J. Mol. Biol. (1997) **272**, 336–347

JMB

Strategies for the *in vitro* Evolution of Protein Function: Enzyme Evolution by Random Recombination of Improved Sequences

Jeffrey C. Moore, Hua-Ming Jin, Olga Kuchner and Frances H. Arnold*

Division of Chemistry and Chemical Engineering, Mail Code 210-41, California Institute of Technology Pasadena, CA 91125, USA

**Corresponding author*

Sets of genes improved by directed evolution can be recombined *in vitro* to produce further improvements in protein function. Recombination is particularly useful when improved sequences are available; costs of generating such sequences, however, must be weighed against the costs of further evolution by sequential random mutagenesis. Four genes encoding para-nitrobenzyl (pNB) esterase variants exhibiting enhanced activity were recombined in two cycles of high-fidelity DNA shuffling and screening. Genes encoding enzymes exhibiting further improvements in activity were analyzed in order to elucidate evolutionary processes at the DNA level and begin to provide an experimental basis for choosing *in vitro* evolution strategies and setting key parameters for recombination. DNA sequencing of improved variants from the two rounds of DNA shuffling confirmed important features of the recombination process: rapid fixation and accumulation of beneficial mutations from multiple parent sequences as well as removal of silent and deleterious mutations. The five to sixfold further enhancement of total activity towards the para-nitrophenyl (pNP) ester of loracarbef was obtained through recombination of mutations from several parent sequences as well as new point mutations. Computer simulations of recombination and screening illustrate the trade-offs between recombining fewer parent sequences (in order to reduce screening requirements) and lowering the potential for further evolution. Search strategies which may substantially reduce screening requirements in certain situations are described.

Keywords: directed evolution; DNA shuffling; random mutagenesis; para-nitrobenzyl esterase

Introduction

Enzymes can be evolved *in vitro* to exhibit new and useful functions. A sampling of the local sequence space of the enzyme is created by mutagenesis; screening or selection directs the evolution towards the desired features. A successful strategy for improving enzyme activity in non-natural environments (Chen & Arnold, 1993) and on non-natural substrates (Moore & Arnold, 1996) has been to accumulate amino acid substitutions over multiple generations of random mutagenesis and screening. In practice, the best variant identified in each generation is chosen to parent the subsequent generation. Other potentially useful variants are set aside, and their mutations must be rediscovered in the evolved protein background in order to become incorporated. Because there is no mechanism other than back mutation for deleting mutations, this approach can also accumulate deleterious mutations, leading to premature termination of an evolving lineage. These are the classical arguments for the benefits of recombination (sex) in evolution (Maynard Smith, 1988). Recombination allows more rapid accumulation of beneficial mutations present in a population. It also makes possible the removal of deleterious mutations which would otherwise accumulate in an asexual population, a phenomenon known to geneticists as Müller's ratchet (Müller, 1932). Recombination can

Present addresses: J. C. Moore, Merck & Co., Inc., P.O. Box 2000, RY80Y-110, Rahway, NJ 07065, USA; H.-M. Jin, SmithKline Beecham, Mail Code UE0548, 709 Swedeland Rd., King of Prussia, PA 19406, USA.

Abbreviations used: pNB, para-nitrobenzyl; pNP, para-nitrophenyl.

0022–2836/97/380336–12 $25.00/0/mb971252

provide similar benefits for *in vitro* molecular evolution (Stemmer, 1994a,b).

Bacillus subtilis p-nitrobenzyl (pNB) esterase catalyzes the hydrolysis of the para-nitrobenzyl esters of various cephalosporin-type antibiotics, a necessary step in their large-scale synthesis (Zock *et al.*, 1994). Using four generations of sequential random mutagenesis and screening, we evolved a series of pNB esterases up to 30 times more active towards hydrolysis of the pNB ester of loracarbef (LCN-pNB) in aqueous dimethylformamide (Moore & Arnold, 1996). During the fourth generation, a large number (~7500) of pNB esterase clones were screened and partially characterized in order to validate the rapid screening assay. Sixteen improved pNB esterase clones were identified, from which the five most active enzymes (>50% enhancements in activity over the parent enzyme) were characterized. DNA sequencing revealed four unique pNB esterases (Table 1). Due to the limitations of screening, evolved sequences are generated using a low rate of point mutagenesis and typically accumulate a single beneficial mutation per generation. A simple restriction/ligation experiment demonstrated that recombination of mutations present in at least two of those sequences could further improve pNB esterase activity. Recombining gene segments from two improved pNB esterase variants yielded an enzyme twice as active as the best parent. DNA sequencing demonstrated that mutations from each of the two parents were combined in the new sequence (I60V and L334S), while one neutral or slightly deleterious mutation was deleted (K267R; Moore & Arnold, 1996).

Stemmer recently introduced the technique of "DNA shuffling" to create novel genes by recombination of closely-related DNA sequences (Stemmer, 1994b). Because it also introduces new point mutations during reassembly of the DNA fragments, DNA shuffling alone has been effective for directed protein evolution starting from a single sequence (Stemmer, 1994a; Crameri *et al.*, 1996). Questions arise as to how this approach is best implemented and integrated with other *in vitro* evolution approaches such as sequential random mutagenesis. Issues include optimizing the point mutagenesis rate associated with DNA shuffling, determining appropriate screening sample sizes and how many parental genes to recombine, and deciding when to use recombination. Here we investigate the further evolution of pNB esterase by DNA shuffling of the improved sequences generated by random mutagenesis and screening. By following how the genes evolve during cycles of DNA shuffling and screening, we can elucidate the mechanisms contributing to the evolution of function and begin to optimize strategies for *in vitro* evolution. An analysis of the recombination process identifies some of its benefits and limitations for directed evolution and allows a rational choice of mutagenesis and screening strategies.

Results and Discussion

Recombination statistics and screening requirements

To comment on the utility of DNA shuffling in directed evolution, a review of the statistics of recombination of multiple parent sequences is useful. For this discussion, we will assume that the mutations are unique and distributed far enough from one another on the genes that recombination occurs freely between any two. Furthermore, equal amounts of the initial DNA sequences are recombined. Consider the random recombination of three parent sequences, each of which contains a single mutation. Any given mutation will be incorporated into a progeny sequence with a probability of 1/3; the probability of generating the wild-type sequence is 2/3 at each mutation site. This highlights an important consequence of shuffling multiple sequences: there is a statistical preference for the absence of mutation in the progeny. The overall probability of picking a completely wild-type sequence from the recombined library is $(2/3)^3 = 0.30$. The probability of generating a sequence containing a single mutation (a parent sequence) is $1/3 \times (2/3)^2 = 0.15$. Because there are $C_1^3 = 3!/1!2!$, or three such sequences, the overall fraction of parent sequences in the library is 0.45. Thus fully 75% of the sequences in the recombined library are variants already in the evolutionist's possession.

