Combinatorial protein design by *in vitro* recombination
Lori Giver and Frances H Arnold*

DNA recombination is a powerful engine for the creation of new phenotypes. Recently, methods for *in vitro* DNA recombination (DNA shuffling) have been developed and applied to the evolution of novel molecules in the laboratory. An exciting new development is the shuffling of homologous genes to create diversity for directed evolution.

Addresses
Division of Chemistry and Chemical Engineering 210-41, California Institute of Technology, Pasadena, CA 91125, USA
*e-mail: frances@cheme.caltech.edu

Abbreviations
GFP   green fluorescent protein
StEP   staggered extension process

Introduction
The generation of genetic diversity through DNA recombination is a key process in adaptation and evolution. Most notably, the result of sexual reproduction is offspring with novel combinations of the parental genes. Although the driving force behind the origin and maintenance of sexual recombination is still debated [1], the benefits of a process that allows the joining of beneficial mutations and the removal of deleterious ones appear to outweigh the costs. There are other *in vivo* mechanisms for DNA recombination in addition to genomic DNA recombination during sexual reproduction. For example, these mechanisms generate high diversity in viral populations [2] and the vertebrate immune system [3]. This powerful feature of natural evolution has been exploited in recent years by scientists trying to produce novel biological molecules by directed evolution. This review focuses on *in vitro* methods for DNA recombination (often referred to as 'DNA shuffling') and their application to the generation of gene libraries for directed evolution, which is a highly combinatorial approach to protein design. To date, these methods have been used primarily to evolve novel proteins, particularly enzymes (for a recent review of directed enzyme evolution, see [4]). For a different perspective on this topic and for a review of significant as yet unpublished work on applications of DNA shuffling for protein pharmaceuticals and DNA vaccines, the readers are directed to the recent review of Patten *et al.* [5].

In directed evolution, key processes of Darwinian evolution — random mutagenesis, recombination and selection (or screening) — are mimicked in order to evolve novel biological molecules. The power of, and need

for, evolutionary approaches to solve complex biological design problems have been well described [5,6]. Before Stemmer introduced DNA shuffling in 1994 [7,8], genetic diversity for directed protein evolution was generated primarily by point mutagenesis (error prone PCR, chemical mutagenesis, mutator strains, etc.) or by combinatorial cassette mutagenesis [9]. Recursive ensemble mutagenesis, a theoretically elegant although experimentally somewhat cumbersome approach, represented an early attempt to exploit features of genetic algorithms in directed evolution [10,11].

Evolution by sequential random mutagenesis is the equivalent of an asexual evolutionary process. If deleterious mutations appear in conjunction with beneficial mutations, they become fixed and can even limit the evolutionary potential (similar to the process known to population geneticists as Müller's ratchet [12]). In addition, all improved enzymes except the one chosen to parent the next generation are discarded. Beneficial mutations present in the other genes are therefore lost until they are re-discovered in subsequent generations.

DNA shuffling
In 1994, Stemmer [7,8] introduced the method of DNA shuffling for the *in vitro* formation of recombinant genes from a set of parental genes. DNA shuffling has already had a significant impact on the nascent field of directed evolution; each new demonstration has expanded the method's utility, generating broad academic and commercial interest. The Stemmer DNA shuffling method is well described in several places [5,7,8]. Briefly, the parental genes are randomly fragmented by DNase I. The purified fragments are extended by repeated cycles of overlap extension into full-length genes that contain novel combinations of the parental mutations. Recombination comes about when fragments derived from different parental genes prime one another. Point mutations are also introduced during the shuffling process.

Crameri *et al.* [13] used DNA shuffling to evolve green fluorescent protein (GFP) variants with increased fluorescence. After three rounds of DNA shuffling and screening, a clone was isolated with an average fluorescence intensity more than 40 times that of wild type. The amino acid substitutions did not change the chromophore, but instead increased the solubility of GFP in both the *Escherichia coli* and Chinese hamster ovary host cells. The best first-generation mutant had seven nonsynonymous mutations, four of which reverted in the second and third generations. Because introducing new mutations is not always desirable, especially at this high rate, variations of the shuffling method have been developed to control the mutagenesis rate [14,15].

