1448–1459 *Nucleic Acids Research, 2004, Vol. 32, No. 4*
DOI: 10.1093/nar/gkh315

SURVEY AND SUMMARY

Chemical and biochemical strategies for the randomization of protein encoding DNA sequences: library construction methods for directed evolution

Cameron Neylon*

School of Chemistry, University of Southampton, Highfield SO17 1BJ, UK

Received January 6, 2004; Revised and Accepted February 6, 2004

ABSTRACT

Directed molecular evolution and combinatorial methodologies are playing an increasingly important role in the field of protein engineering. The general approach of generating a library of partially randomized genes, expressing the gene library to generate the proteins the library encodes and then screening the proteins for improved or modified characteristics has successfully been applied in the areas of protein–ligand binding, improving protein stability and modifying enzyme selectivity. A wide range of techniques are now available for generating gene libraries with different characteristics. This review will discuss these different methodologies, their accessibility and applicability to non-expert laboratories and the characteristics of the libraries they produce. The aim is to provide an up to date resource to allow groups interested in using directed evolution to identify the most appropriate methods for their purposes and to guide those moving on from initial experiments to more ambitious targets in the selection of library construction techniques. References are provided to original methodology papers and other recent examples from the primary literature that provide details of experimental methods.

INTRODUCTION

Directed molecular evolution has earned a secure position in the range of techniques available for protein engineering [for recent general reviews see (1–5)]. Despite continued advances in our understanding of protein structure and function, it is clear that there are many aspects of protein function that we cannot predict. It is for these reasons that ‘design by statistics’ or combinatorial strategies for protein engineering are appealing. All such combinatorial optimization strategies require two fundamental components: a library, and a means of screening, or selecting from, that library. The application of combinatorial strategies to protein engineering therefore requires, above all else, the construction of a library of variant proteins. The most straightforward method of constructing a library of variant proteins is to construct a library of nucleic acid molecules from which the protein library can be translated. This also has the advantage that (as long as a link between protein and nucleic acid is maintained) the identity of any selected protein can be directly determined by DNA sequencing. Much of the appeal of directed evolution of proteins lies in the fact that the coding information is held in a molecular medium which is straightforward to amplify, read and manipulate, while the functional molecule, the protein, has a rich chemistry that provides a wide range of possible activities. The methodology of choice in a directed evolution experiment is therefore to construct a library of variant genes, and screen or select from the protein products of these genes. Advances in screening methodology have been reviewed elsewhere and will not be discussed here [see (6) for a recent review and papers in (7) for detailed protocols]. A wide variety of methods have been developed for the construction of gene libraries. The most recent collection of detailed protocols may be found in (8). The purpose of this review is to give an overview of the different methods available and how they relate to each other, as well as how they may be combined.

Methods for the creation of protein-encoding DNA libraries may broadly be divided into three categories (Fig. 1). The first two categories encompass techniques that directly generate sequence diversity in the form of point mutations, insertions or deletions. These can be divided in turn into methods where changes are made at random along a whole gene and methods that involve randomization at specific positions within a gene sequence. The first category, randomly targeted methods, encompasses most techniques in which the copying of a DNA sequence is deliberately disturbed. These methods, which include the use of physical and chemical mutagens, mutator strains and some forms of insertion and deletion mutagenesis as well as the various forms of error-prone PCR (epPCR), generate diversity at random positions within the DNA being copied. The second category of methods targets a controlled level of randomization to specific positions within the DNA sequence. These methods involve the direct synthesis of mixtures of DNA molecules and are usually based on the