In general, for a recombination system consisting of N sequences and M total mutations, the probability of generating progeny sequences containing μ mutations equals the number of ways a μ-mutation sequence can be generated (C_μ^M) multiplied by the probability of generating any single μ-mutation sequence:

$$P_\mu = C_\mu^M \left(\frac{1}{N}\right)^\mu \left(\frac{N-1}{N}\right)^{M-\mu}$$

$$= \frac{M!}{(M-\mu)!\mu!} \left(\frac{1}{N}\right)^\mu \left(\frac{N-1}{N}\right)^{M-\mu}$$

Figure 1 summarizes the analysis for recombination of single-mutation parent sequences ($N = M$). The probability that recombination will return the zero-mutation "grandparent" or single-mutation parent sequences remains constant between 73 and 75%; only ~25% of the clones screened have sequences that have not already been examined. The probability of creating individual sequences declines dramatically with increasing numbers of parents. The least frequent sequences are those containing the majority of mutations from the parent population, and the sequence containing all the mutations ($\mu = M$) is of course the rarest. The probability P_M of generating the rarest sequence is $1/N^M$.

Because we are interested in the evolution of function, we need consider only those mutations responsible for functional differences among pro-

Table 1. DNA and amino acid substitutions in fourth, fifth and sixth generation evolved pNB esterases

Mutation	Amino acid substitution	4-54B9	4-38B9	4-53D5	4-43E7	5-6C8	5-5E4	5-4H4	5-4G2	5-4D12	5-2D3	6-10F1	6-1D12	6-1C7	6-1A6
ATC 27 → ATA									YES	yes					
CCT 33 → CCC		yes													
ATT 60 → GTT	I60V		**yes**			**yes**	**yes**		**yes**	**yes**	**yes**	**yes**	**yes**	**yes**	**yes**
GAT 81 → GAC															yes
TAT 84 → TAC				yes		yes									
AGT 94 → GGT	S94G				**yes**										
GGA 127 → GGG															yes
TCG 148 → TCC						yes									
TTT 149 → TTC									yes	yes					
GGC 227 → GGG	A227G										**yes**				
ATT 239 → ATC				yes			yes								
AGA 246 → AGG															yes
GCG 252 → GCT							yes								
AAA 267 → AGA	K267R	**yes**													
CCG 317 → TCG	P317S					**yes**						**yes**	**yes**	**yes**	**yes**
TTA 334 → GTA	L334V				**yes**										
TTA 334 → TCA	L334S	**yes**				**yes**	**yes**	**yes**	**yes**	**yes**	**yes**	**yes**	**yes**	**yes**	**yes**
GCT 343 → GTT	A343V			**yes**				**yes**							**yes**
CAT 356 → CGT	H356R								**yes**	**yes**		**yes**		**yes**	
ACT 359 → GCT	T359A												**yes**		
ATT 464 → GTT	I464V								**yes**	**yes**		**yes**		**yes**	

DNA substitutions are identified in the context of the three-base codon of the encoded enzyme sequence. Grey background indicates mutations from fourth generation parent sequences. White background indicates new mutations which arose during DNA shuffling. Bold face type indicates translated DNA mutations.

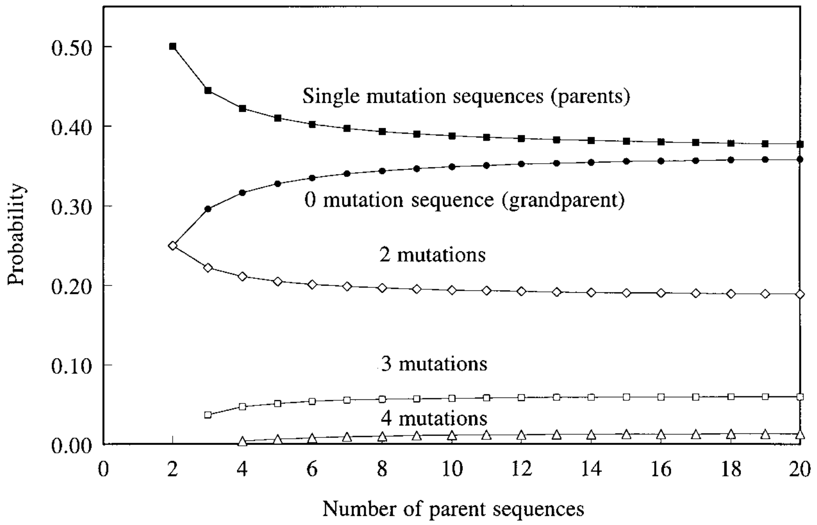

Figure 1. Probabilities of generating sequences containing different numbers of mutations by random recombination, based on recombining single-mutation parent sequences. Novel variants (not grandparent or parent sequences) are shown with unfilled symbols.

tein variants. Neutral mutations by definition do not affect function; their distribution among progeny sequences is determined statistically, even in the screened population (Zhao & Arnold, 1997b). Thus for the purposes of this discussion of recombination libraries and screening requirements, *M* is the number of mutations that affect the targeted function (either beneficial or deleterious).† By screening enough clones to ensure that the rarest sequence, that is, containing all *M* mutations, has been examined, one can be sure that the best variant will be discovered. This is true even if the best variant does not contain all the functional mutations (as would be expected if some mutations were deleterious or if the effects of mutations are not cumulative).

In practice, of course, oversampling is required to ensure that a particular variant has been examined during the course of screening. To be 95% confident that the most active combination variant has been examined, we must be 95% confident the rarest variant has been examined. If *S* is the number of clones sampled, then

$$(1 - P_{\mathrm{M}})^{S} < 1 - \text{confidence limit}$$

describes how the probability of not sampling the rarest variant changes with increasing *S*. This allows calculation of the number of samples required for a given confidence limit. The oversampling is then how many more samples must be screened over the theoretical minimum. When one clone is required with 95% confidence, the oversampling will be between 2.6 and 3.0 (for larger numbers of parents). Even a relatively low rate of background point mutagenesis, however, can introduce significant confounding effects. Non-neutral point mutations obscure recombination events and increase the amount of screening required to find the best sequences (*vide infra*). Thus, in practice, it may be impossible to screen sufficient numbers of clones to be sure of finding the best recombinant, particularly when the point mutation rate is high and a large number of functional mutations are being recombined. Alternative strategies which can reduce screening requirements under special conditions will be discussed further on.