That the solution to the problem posed for the evolution of GFP fluorescence intensity was to increase protein solubility in the host cell rather than alter the chromophore underscores two recurring themes of directed evolution: you get what you screen for [16] and the mutations that arise are rarely those that would have been made in a rational design effort. Often located far from an active site, the mutations accumulated during directed evolution influence protein structure and function in ways that are not easy to discern, although their effects on protein function are readily observed in screening assays. The great power of directed evolution is that it provides a means to engineer poorly understood functions and requires no knowledge of the protein structure.

Substrate specificity and activity towards novel substrates have been evolved by DNA shuffling alone [17], by DNA shuffling in combination with random mutagenesis [18•] and, most recently, by DNA shuffling contaminated with rational design [19•]. Schultz and co-workers [19•] changed the specificity of a tRNA synthetase for use in the site-specific incorporation of non-natural amino acids *in vivo*. Based on the crystal structure of the wild type tRNA–synthetase complex, these workers incorporated degenerate oligos into the shuffling reactions (as described in [8]) in order to target mutations to sites believed particularly important for recognition of the tRNA by the synthetase. It is unclear whether the rational targeting of specific sites played a beneficial role in the generation of the desired synthetases, which exhibit reduced affinity for the wild type tRNA and increased affinity for the suppressor tRNA. In fact, the synthetase with the highest activity on the suppressor tRNA had no mutations at four of the five targeted positions.

Earlier, Crameri *et al.* [20] used a similar approach, incorporating synthetic oligonucleotides that correspond to the complementarity-determining regions of the Fv fragment into the DNA shuffling reactions in order to construct a diverse antibody library. This library was displayed on phage; antibodies that tightly and specifically bound human granulocyte colony-stimulating factor (G-CSF) receptor were isolated. Although these clones may be useful as diagnostic tools, the fact that they contain many mutations, even after attempts to remove neutral mutations by back-crossing (recombining the evolved sequences with the original), and do not bind the receptor when in soluble form (you get what you screen for) make them less attractive for therapeutic applications. The use of degenerate oligonucleotides in DNA shuffling to increase library diversity at specific sites may be useful under some circumstances; however, a very high level of mutations clouds the ability to identify functional mutations and increases the potential for immunogenic reactions in therapeutic applications. Furthermore, our understanding of these molecules is often so limited that attempts to rationally target specific sites may instead end up impeding the evolutionary progress.

Although most of the DNA shuffling experiments reported to date have been carried out on small, monomeric enzymes, neither the technique itself nor directed evolution are limited in this way. Zhang *et al.* [17] evolved a galactosidase into a more efficient fucosidase in seven rounds of shuffling and screening. The wild type galactosidase is a complex composed of four 1023 amino acid subunits. Mutants with significant shifts in substrate specificity were identified by screening moderately sized libraries (~40,000 variants) using a colorimetric plate assay. Gene operons have also been evolved by DNA shuffling. The three-gene, 2.3 kilobase arsenic resistance operon was evolved to confer 40-fold increased resistance to arsenate and fivefold increased resistance to arsenite in three rounds of DNA shuffling and selection [21].

