close to one another in both the primary and the tertiary protein structure and predicted which interaction subsets could be manipulated in a way that minimally disrupted protein structure and function. This analysis results in a 'schema profile' for the proteins, which indicates the amount of disruption to the schemas that recombination at each point along the sequence will cause (Figure 1b). Several proteins that had previously been evolved *in vitro* by family shuffling were evaluated, and the schema profiles of these proteins correlated well with the experimentally determined crossover points [14].

This algorithm was then used to generate schema profiles between two β-lactamases, TEM-1 and PSE-4, which confer
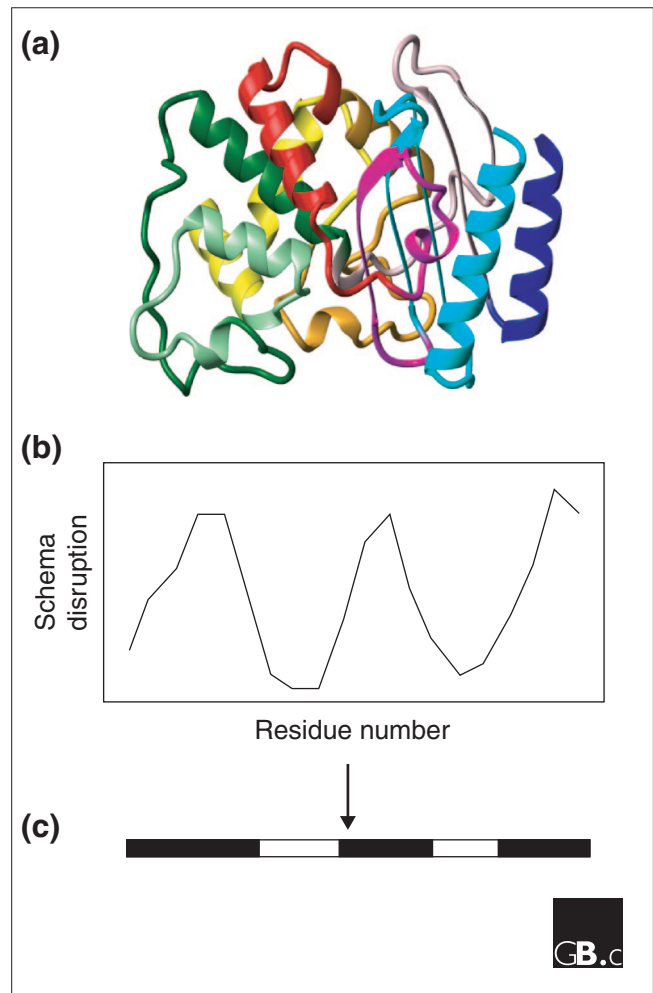
Figure 1

A graphical representation of the relationship between protein structure and schemas. **(a)** The β-lactamase protein is shown divided into different colored substructures (schemas), which are derived from the schema profile of the protein. **(b)** An example of a schema profile for a (simpler) hypothetical protein. Peaks correlate with positions in the protein where recombination will be maximally disruptive; valleys correlate with positions that are predicted to minimally disrupt the structure and function of the protein. **(c)** Intron structure may correlate with schema structure. To the extent it is now possible to calculate schema profiles, it can be hypothesized that introns (white) may generally fall at minima while exons (black) may generally contain larger disruption values.

ampicillin resistance and share only 40% amino-acid sequence identity. Hybrid enzymes that had varying degrees of recombination between schemas were then constructed, and the recombined variants were transformed into bacteria, which were assayed for ampicillin resistance. The most resistant hybrids contained recombined genes with crossovers that had been predicted in advance to occur between schemas [10].

What is particularly surprising is not that DNA shuffling occurs between domains; even a brief observation of the

three-dimensional structures of proteins immediately suggests that recombinational breakpoints will probably have the smallest effect on protein function if they occur outside of major structural units found by Voigt *et al.* [10] (although certain breakpoints between structural subunits, such as in the middle of α helices, would probably not have been predicted without schema profiling). Rather, the amazing thing is that proteins have evolved so that they are by and large composed of structural domains that can undergo recombination. As Voigt *et al.* [10] point out, Gô [17,18] found a correlation between intron locations and structural domains. This was expanded on by Gilbert and his co-workers [19], who advanced the notion that proteins could be modularly constructed from structural domains as an attempt to explain the origin of introns. Although the 'introns early' hypothesis has long since been shown to be implausible [20-22], the original notion that introns could act as buffers for recombination is still intellectually compelling, and it may be consistent with the results of Voigt *et al.* [10].

Interestingly, to the extent that proteins have evolved as modular machines that are capable of taking advantage of recombination during their evolutionary history, the very mathematical models propagated by Joern *et al.* [9] and Voigt *et al.* [10] may be unnecessary. 'Blind' DNA shuffling between closely related proteins may already be more than good enough to generate proteins with novel phenotypes. For example, we have evolved a β-glucuronidase *in vitro* to switch its substrate specificity from β-glucuronides to β-galactosides and have achieved an over 500-fold increase in activity towards the new substrate [23]. This catalytic conversion was achieved in three rounds of shuffling and screening, but further rounds of selection failed to achieve greater cleavage of β-galactosides. The initial library of this selection was constructed using mutagenic PCR, and a large fraction of the population was inactive, yet the catalytic specificity of the selection produced a switch of over 52-million-fold in substrate preference.