DNA shuffling of evolved pNB esterases

An effect of forcing DNA polymerase to synthesize full length genes from the pool of small DNA fragments generated during DNA shuffling is additional background point mutagenesis. A high rate of point mutagenesis can severely inhibit the discovery of novel combinations of existing mutations within a population. Because most mutations are deleterious (in a screening assay sensitive to small changes in the screening variable), beneficial recombinations and rare beneficial point mutations are masked by the negative background. DNA shuffling with a 0.7% mutagenesis rate, for example, would yield an average of 10-11 point mutations in the 1470 bp pNB esterase gene. This is substantially more than the optimal mutation frequency (~three mutations per gene) for directed evolution of pNB esterase (Moore & Arnold, 1996). In fact, when the four evolved pNB esterase gene sequences were shuffled using *Taq* polymerase, fully 90% of the clones in the resulting library exhibited essentially no esterase activity during screening (data not shown). In a parallel study, we observed that 80% of the clones generated by DNA shuffling of subtilisin E exhibited no activity (Zhao & Arnold, 1997a).

In an effort to reduce the background mutagenesis rate, a proofreading polymerase (Pwo) was used during fragment reassembly. With Pwo, 50 to 100 base-pair fragments could be reassembled to create a library in which fully 80% of the clones

† A mutation that is neutral in one context (i.e. in the wild-type background), but becomes functional in a different context, would be considered a functional mutation.

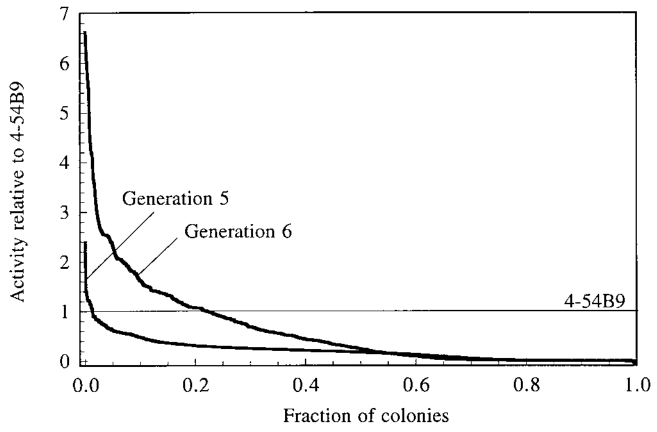

Figure 2. Activity profiles of generations 5 and 6 determined by screening libraries created by DNA shuffling of unique fourth and fifth generation variants. Activities were sorted from best to worst. Profiles are normalized by the number of clones screened.

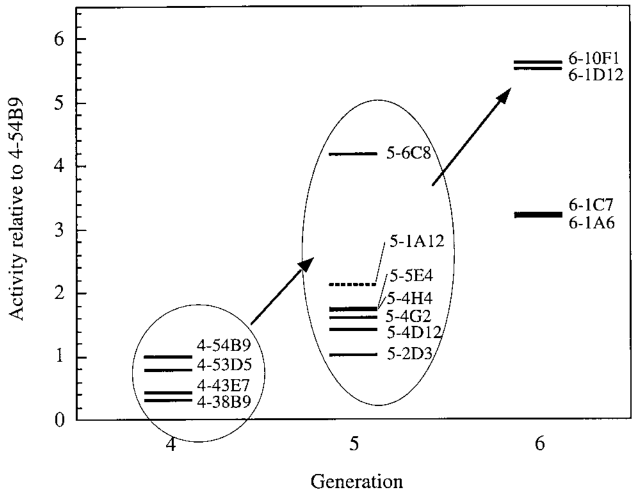

Figure 3. Activities of fourth, fifth and sixth generation pNB esterase variants (Table 1) in screening assay. Fourth generation variants were recombined and screened to identify improved enzymes in generations 5 and 6.

retained activity. Inserts from 13 randomly picked colonies were partially sequenced in order to determine the point mutation rate. Five mutations not present in any of the parent sequences were found in 12,000 nucleotides sequenced, for an overall mutagenic rate of ~0.04%. These minimally mutagenic conditions were used for DNA shuffling. A subsequent, in-depth investigation of the various steps involved in DNA shuffling has allowed us to identify a set of recombination protocols with a wide range of point mutagenesis rates (Zhao & Arnold, 1997a).

Four unique fourth generation improved pNB esterase variants were chosen as the starting point for further directed evolution by DNA shuffling. Two cycles of DNA shuffling and screening for activity towards the *p*-nitrophenyl ester of loracarbef (pNP-LCN) in 25% dimethylformamide (DMF) were performed. The activity profiles of the resulting populations (generations 5 and 6) are shown in Figure 2. To generate these profiles, activities of the individual clones measured in the 96-well plate screening assay were normalized by cell density (A_{600}) and plotted in descending order. Approximately 2% of the 948 generation 5 clones screened exhibit more total activity than the most active parent (4-54B9). The screened population was sufficiently large to give a high level of confidence that the most active variant that can be generated by simple recombination of the fourth generation sequences has been found.† The six most active variants from generation 5 were collected and shuffled again to create generation 6. Fully 20% of the 474 clones screened were more active than 4-54B9. Only 20 to 25% of the clones were inactive, as expected using the high fidelity Pwo-only shuffling conditions.

Figure 3 summarizes the activities of the four fourth generation parents and the best variants identified in generations 5 and 6. The improvement in enzyme activity as a result of shuffling is already apparent in the fifth generation, which includes one variant (5-6C8) fourfold more active than 4-54B9 and twice as active as variant 5-1A12 previously generated by ligation recombination (Moore & Arnold, 1996). The sixth generation contains two clones with yet higher activities than 5-6C8. The best one, 6-10F1, represents a five to sixfold improvement over 4-54B9 and is ~150 times more active than the wild-type.

Activities of the fifth and sixth generation variants towards the *p*-nitrobenzyl ester of loracarbef (LCN-pNB) were also determined, using a modified HPLC assay as described in Materials and Methods. The best pNB esterase is 5-6C8, which exhibits a threefold increase in total activity over 4-54B9. This clone is now ~100 times more active than wild-type pNB esterase towards LCN-pNB in 25% DMF. The sixth generation variants exhibited no further improvement in activity towards this substrate, a clear reflection of the use of the pNP ester during screening and the first law of random mutagenesis: "You get what you screen for" (You & Arnold, 1996).

† When shuffling four parent sequences each of which contains one beneficial mutation, 765 clones must be screened to be 95% confident that all combinations have been examined (assuming recombination occurs freely between mutations and no point mutagenesis). A 0.04% rate of point mutagenesis translates to less than 0.6 new mutations per sequence, of which only a fraction will affect function (estimated from the activity profile of a library created by error-prone PCR to be ~0.5, data not shown).