The evolutionary experiments described above all began with a single parental sequence. Mutations for recombination were produced by the point mutagenesis associated with DNA shuffling. Alternatively, they can be produced separately under controlled conditions, for example by PCR mutagenesis [18•]. Another, potentially very rich, source of diversity, however, are homologous genes from nature. Provided these genes share enough identity to cross-prime one another, they can be 'family shuffled.' Crameri *et al.* [22••] recently shuffled four cephalosporinase genes (57–82% identical at the DNA sequence level) to create a library of mosaic genes from which enzymes conferring high resistance to moxalactam could be isolated. An evolved enzyme exhibiting this activity contained sequences from three of the four parent cephalosporinases. Surprisingly, it also contained 33 new amino acid mutations (i.e. amino acids not found in the parents). It is not yet clear how these mutations came about, nor how such a high level of mutagenesis was tolerated by the evolutionary process. It was observed that shuffling the four genes together evolved the new activity faster than DNA shuffling performed on individual genes. Whether this is generally true for the evolution of novel activities is a fundamentally important question. Recent theoretical studies have proposed that neutral mutations can play a key role in adaptation by providing access to much larger fractions of sequence space [23]. It is proposed that evolution along a set of neutral sequences will eventually lead to a 'jumping off' point for dramatic improvements in function.

Naturally occurring homologous genes are clearly a rich source of neutral or nearly neutral mutations; however, it is not yet clear how best to exploit this diversity in laboratory evolution. One major problem is that screening limitations are often severe. Even a good selection is limited to searching an infinitesimally small fraction of the sequences that can be generated by DNA shuffling;

for example, the degree to which evolutionary potential is increased by recombining homologous genes is still unknown, although it is likely to be considerable. The additional screening requirement generated by diluting beneficial mutations with neutral ones, however, is a real problem in the laboratory!

These fundamentally important strategic issues are only beginning to be addressed. A recent simple analysis of DNA shuffling of single mutants of a gene illustrated how the screening required to identify combinations of mutations increases very rapidly with the number of parental genes and mutations [18•]. The more beneficial mutations present in the parental genes, the greater the improvements in phenotype that can be generated; however, as the number of mutations present in the shuffling reaction increases, so does the number of progeny with phenotypes less fit than the parental enzymes. Strategies for optimizing the success of a directed evolution experiment are discussed [18•].

A major challenge in interpreting the results of both natural and artificial molecular evolution is identifying the functional mutations in a background of neutral mutations. This neutral background is especially high in natural homologous genes. DNA shuffling can be used to identify functional mutations efficiently by back-crossing under high-fidelity conditions [24•].

Stemmer's DNA shuffling method is fairly robust; however, it can prove technically challenging with certain sequences and suffers some limitations. Two new methods for shuffling single- or double-stranded polynucleotide sequences have been described in this review period: random priming recombination (RPR) [25•] and the staggered extension process (StEP) [26•]. Because these methods are based on the *de novo* synthesis of recombinant products, they require a much smaller amount of template and do not need the DNAse I fragmentation step. In the RPR method, short random primers are annealed to the template and extended by polymerase. The resulting fragments, the length of which can be controlled by altering the conditions of the annealing and extension reaction, are then separated from the initial template and unextended primers. These fragments are assembled into full length genes by cycles of overlap extension, similar to the Stemmer protocol. The StEP method, illustrated in Figure 1, uses template switching during synthesis to form the desired chimeric genes. The templates are mixed with one or more primers and subjected to repeated cycles of denaturation and short annealing/extension steps. Because the growing fragments can anneal to different templates, the resulting full length sequences contain sequence information from different parents. Because this method does not require a purification step, it is probably the simplest method technically for *in vitro* recombination. A thorough comparison of the various recombination methods, however, has not been performed.