---

*Tel: +44 23 8059 4164; Fax: +44 23 8059 6805; Email: D.C.Neylon@soton.ac.uk

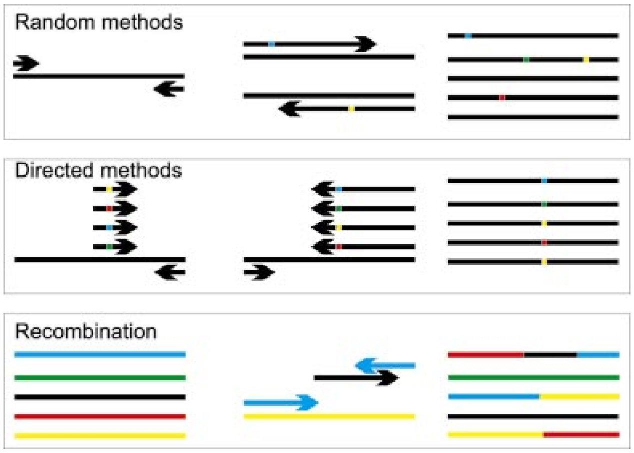

Figure 1. Overview of methods for the randomization of DNA sequences. Random methods introduce changes at positions throughout the gene sequence. Directed methods will randomize only a specific position or positions. Recombination methods bring existing sequence diversity, either from point mutants or from different parental DNA sequences, together in novel combinations.

incorporation of partially randomized synthetic DNA cassettes into genes via PCR or direct cloning. The key to these methods is the introduction of diversity at specific positions within the synthetic DNA. Thus these two approaches to generating diversity are complementary. The third category of techniques for library construction are those that do not directly create new sequence diversity but combine existing diversity in new ways. These are the recombination techniques, such as DNA shuffling (9,10) and the staggered extension process (11) that take portions of existing sequences and mix them in novel combinations. These techniques make it possible to bring together advantageous mutations while removing deleterious mutations in a manner analogous to sexual recombination. Also belonging to this category are methods such as iterative truncation for the construction of hybrid enzymes (12) that make it possible to construct hybrid proteins even when the genes have little or no sequence homology. It should be noted that, while these techniques do not in principle produce new point mutations, they are generally dependent on a PCR reconstruction process that can be error prone, and new point mutations are usually produced as a by-product of these techniques.

This review will discuss each of these categories of methodology in turn, highlighting the advantages and disadvantages of each methodology, the characteristics of the libraries that each method would ideally produce and the known or likely deviations from this ideal in reality. The patent status of different techniques will also be discussed.

GENERATING DIVERSITY THROUGHOUT A DNA SEQUENCE

Generation of mutations by directly damaging DNA with chemical and physical agents has been used to dissect biological systems for many years and will not be discussed in detail here. However, it does provide a valuable point of comparison to other methodologies. The basis of mutagenesis by UV irradiation or alkylating agents is that the damaged DNA is incorrectly replicated or repaired leading to mutation. The idea of relaxing the, usually very high, fidelity of DNA replication is also exploited in mutator strains. These bacterial strains have defects in one or several DNA repair pathways leading to a higher mutation rate. Genetic material that passes through these cells accumulates mutations at a vastly higher rate than usual. This is an effective and straightforward way of introducing mutations throughout a DNA construct. However, in common with physical and chemical mutagens the mutagenesis is indiscriminate. Thus the construct carrying the gene of interest as well as the gene itself, and indeed the chromosomal DNA of the host cell, suffers mutation. The process of mutagenesis using mutator strains can also be quite slow as the level of mutagenesis is controlled by the length of time the DNA spends in the strain. Constructing a library with mutagenesis levels of one or two nucleotide changes per gene can require multiple passages through the mutator strain. It is primarily this second disadvantage that has lead to the almost universal use of error-prone PCR methods for the generation of diversity for directed evolution experiments. However the simplicity of mutator strains will appeal to groups entering the area of directed evolution, particularly those with less experience in molecular biology. They may also be the most appropriate methodology when a simple initial experiment is required to generate preliminary results. The XL1-Red strain, commercially available from Stratagene, has been used in most experiments that utilize this strategy [for examples see (13,14)]. The use of mutator strains for library construction has recently been reviewed (15).