Analysis of evolved pNB esterase genes

DNA mutations present in the four parent fourth generation sequences and mutations identified by sequencing the genes encoding the selected fifth and sixth generation variants are summarized in Table 1. By comparing the activities and sequences of these variants with the third-generation parent, four beneficial mutations were identified (leading to amino acid substitutions I60V, L334V, L334S and A343V). The remaining mutations present in the fourth generation sequences are neutral or mildly deleterious (Moore & Arnold, 1996).

Several interesting observations can be made from this Table. It can be seen that a number of mutations increase their frequencies in the subsequent generations. Substitutions I60V in 4-38B9 and L334S in 4-54B9 are each present in a single fourth generation parent. In contrast, I60V is present in five of the six fifth-generation variants, and L334S is present in all six. By the sixth generation both substitutions are fixed in the population. A new substitution at position 317, first found during the fifth generation (5-6C8), also becomes fixed by the sixth. This new mutation probably accounts for the significant increase in activity of variant 5-6C8. The P317S substitution is positioned near the enzyme surface in a loop located on the same side of the entrance to the substrate binding pocket as amino acid substitutions L334S, M358V and A343V (Moore & Arnold, 1996). Removal of a proline at this position may relax conformational constraints on the loop, allowing the substrate freer access to the active site.

The two separate beneficial mutations at position 334 in 4-43E7 and 4-54B9 are mutually exclusive, and a competition exists as to which one will be propagated to successive generations. Variant 4-54B9 has more than twice the activity of 4-43E7 as a result of the mutation at position 334, and the fifth generation recombination progeny in fact show the L334S substitution from 4-54B9 exclusively. Recombination provides a rapid means to identify the most effective mutation among multiple possibilities at any given site.

Related to the observation that beneficial mutation combinations are fixed is the fact that recombination and screening also effectively remove neutral and deleterious mutations. Three of the five mutations present in the fourth generation parents that are synonymous (DNA mutations in codons 33, 84, and 239 that do not lead to amino acid substitutions) or non-synonymous, but believed neutral or mildly deleterious in their effects on total activity (mutations leading to amino acid substitutions S94G and K267R (Moore & Arnold, 1996)), have been removed from the improved pNB esterase population in a single round of shuffling; all five are removed by the sixth generation. The two most active sixth generation enzyme variants, 6-10F1 and 6-1D12, have no synonymous mutations at all and only one mutation (at position 359) not seen in any previous clone. Due to the statistical preference for the absence of mutations the recombination process is highly effective in filtering out neutral (and deleterious) mutations starting from multiple parent sequences.

Table 1 also shows that the DNA shuffling technique can recombine multiple parent sequences to create novel progeny. Recombination between at least three fourth-generation parents is required to create 5-5E4, and at least three fifth-generation parents were recombined to generate clones 6-10F1 and 6-1A6 (based on the presence and absence of the DNA mutations in the sequences compared to the parent sequences).

Finally, it is useful to note that DNA shuffling generates point mutations that are rarely observed during PCR (at least for the low-mutagenesis rate PCR conditions used for directed evolution of longer DNA sequences). Four of the 12 new point mutations identified in the fifth and sixth generation variants, for example, are G → C (and C → G) and G → T (and C → A) transversions, which were not found at all during the first four generations of pNB esterase evolution involving PCR mutagenesis (Moore & Arnold, 1996). These mutations were also generated very rarely during the error-prone PCR mutagenesis of subtilisin (Shafikhani *et al*., 1997). DNA shuffling and error-prone PCR together may provide access to a wider range of amino acid substitutions.

Evolved pNB esterase amino acid sequences

Amino acid substitutions in the evolved pNB esterases are indicated in Table 1; changes in amino acid sequence along the lineage are summarized in Figure 4. The accumulation and fixation of two beneficial amino acid substitutions from the fourth generation, I60V and L334S, is essentially complete in a single generation of DNA shuffling and screening 948 clones. In contrast, A343V, a beneficial mutation found in the fourth generation, no longer appears in the majority of fifth or sixth generation variants. The (5-4H4) recombinant of the parent containing this mutation (4-53D5) with 4-54B9 shows no improvement in activity over 4-54B9 (Figure 3). Substitutions A343V and L334S therefore do not work in concert to improve enzyme activity, and consequently there is little or no driving force to retain A343V in the population. The remaining fifth generation variants, with the exception of 5-6C8, are less active than 5-1A12 (Figure 3), yet they contain the I60V and L334S substitutions while omitting K267R, as does 5-1A12. This suggests that the additional mutations found in those sequences are neutral, or possibly, deleterious. For instance, the amino acid sequences of 5-5E4 and 5-1A12 are identical, and the decreased activity of the former is likely due to the two synonymous mutations in 5-5E4 not present in 5-1A12. Because the screen evaluates the total activity of a clone (normalized by cell density), synonymous mutations can influence the result, for

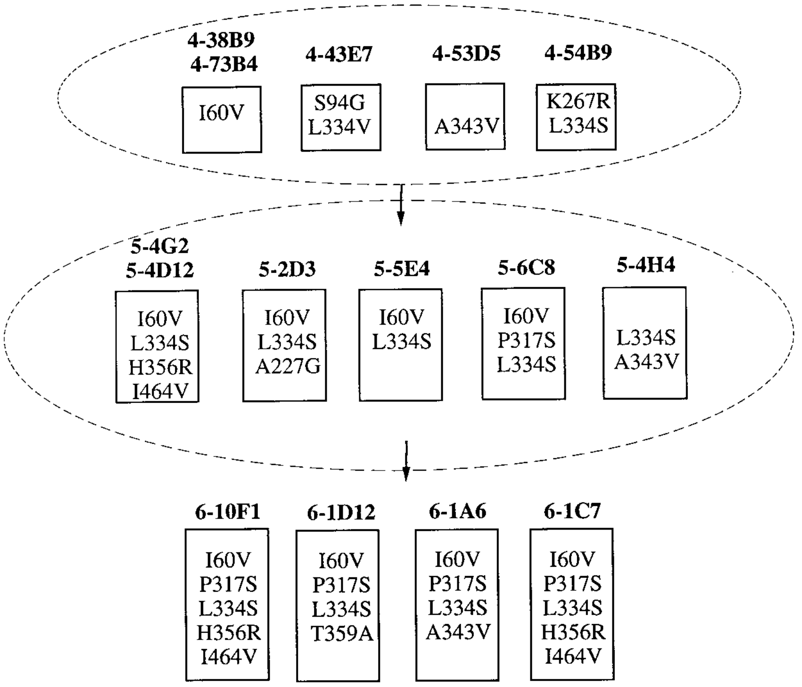

Figure 4. Lineage of pNB esterase variants showing amino acid substitutions accumulated by four generations of sequential random mutagenesis (fourth generation) and by DNA shuffling (fifth and sixth generations) and screening. All variants contain amino acid substitutions H322R, Y370F, M358V and L144M from the third generation parent (Moore & Arnold, 1996).

example, by affecting the amount of active enzyme expressed. The new beneficial mutation that gives rise to the P317S substitution becomes fixed in the sixth generation, and further evolution during that generation primarily arises from point mutation rather than recombination.