Figure 1

Current Opinion in Chemical Biology
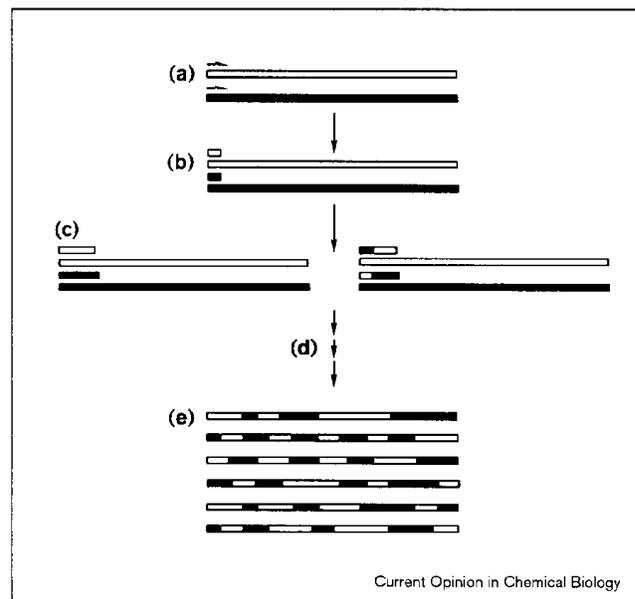
The recombination of two genes to create chimeric sequences using the staggered extension process (StEP) method [26•]. Only one primer and single strands from two template genes are shown. (a) Primers are added to denatured template genes and (b) short fragments are produced by brief polymerase-catalyzed extension. (c) In the next cycle of StEP, these fragments randomly prime the templates (template switching) and extend further. (d) This process is repeated until full-length genes are produced. (e) Full-length genes are purified and (optional) amplified in a PCR reaction with external primers. The original template molecules can be removed from the newly synthesized material by digestion with the restriction enzyme Dpn I, reducing the number of parental-like genes in the screened libraries.

Conclusions

Over the course of millions of years, nature has developed an effective strategy for adaptation and evolution. Researchers in the field are striving to mimic the key features of this design strategy in order to evolve interesting and useful molecules in the laboratory. With the introduction of new methods for *in vitro* gene recombination, we can accumulate beneficial mutations and remove deleterious ones more efficiently. The availability of these new methods now raises the question of how to best implement them.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1.   Hurst LD, Peck JR: **Recent advances in understanding of the evolution and maintenance of sex.** *Trends Ecol Evol* 1996, 11:46-52.

2.   Kuwata T, Miyazaki Y, Igarashi T, Takehisa J, Hayami M: **The rapid spread of recombinants during a natural *in vitro* infection with**

two human immunodeficiency virus type 1 strains. *J Virol* 1997, **71**:7088-7091.

3. Tonegawa S: **Somatic generation of antibody diversity.** *Nature* 1983, **302**:572-581.

4. Kuchner O, Arnold FH: **Directed evolution of enzyme catalysts.** *Trends Biotechnol* 1997, **15**:523-530.

5. Patten P, Howard RJ, Stemmer WPC: **Applications of DNA shuffling to pharmaceuticals and vaccines.** *Curr Opin Biotechnol* 1997, **8**:724-733.

6. Arnold FH: **Design by directed evolution.** *Accounts Chem Res* 1998, **31**:125-131.

7. Stemmer WPC: **Rapid evolution of a protein *in vitro* by DNA shuffling.** *Nature* 1994, **370**:389-391.

8. Stemmer WPC: **DNA shuffling by random fragmentation and reassembly: *in vitro* recombination for molecular evolution.** *Proc Natl Acad Sci USA* 1994, **91**:10747-10751.

9. Black ME, Newcomb TG, Wilson HMP, Loeb LA: **Creation of drug-specific herpes simplex virus type 1 thymidine kinase mutants for gene therapy.** *Proc Natl Acad Sci USA* 1996, **93**:3525-3529.

10. Delagrave S, Youvan DC: **Searching sequence space to engineer proteins: exponential ensemble mutagenesis.** *Bio-Technology* 1993, **11**:1548-1552.

11. Arkin AP, Youvan DC: **An algorithm for protein engineering: simulations of recursive ensemble mutagenesis.** *Proc Natl Acad Sci USA* 1992, **89**:7811-7815.

12. Müller HJ: **The relation of recombination to mutational advance.** *Mutat Res* 1964, **1**:2-9.

13. Crameri A, Whitehorn EA, Tate E, Stemmer WPC: **Improved green fluorescent protein by molecular evolution using DNA shuffling.** *Nat Biotechnol* 1996, **14**:315-319.