Error-prone PCR

The error-prone nature of the polymerase chain reaction has been an issue almost since its initial development. However, even the relatively low fidelity Taq DNA polymerase is too accurate to be useful for the construction of combinatorial libraries under standard amplification conditions. Increases in error rates can be obtained in a number of ways. One of the most straightforward and popular methods is the combination of introducing a small amount of Mn^{2+} (in place of the natural Mg^{2+} cofactor) and including biased concentrations of dNTPs (16,17). The presence of Mn^{2+} along with an over-representation of dGTP and dTTP in the amplification reaction leads to error rates of ~1 nt/kb in the final library [see (17–20) for examples that give detailed protocols]. The level of mutagenesis can be controlled within limits by the proportion of Mn^{2+} in the reaction or by the number of cycles of amplification (17). For higher rates of mutagenesis Zaccolo *et al.* (21) report a number of nucleoside triphosphate analogues that lead to high and controllable levels of misincorporation. Performing the reaction in the presence of these modified bases leads to mutation rates of up to one in five bases (21,22). In addition to these ‘home-made’ approaches, kits are available from both Clontech (Diversify PCR Random Mutagenesis Kit) and Stratagene (GeneMorph System). The Clontech system is based on Taq polymerase and provides ready-mixed reagents for modifying mutation rates by changing the concentrations of Mn^{2+} and dGTP. The Stratagene GeneMorph systems is based on a highly error-prone polymerase. This kit is straightforward to use and comes with detailed instructions, and therefore is appealing to those entering the area. The level of mutagenesis is controlled by the concentration of template used and the number of serial amplification reactions performed.

The bias problem

The methodologies for error-prone PCR all involve either a misincorporation process in which the polymerase adds an incorrect base to the growing daughter strand and/or a lack of proofreading ability on the part of the polymerase. The inherent characteristics of the polymerase used mean that some types of error are more common than others (17,21). The appeal of the error-prone PCR approach is that it leads to randomization along the length of the DNA sequence, ideally leading to a library in which all potential mutations are equally represented. However, the construction of such an ideal library relies on all possible mutations occurring at the same frequency. The fact that specific types of error in the amplification process are more common than others means that specific mutations will occur more often than others, leading to a bias in the composition of the library. This 'error bias' means that libraries have non-random composition with respect to both the position and the identity of changes. Error-prone PCR using Taq and the Stratagene GeneMorph kit have different biases, making it possible to use a combination of techniques to construct a less biased library (23,24).

There are two other sources of bias in libraries constructed by error-prone PCR. The first of these is a 'codon bias' that results from the nature of the genetic code. Error-prone PCR introduces single nucleotide mutations into the DNA sequence. Even without error bias single mutations will lead to a bias in the variant amino acids that the mutated DNA encodes. For example, single point mutations in a valine codon are capable of encoding phenylalanine, leucine, isoleucine, alanine, aspartate or glycine. To access the codons for other amino acids either two point mutations (C, S, P, H, R, N, T, M, E, Y) or even three (Q, W, K) are required. The result of this codon bias is that specific amino acid changes will be much less common than others in any library constructed by error-prone PCR. It can be argued that this bias in the genetic code is optimized to ensure that amino acid substitutions are biased towards those that are less likely to cause loss of function (25). However, where the object is to efficiently screen a specific subset of possible mutations there is no advantage in being forced to screen 100 valine to alanine conversions to be sure that the valine to tryptophan mutation *is* in fact deleterious.

The final source of bias, 'amplification bias', is a characteristic of any mutagenesis protocol that involves an amplification step, particularly PCR amplification. PCR is by its nature an exponential amplification process. If, in an imaginary PCR amplification reaction from a single molecule, a mutation is introduced in the first copying step, then this mutation will be present in 25% of the product molecules. Such an extreme situation is unlikely to arise in a real experiment but the point is clear. Any molecule that is copied early in the amplification process will be over-represented in the final library. Owing to the exponential nature of the amplification such an over-representation can be serious. However, such bias can be difficult to detect by sequence analysis, particularly in large libraries. The problem can to some extent be overcome by performing several separate error-prone PCRs and combining these to construct the final library. Another strategy is to reduce the number of amplification cycles, but changing the number of amplification cycles is also one of the most straightforward ways of controlling the level of mutagenesis. A combination of multiple amplification reactions and reducing the number of amplification cycles is the most effective means of combating this form of bias. Amplification bias could be a serious problem when statistical conclusions are being sought from experiments. It is not, however, a significant issue when the aim is the improvement of protein characteristics (as long as results of the desired quality are obtained).