Clones 5-4G2 and 5-4D12, whose DNA sequences are identical, both contain amino acid substitutions H356R and I464V. These two substitutions are seen together again in 6-10F1 and 6-1C7. Because 6-10F1 and 6-1D12 have almost identical activity, we can reasonably infer that the I60V, P317S, and L334S substitutions are responsible for that activity, while the mutations leading to H356R and I464V from the fifth generation as well as a new mutation, T359A, in 6-1D12 are neutral. The three mutations believed responsible for enhanced activity are also present in 6-1A6, along with the last mutation in this system known to enhance activity, A343V. That 6-1A6 has lower activity than 6-10F1 and 6-1D12 is therefore attributable to either the three synonymous mutations in 6-1A6 (Table 1) or antagonism between amino acid substitutions A343V and P317S or I60V.

The new point mutations that arose during the minimally mutagenic DNA shuffling increased (P317S) and decreased enzyme activity. The effects of individual mutations can be ascertained with confidence because the sequences differ from one another at very few positions. We have recently demonstrated a method that allows one to distinguish clearly beneficial, neutral and deleterious mutations in evolved sequences by random recombination with ancestor sequences (Zhao & Arnold, 1997b). This method will be particularly useful for identifying mutations responsible for functional changes in proteins in a background of neutral mutations (as happens when multiple new mutations are present).

Only 2% of the fifth generation clones are more active than the most active parent, 4-54B9 (Figure 2). Although 25% of the progeny should be novel, the combination I60V + L334S predominates in the most active variants (Figure 4), suggesting that many of the remaining combinations lead to lower activity than in 4-54B9. Additionally, while there is no mechanism for recombination alone to generate inactive clones, ~25% of the variants in Figure 2 are inactive, presumably as a result of background point mutation. This implies that the frequency of enhanced-activity recombinants is reduced by point mutation and emphasizes the importance of minimizing the mutagenesis rate when recombining positive mutations.

Developing strategies for directed evolution

Recombination versus *random mutagenesis*

Recombination is only useful if a population of sequences is available from which new combinations of mutations can be generated. Homologous proteins with similar sequences could provide such a starting population (Stemmer, 1994b). (Note, however, that a high level of sequence identity may be required for DNA shuffling.) Populations of sequences can also be created by the background point mutagenesis feature of DNA shuffling (Crameri *et al.*, 1996). Alternatively, they can be generated by random mutagenesis and screening experiments, as they have been for the current study. When interesting sequences already exist, recombination offers an efficient means to use that information. If the sequences must be generated, however, then one should consider that cost in the overall cost of evolution by recombination as compared to, for example, evolution by sequential generations of random mutagenesis and screening.

In theory, the sequential (or ''asexual'') approach requiring the least labor in terms of screening is to screen randomly mutagenized clones until a positive is identified and then use that as the template for the next generation. The process is a random walk in which the first uphill step encountered is taken. To take a simple illustration, consider three mutations A, B and C that each contribute in a cumulative, if not additive, manner when combined. A, B and C could be collected in the ABC variant in three sequential generations of mutagenesis and screening. Alternatively, if A, B and C all contribute to the desired feature in the wild-type background (as they often do; see, for example, Chen & Arnold, 1993), they could be found separately and then recombined to make ABC. Finding the single-mutation sequences A, B, and C, however, requires screening the same number of colonies as finding ABC by sequential evolution. Recombining the A, B, and C sequences to make ABC requires additional screening. Of course, the sequential pathway requires three random mutagenesis steps, while the recombination pathway requires only one mutagenesis step and one DNA shuffling step. The advantages of one approach over the other then depend on the costs of screening relative to the DNA manipulations.

Note that the severe limitations screening places on the number of colonies that can be sampled makes it difficult to accept downhill steps in the hope that further improvements can be found further out in sequence space (Moore & Arnold, 1997). It also means that extremely rare events such as the recombination of neutral or slightly deleterious mutations to make a beneficial combination will probably not contribute in any significant fashion to the evolutionary process.

The pNB esterase evolution provides a concrete example for analysis. Approximately one in every 1500 to 2000 randomly mutagenized pNB esterase clones screened was positive (showing 50% or greater enhancement in activity over the parent; Moore & Arnold, 1996). To generate the population of four unique positives for DNA shuffling, we examined a total of 7500 clones. Finding the best combination variant required additional DNA shuffling experiments, and ~1400 additional colonies were screened. Thus a total of 9000 clones were screened in going from generations 3 to 6. There is no guarantee that the sequences chosen for recombination are unique: in fact, the original fourth generation clones contained five variants, two of which were identical (4-38B9 and 4-54B9) and two of which contained mutations in the same codon (4-43E7 and 4-54B9), precluding recombination between these variant pairs. It is very likely that variants of comparable or even greater activity could also have been created by continuing random mutagenesis and screening for three generations from the first fourth generation variant identified. The total screening requirement would be the same.

In practice, however, the uphill climb often involves identification of multiple positives during each generation. Everything but the one chosen to parent the next generation is discarded in the random uphill walk of the ''asexual'' evolution. During the pNB esterase evolution, we often identified four or five potential positives during the rapid screen on the LCN-pNP colorimetric substrate. Those were either verified or not during a second level screen on the *p*-nitrobenzyl (LCN-pNB) substrate, and it was often the case that more than one sequence was a true positive (Moore & Arnold, 1996). The other improved sequences could of course be collected and recombined at any time and at relatively little screening cost. A significant advantage of the DNA shuffling method is its ability to utilize these available positive sequences.

Computer simulations of random recombination and screening

The statistical model can be used to optimize the number of parent sequences chosen for DNA shuffling. Screening during the fourth generation actually resulted in the identification of 16 clones measurably more active than the parent, of which five were at least 50% more active (Moore & Arnold, 1996). An attempt to recombine all 16 sequences yielded no clones more active than 4-54B9 (~1000 clones screened). This result can be understood when we consider the dramatically lower probability of finding the best combination(s) as the number of sequences increases. If the screening sample size is limited to a few thousand clones, there is little chance that the best sequences, or even sequences better than the best parent, will be found by screening a library created from 16 parents.