14. Zhao H, Arnold FH: **Optimization of DNA shuffling for high fidelity recombination.** *Nucleic Acids Res* 1997, **25**:1307-1308.

15. Lorimer IAJ, Pastan I: **Random recombination of antibody single chain Fv sequences after fragmentation with DNase I in the presence of Mn^{2+}.** *Nucleic Acids Res* 1995, **23**:3067-3068.

16. Zhao H, Arnold FH: **Combinatorial protein design: strategies for screening protein libraries.** *Curr Opin Struct Biol* 1997, **7**:480-485.

17. Zhang J, Dawes G, Stemmer WPC: **Evolution of an effective fucosidase from a galactosidase by DNA shuffling and screening.** *Proc Natl Acad Sci USA* 1997, **94**:4504-4509.

18. Moore JC, Jin HM, Kuchner O, Arnold FH: **Strategies for the**
   • ***in vitro* evolution of protein function: enzyme evolution by random recombination of improved sequences.** *J Mol Biol* 1997, **272**:336-347.
   DNA shuffling was used to recombine improved *para*-nitrobenzyl esterases. It is shown how screening requirements increase rapidly as more mutations and more parent sequences are recombined. A simple statistical analysis helps to compare various evolution strategies.

19. Liu DR, Magliery TJ, Pastrnak M, Schultz PG: **Engineering a**
   • **tRNA and aminoacyl-tRNA synthetase for the site-specific incorporation of unnatural amino acids into proteins *in vivo*.** *Proc Natl Acad Sci USA* 1997, **94**:10092-10097.
   Used DNA shuffling and selection to evolve a tRNA synthetase that efficiently acylates an orthogonal tRNA. In addition, a significant decrease in the recognition of the wild type tRNA was observed, although no negative selection against wild type tRNA recognition was employed, which implies that the functions (activities towards the two tRNAs) are inversely coupled.

20. Crameri A, Cwirla S, Stemmer WPC: **Construction and evolution of antibody-phage libraries by DNA shuffling.** *Nat Med* 1996, **2**:100-102.

21. Crameri A, Dawes G, Rodriguez E, Silver S, Stemmer WPC: **Molecular evolution of an arsenate detoxification pathway by DNA shuffling.** *Nat Biotechnol* 1997, **15**:436-438.

22. Crameri AC, Raillard S, Bermudez E, Stemmer WPC: **DNA**
   •• **shuffling of a family of genes from diverse species accelerates directed evolution.** *Nature* 1998, **391**:288-291.
   DNA 'family' shuffling of homologous genes yielded a sequence that confers a dramatic increase in antibiotic resistance. It is proposed that a sparse sampling of sequence space around multiple parents is more powerful than a more exhaustive sample of sequence space around any one parent.

23. Huynen MA, Stadler PF, Fontana W: **Smoothness within ruggedness: the role of neutrality in adaptation.** *Proc Natl Acad Sci USA* 1996, **93**:397-410.

24. Zhao H, Arnold FH: **Functional and non-functional mutations**
   • **distinguished by random recombination of homologous genes.** *Proc Natl Acad Sci USA* 1997, **94**:7997-8000.
   This paper outlines an efficient method for identifying mutations responsible for a difference in the function of two proteins by high-fidelity DNA shuffling.

25. Shao Z, Zhao H, Giver L, Arnold FH: **Random-priming *in vitro***
   • **recombination: an effective tool for directed evolution.** *Nucleic Acids Res* 1998, **26**:681-683.
   Describes a method for *in vitro* recombination in which fragments for reassembly are produced by extension of random-sequence primers.

26. Zhao H, Giver L, Shao Z, Affholter JA, Arnold FH: **Molecular**
   • **evolution by staggered extension process (StEP) *in vitro* recombination.** *Nat Biotechnol* 1998, **16**:258-261.
   Describes a method for *in vitro* recombination based on template switching during DNA synthesis. This technically simple method is used to evolve a thermostable subtilisin.