Similarly, new experiments from Zhang *et al.* [11] provide additional evidence that blind shuffling is fully capable of functional improvement, not just at the protein level, but even at the organismal level. These researchers coupled classical strain improvement (mutation and selection) with genetic recombination. Protoplast fusion results in very efficient recombination between the genomes of *Streptomyces* species, and iterative protoplast fusion results in the reassortment of multiple markers between species. To show the power of this new method, a *Streptomyces* strain producing the complex polyketide antibiotic tylosin was selected for improved function and was then forced to undergo the equivalent of sexual reproduction. The genomes of several surviving mutants were shuffled after every round of selection to generate a combinatorial library of organisms that could again be screened for improved function. A strain generated by only two rounds of shuffling could produce tylosin at a rate comparable to strains that had undergone 20 rounds of classical selection. These results demonstrate that genome shuffling will probably lead to changes and improvements in organismal function as radical as those that have previously been observed for proteins.

Overall, these results further support the idea that evolution can act reflexively - that is, to enhance its own ability to act. From the results of Arnold and co-workers [9,10], it is possible that regions that fall between predicted schema might be conserved in sequence in order to facilitate recombination; this hypothesis could be checked directly by database analysis. The application of the techniques described by Arnold and co-workers [9,10], del Cardayré and co-workers [11], and others may allow researchers to more effectively design libraries for screening. A large fraction of the products generated in traditional screening or even shuffling reactions are nonfunctional. Schema profiling and pathway shuffling may eventually make it possible to design directed evolution experiments in which structural and metabolic subunits are preserved, thereby limiting the exploration of sequence space largely to functional molecules. Ultimately, these advances should expand our understanding of natural genetic processes and thereby allow biologists to generate novel proteins and pathways in a fraction of the time that nature or conventional breeding would take.

Acknowledgements

References

1. Freeland SJ, Hurst LD: **The genetic code is one in a million.** *J Mol Evol* 1998, **47:**238-248.
2. Radman M, Matic I, Taddei F: **Evolution of evolvability.** *Ann NY Acad Sci* 1999, **870:**146-155.
3. Radman M, Taddei F, Matic I: **Evolution-driving genes.** *Res Microbiol* 2000, **151:**91-95.
4. Tenaillon O, Taddei F, Radmian M, Matic I: **Second-order selection in bacterial evolution: selection acting on mutation and recombination rates in the course of adaptation.** *Res Microbiol* 2001, **152:**11-16.
5. Sniegowski PD, Gerrish PJ, Lenski RE: **Evolution of high mutation rates in experimental populations of *E. coli*.** *Nature* 1997, **387:**703-705.
6. Guttman DS, Dykhuizen DE: **Clonal divergence in *Escherichia coli* as a result of recombination, not mutation.** *Science* 1994, **266:**1380-1383.
7. Farinas ET, Bulter T, Arnold FH: **Directed enzyme evolution.** *Curr Opin Biotechnol* 2001, **12:**545-551.
8. Kolkman JA, Stemmer WP: **Directed evolution of proteins by exon shuffling.** *Nat Biotechnol* 2001, **19:**423-428.
9. Joern JM, Meinhold P, Arnold FH: **Analysis of shuffled gene libraries.** *J Mol Biol* 2002, **316:**643-656.
10. Voigt CA, Martinez C, Wang ZG, Mayo SL, Arnold FH: **Protein building blocks preserved by recombination.** *Nat Struct Biol* 2002, **9:**553-558.
11. Zhang YX, Perry K, Vinci VA, Powell K, Stemmer WP, del Cardayré SB: **Genome shuffling leads to rapid phenotypic improvement in bacteria.** *Nature* 2002, **415:**644-646.
12. Stemmer WP: **DNA shuffling by random fragmentation and reassembly: *in vitro* recombination for molecular evolution.** *Proc Natl Acad Sci USA* 1994, **91:**10747-10751.

13. Stemmer WP: **Rapid evolution of a protein *in vitro* by DNA shuffling.** *Nature* 1994, **370:**389-391.
14. Matsumura I, Ellington AD: **Mutagenic PCR of protein-coding genes for *in vitro* evolution.** In *In Vitro Mutagenesis Protocols.* Edited by Braman J. Totowa, NJ: Humana; 2001.
15. Crameri A, Raillard SA, Bermudez E, Stemmer WP: **DNA shuffling of a family of genes from diverse species accelerates directed evolution.** *Nature* 1998, **391:**288-291.
16. Ness JE, Welch M, Giver L, Bueno M, Cherry JR, Borchert TV, Stemmer WP, Minshull J: **DNA shuffling of subgenomic sequences of subtilisin.** *Nat Biotechnol* 1999, **17:**893-896.
17. Gô M: **Correlation of DNA exonic regions with protein structural units in haemoglobin.** *Nature* 1981, **291:**90-92.
18. Gô M: **Modular structural units, exons, and function in chicken lysozyme.** *Proc Natl Acad Sci USA* 1983, **80:**1964-1968.
19. Gilbert W, Glynias M: **On the ancient nature of introns.** *Gene* 1993, **135:**137-144.
20. Palmer JD, Logsdon JM Jr: **The recent origins of introns.** *Curr Opin Genet Dev* 1991, **1:**470-477.
21. Cavalier-Smith T: **Intron phylogeny: a new hypothesis.** *Trends Genet* 1991, **7:**145-148.
22. Rogers JH: **The role of introns in evolution.** *FEBS Lett* 1990, **268:**339-343.
23. Matsumura I, Ellington AD: ***In vitro* evolution of beta-glucuronidase into a beta-galactosidase proceeds through non-specific intermediates.** *J Mol Biol* 2001, **305:**331-339.