We have used a computer simulation of the random sampling of the two recombined libraries obtained by shuffling five and ten sequences to illustrate the advantage of choosing fewer parents when screening is limited. Recombining all ten parents becomes advantageous, however, when large numbers of clones can be examined. (Of course, the larger sampling requirement should then be compared to the potential for continued evolution by random mutagenesis.) Assuming that the ten parent sequences each contain a unique, single beneficial mutation ($N = M$) and that they can be recombined to give all possible combinations, we calculated P_μ for $\mu = 0$ through 10. Since $\Sigma P_\mu = 1$, these were organized into a cumulative distribution from 0 to 1, and a random number generator was used to pick a point on the cumulative distribution, thereby identifying μ (number of mutations per sequence). A second random number generator was used to pick one of the C_μ^M possible sequences containing μ substitutions using an evenly spaced distribution of possible combinations. The activity of the sequence chosen was then calculated by assuming that the free energies of activation of the variants (proportional to the natural logarithms of their activities) are additive.

The results of this simulation are shown in Figure 5, using the activity data from the fourth generation pNB esterase variants. Figure 5(a) shows the averages of the highest values of mutant activities obtained over 15 separate trials for each (screening) sample size. The results obtained by shuffling the ten best mutants (black diamonds) can be seen to be slightly worse than those obtained by shuffling the five best mutants (white squares), for sample sizes up to about 10,000 to 15,000. That is, the average expected best mutant is higher for shuffling five parents at a time for small sample sizes. Figure 5(b) and (c) show the range of values of the highest mutant activity obtained on each of 15 separate trials for each sample size. Here, the highest values obtained from recombining the best ten variants (black diamonds) become better than the values obtained from shuffling the best five (white squares) at sample sizes greater than about 1000. Although shuffling the top ten mutants for this set of data can yield higher final activities, the simulation shows that the outcome is much more risky when screening capabilities are limited to a few thousand clones.

Simulations also show that the results of the comparison of shuffling five *versus* ten parents is highly sensitive to the values of the activities. For instance, if the activities of mutants 6 through 10 are decreased, then the sample size at which recombining all ten mutants becomes preferable becomes much higher. Moreover, the simulation can be adapted for cases in which some or all of the parent sequences have two or more mutations, which may or may not be recombinable. Thus this simulation approach can be used to determine the optimal number of sequences to recombine for any given set of activity values and any given sample size.

The simple additivity assumption on which these simulations are based† is a reasonable first approximation of the behavior of combined mutations in proteins (Wells, 1990) and is useful for a first exploration of strategic issues in *in vitro* protein evolution. The real behavior is often more complex and will depend on the property of interest as well as the particular protein. However, it is likely that deviations from simple additivity are governed by non-linear functions of the number and magnitude of changes; values will certainly depend on which subset of mutations is recombined. While it is possible to modify the simulation to take into account deviations from additivity, very little data are available on the effects of large numbers of mutations. We have therefore not

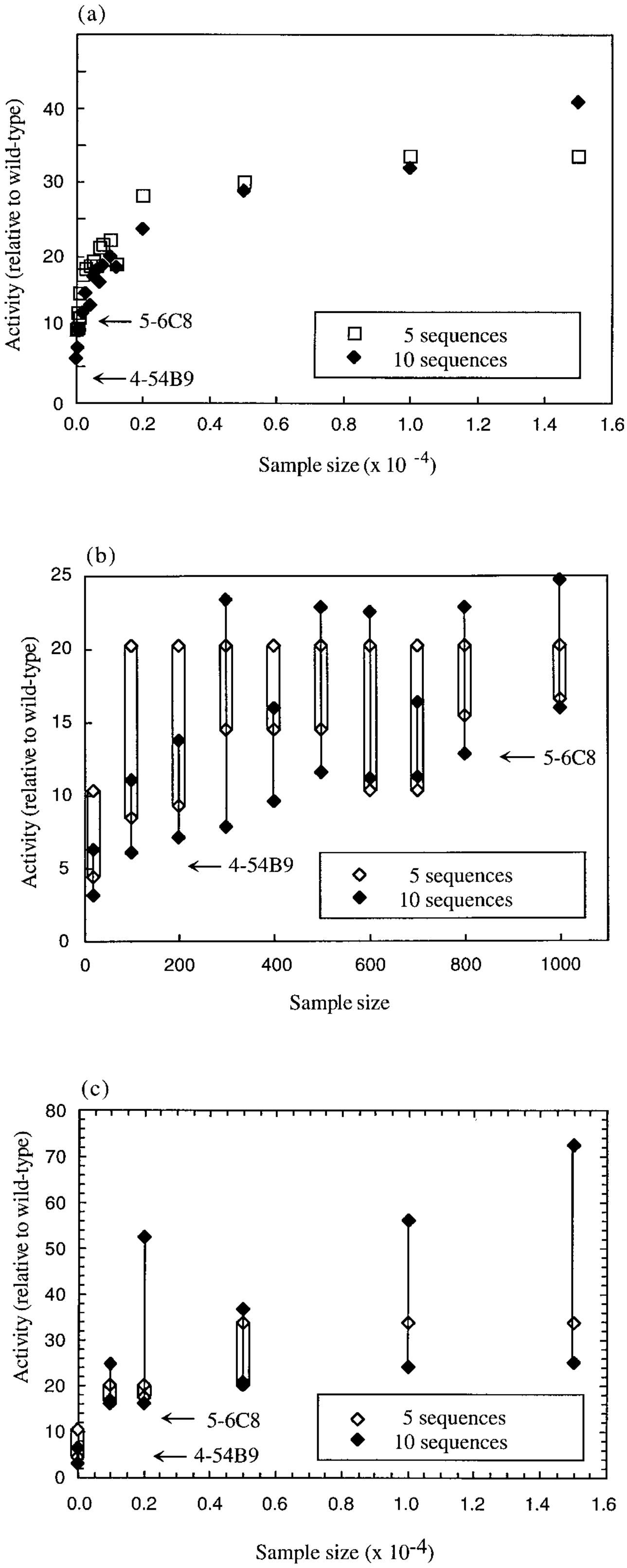

Figure 5. (a) Averages of highest values of mutant activities obtained over 15 separate trials of simulated random recombination of five and ten parent sequences. (b) and (c) Range of values of mutant activities obtained over 15 separate trials. Activities of best fourth-generation parent (4-54B9) and highest-activity fifth generation clone identified (5-6C8) are indicated for comparison.

† Both beneficial and deleterious mutations can be accommodated in this framework.

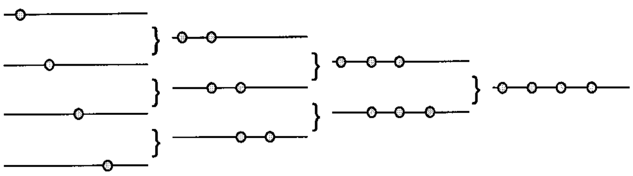

Figure 6. Pairwise recombination can reduce screening requirements, provided effects of mutations are cumulative. By shuffling two sequences at a time, sequences containing two mutations represent 25% of the recombined library. This example involves six recombination experiments.

attempted to include deviations from additivity in the current simulations. Figures 5(a), (b), and (c) show the activities of the best fourth generation parent (4-54B9) and the best fifth generation clone identified (5-6C8) by screening the shuffled library. That the activity of 5-6C8 is ~twofold less than the average expected for screening 948 clones reflects the fact that (i) only four of the original five positive clones identified during generation 4 were unique, (ii) two mutations were on the same codon and could not be recombined, and (iii) the mutations combine with significantly less than 100% additivity.

Alternative search strategies

Finally, we will briefly consider two other search strategies that might be used to minimize screening requirements. One approach to producing a multiple-mutation variant which requires the screening of far less clones is multiple-step pairwise recombination. This strategy is illustrated in Figure 6 for the simple case of recombining four (beneficial) mutations from four separate parents. Pairs of parents are mated. As each progeny is a double mutant 25% of the time, only 12 clones are required to find all the double mutants, assuming the effects of the mutations are cumulative. The double mutants are then similarly mated, and screening only eight clones will identify the triple mutants. Mating and screening four clones will generate the quadruple mutant. Thus a total of only 62 clones (24 × 2.6 times oversampling to be 95% confident at each step) must be screened, as compared to the 765 required to generate the quadruple mutant in a single recombination step. Such an approach requires considerable DNA manipulation and would be most useful when screening is extremely difficult. (An attractive alternative at this point may be sequencing the parents and recombination by site-directed mutagenesis.) A further cost of this approach is that the search space is very limited. The assumption is that each activity-enhancing mutation will contribute to the overall activity, so that the quadruple mutant is the best performer of this population. If a particular double or triple mutant is the best performer, it may or may not be found, since not all of these intermediate mutants will have been examined.

A compromise method that works well, at least in theory, can be described as ''population recombination.'' The idea is to shuffle all four parent sequences at once and screen enough clones to see all the double mutants. Because each double mutant occurs 3.5% of the time, 28 clones must be screened. This examines all of the pair-wise interactions between mutations and eliminates those which are not cumulative. The double mutant population is recombined to produce all of the triple mutants and the quadruple mutant (requires screening 16 clones). If the mutations were at least cumulative in their effects, screening 132 (44 × 3.0 times oversampling) clones would search the space completely for the best (quadruple) mutant. This approach most closely describes how recombination/selection experiments operate (Stemmer, 1994a) where all of the clones that survive a particular selection criterion are recombined (often 100 clones or more serving as the parent population for the next generation).

Conclusions

Recombination is an important tool for directing the evolution of proteins. Beneficial mutations can be recombined, while neutral and deleterious mutations are eliminated. The need to screen rather than select for many important enzyme functions, however, severely limits the ability to search for useful combinations. It is therefore imperative to analyze various recombination strategies. Mutagenic rates associated with the recombination process must be low so that beneficial mutations are not lost in a background of deleterious ones. Although a new beneficial amino acid substitution was found as a result of the DNA shuffling of pNB esterase, DNA shuffling may be less efficient for discovery of new mutations compared to a controlled mutagenesis technique (a beneficial mutation can be masked in the background of recombined sequences). Utilizing more than two parents for recombination introduces a statistical preference for not incorporating mutations in progeny, and this has several consequences especially with respect to screening. Recombination favors the dilution of progeny containing the most mutations, which has the effect of exponentially increasing the number of progeny that must be screened in order to find the rarest ones. Because shuffling large numbers of parent sequences can yield many possible combinations, it may also be necessary to strictly limit the number of parent sequences in any given recombination experiment. We have described two alternative search strategies which reduce the required number of variants examined, at the cost of possibly missing intermediate beneficial combinations.

Finally, recombination requires a population of positive variants for efficient enzyme improvement. If a population of positive variants must first be generated, sequential random mutagenesis may require less effort to produce sequences containing multiple mutations. Multiple positive variants are often generated, however, during a single cycle of random mutagenesis and screening. Recombination of these positives can provide substantial improvements at relatively little cost.

Materials and Methods

DNA shuffling

DNA shuffling was performed as described by Stemmer (1994b) with modifications. The 2 kb DNA fragment encoding the *B. subtilis* pNB esterase gene was amplified using PCR (forward primer 5′-CAATCTAGAGGGTATTAATAATG-3′ and reverse primer 5′-CGCGGGATCCCCGGGTACCGGGC-3′). The amplified DNA was purified by gel electrophoresis and extraction using Qiaex kit (Qiagen, Chatsworth, CA). A total quantity of ~10 μg DNA, either from a single parent (non-recombinatorial) or from a mixture of multiple parent sequences (recombinatorial), was digested with DNase I (0.0015 units/μl) at room temperature for 20 minutes in a 100 μl reaction. After ethanol precipitation, the digested DNA was electrophoresed as a smear in a 3% low melting temperature gel of NuSieve GTG Agarose (FMC Bio Products, Rockland, ME). DNA fragments in specified molecular size ranges were collected onto DE-81 filter paper disks (Whatman, Maidstone, England) by electrophoresis and eluted from the filter paper with 400 μl of 10 mM Tris/1 mM EDTA buffer (pH 8.0) containing 1 M NaCl. The DNA fragments were ethanol precipitated and redissolved to approximately 20 ng DNA/μl in 1 × Pwo DNA polymerase buffer (Boerhinger Mannheim, Indianapolis, IN) containing 2 mM $MgSO_4$ and 0.2 mM each of the four dNTPs. A 5 unit/μl Pwo DNA polymerase solution (Boehringer Mannheim) was diluted tenfold, and 0.5 μl was added to a 5 μl redissolved DNA reaction solution. Reassembly of DNA fragments was conducted by PCR, using the conditions 94°C for 40 seconds., then 70 cycles of 94°C for 30 seconds, 50°C for 30 seconds, 72°C for 30 seconds, followed by a final extension step at 72°C for five minutes. A second 0.5 μl of Pwo polymerase was added at the annealing step of the 35th cycle. The reassembled DNA fragments were amplified in a conventional PCR (25 cycles) with the dilution of 1 μl reassembled DNA fragments in a 100 μl reaction. Once the success of the reassembly/amplification reactions was verified by gel electrophoresis, the reassembled product was purified with a Wizard PCR prep kit (Promega Corp., Madison, WI), digested with *Bam*HI and *Xba*I, concentrated by ethanol precipitation, and electrophoresed in an agarose gel. The 1.8 kb product was cut from the gel and the DNA extracted using a Qiaex kit. The final products were ligated with the vector generated by *Bam*HI-*Xba*I digestion of pNB106R (Zock *et al.*, 1994). This library was used to transform competent *E. coli* TG1 cells, as described (Moore & Arnold, 1996).

Screening a pNB esterase library

Screening was based on the assay described previously (Moore & Arnold, 1996), using the *p*-nitrophenyl ester of the loracarbef nucleus (LCN-pNP) as substrate. *E. coli* TG1 containing the plasmid library were grown on LB/tetracycline (20 μg/ml) plates. After 36 hours at 30°C single colonies were picked into 96-well plates containing 100 μl LB/tetracycline medium per well. These plates were shaken and incubated at 30°C for 12 hours to let the cells grow to saturation. Aliquots (20 μl) of the cultures were inoculated into a fresh plate containing 100 μl media per well; these were incubated at 40°C for ten hours with shaking to induce the expression of pNB esterase. Esterase activities were then measured by transferring 20 μl aliquots of the cell cultures into a fresh set of plates where they were mixed with 200 μl of 0.1 M Tris-HCl (pH 7.0) 25% DMF and 2 mM LCN-pNP. Reaction velocities were measured at 450 nm over ten minutes. (11 data points) in a ThermoMax microplate reader (Molecular Devices, Sunnyvale CA). Activities were normalized by the cell densities of the original wells measured at 600 nm to control for variations in cell quantities.

For each round of screening, the clones that showed the highest activities were re-streaked on LB/tetracycline agar plates, and single colonies derived from these plates (three to four colonies from each clone) were inoculated simultaneously into 96-well plates and tube cultures. The former were used to repeat the activity assay, as described above, and the latter were used for glycerol stock and plasmid preparation (Qiawell kit, Qiagen).

Assay of pNB esterase activity on LCN-pNB

A modified HPLC assay was used to determine enzyme activity towards the LCN-pNB (*p*-nitrobenzyl ester) substrate (Chen *et al.*, 1995). The bacterial cells were incubated at 30°C with shaking for 12 hours and then at 40°C for ten hours to induce expression of pNB esterase. Aliquots of cells (200 μl) were incubated with 300 μl reaction buffer for 30 minutes at room temperature. The final reaction mixtures contained 0.1 M Tris-HCl (pH 7.0) 25% DMF and 2 mM LCN-pNB. The reactions were stopped by addition of 500 μl acetonitrile and passed through a nylon syringe filter (Micron Separations, Inc., Westboro, MA) with a pore size of 0.45 μm. Aliquots of each sample (50 μl) were analyzed by HPLC on a 250 mm × 4.6 mm C18 reverse-phase column (Vydac, Hesperia, CA) at room temperature using a linear gradient starting with 50:50 of A:B (A = 5% methanol/95% 1 mM triethylamine, pH 2.5; B = 100% methanol) and ending with pure B in eight minutes (flow rate of 1 ml per minutes). Product and substrate were detected at 270 nm. The area of the *p*-nitrobenzyl alcohol product peak was calculated and subtracted from the area of the same peak from a sample containing *E. coli* without a pNB esterase gene. This controls for the small quantities of free product in the substrate preparation and any interference from bacterial contamination. This final area was used as a measure of total activity, which was normalized by cell density.

Acknowledgments

The authors thank Dr W. P. C. Stemmer (Maxygen) for many helpful discussions, Dr Steve Queener (Eli Lilly & Co.) for providing us the wild-type pNB esterase and the challenge to evolve it, and Ms Rebecca Little (Eli Lilly & Co.) for assistance with DNA sequencing. This research is supported by the US Department of Energy's program in Biological and Chemical Technologies Research within the Office of Industrial Technologies, Energy Efficiency and Renewables. O. K. is supported by an NIH predoctoral training fellowship from the National Institute of General Medical Sciences, NRSA Award 1 T32 GM 08346-01.

References

Chen, K. & Arnold, F. (1993). Tuning the activity of an enzyme for unusual environments: sequential random mutagenesis of subtilisin E for catalysis in dimethylformamide. *Proc. Natl Acad. Sci. USA,* **90**, 5618–5622.

Chen, Y., Usui, S., Queener, S. W. & Yu, C. (1995). Purification and properties of *p*-nitrobenzyl esterase from *Bacillus subtilis*. *J. Ind. Micro.* **15**, 10–18.

Crameri, A., Whitehorn, E. A., Tate, E. & Stemmer, W. P. C. (1996). Improved green fluorescent protein by molecular evolution using DNA shuffling. *Nature Biotech.* **14**, 315–319.

Maynard Smith, J. (1988). The Evolution of Recombination. In *The Evolution Of Sex: An Examination Of Current Ideas,* pp. 106–125, Sinauer Associates, Inc, Sunderland, Mass.

Moore, J. C. & Arnold, F. H. (1996). Directed evolution of a para-nitrobenzyl esterase for aqueous-organic solvents. *Nature Biotech.* **14**, 458–467.

Moore, J. C. & Arnold, F. H. (1997). Optimization of industrial enzymes by directed evolution. *Advan. Biochem. Eng.* **58**, 1–14.

Müller, H. J. (1932). Some genetic aspects of sex. *Amer. Nature,* **66**, 118–138.

Shafikhani, S., Siegel, R. A., Ferrari, E. & Schellenberger, V. (1997). Generation of large libraries of random mutants in *Bacillus subtilis* by PCR-based plasmid multimerization. *Biotechniques*, in the press.

Stemmer, W. P. C. (1994a). Rapid evolution of a protein *in vitro* by DNA shuffling. *Nature,* **370**, 389–391.

Stemmer, W. P. C. (1994b). DNA shuffling by random fragmentation and reassembly: *in vitro* recombination for molecular evolution. *Proc. Natl Acad. Sci. USA,* **91**, 10747–10751.

Wells, J. A. (1990). Additivity of mutational effects in proteins. *Biochemistry,* **29**, 8509–8517.

You, L. & Arnold, F. H. (1996). Directed evolution of subtilisin E in *Bacillus subtilis* to enhance total activity in aqueous dimethylformamide. *Protein Eng.* **9**, 77–83.

Zhao, H. & Arnold, F. H. (1997a). Optimization of DNA shuffling for high fidelity recombination. *Nucl. Acids Res.* **25**, 1307–1308.

Zhao, H. & Arnold, F. H. (1997b). Functional and non-functional mutations distinguished by random recombination of homologous genes. *Proc. Natl Acad. Sci. USA,* **94**, 7997–8000.

Zock, J., Cantwell, C., Swartling, J., Hodges, R., Pohl, T., Sutton, K., Rosteck, P., Jr, McGilvray, D. & Queener, S. (1994). The *Bacillus subtilis* pnbA gene encoding *p*-nitrobenzyl esterase: cloning, sequence and high-level expression in *Escherichia coli*. *Gene,* **151**, 37–43.

Edited by J. Wells

(Received 10 March 1997; received in revised form 18 June 1997; accepted 9 July 1997)