Introducing controlled deletions and insertions at random locations: a new type of sequence diversity

Error-prone PCR protocols are effective at changing the DNA sequence; converting one nucleotide to another. In contrast, single nucleotide insertions, and more commonly deletions, are produced but at a much lower rate. This is desirable for library construction as single nucleotide insertions and deletions lead to frameshifts which complicate analysis and screening, and often simply produce truncated proteins. However, the insertion and deletion of amino acids in the protein structure is clearly a desirable type of diversity to explore, providing a new 'dimension' in the protein sequence search space. Although it is possible to introduce the codons for amino acids at specific positions within the protein-encoding DNA by oligonucleotide-based methods (see below), it is only recently that techniques have been reported for the insertion and deletion of codons at random locations throughout the gene. There has therefore been relatively little investigation of the value of random insertion and deletion in directed evolution experiments to date. The availability of some recently developed techniques (26,27) now makes these investigations possible.

The introduction of modified transposons into DNA sequences as a means of creating controlled insertions at random locations has been reviewed by Hayes and Hallet (28). These methods require specialist DNA constructs and are limited in that only specific sequences can be introduced. A general method for creating deletions and repeats at random locations and of random lengths is described by Pikkemaat and Janssen (29). This method utilizes Bal31 nuclease to delete DNA from one end of the template gene. The 5′- and 3′-ends of the gene are treated in separate pools and then recombined by ligation. The ligated products will either contain deletions or sequence repeats. The process is relatively straightforward and easy to perform. One disadvantage of this approach is that the majority of 5′- and 3′-fragments will be ligated out of frame, leading to nonsense mutations. The other is that sequence material is limited to that of the source DNA. However, as the authors argue (29), this is a known pathway for natural evolution, and in particular it is known to be important in the evolution of their system, the haloalkane dehalogenases.

In 2002 an elegant and general methodology for inserting or deleting sequences of defined length and identity was reported by Murakami *et al.* (26,27). This method, termed random insertion/deletion (RID) mutagenesis, enables the deletion and subsequent insertion of an arbitrary number of bases of arbitrary sequence at random along a target gene sequence. In the format as described up to 16 bases can be inserted or

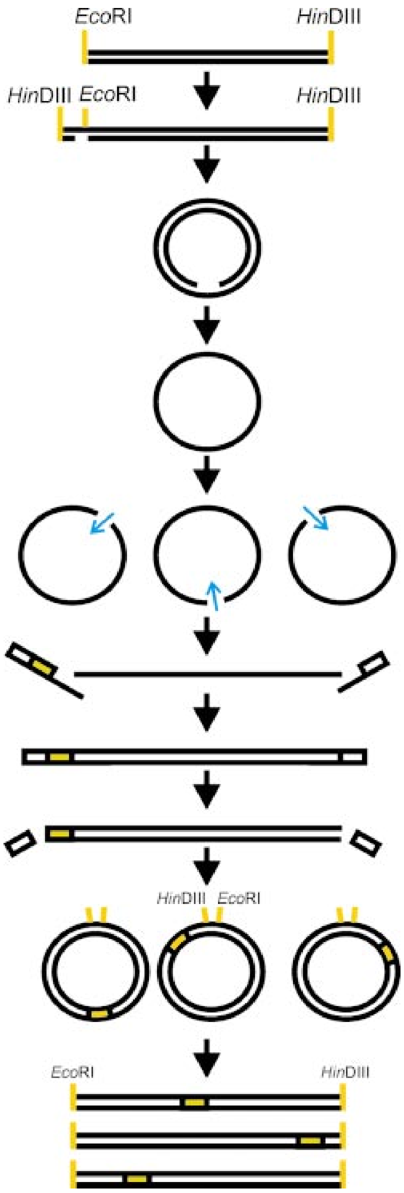

Figure 2. Random insertion/deletion mutagenesis (RID). The template DNA is converted to a covalently closed single-stranded circle which is cleaved at random sites by Ce(IV)-EDTA treatment. Linker fragment cassettes are then annealed to each end of the cleaved single-stranded DNA via a 10 nt random tail. The construct is amplified using primer sites in the cassettes to produce the second DNA strand. Finally the cassettes are cleaved off using a type II restriction enzyme (recognition site in the cassette) to leave the insertion or deletion behind. The remaining construct containing the modification is converted back to double-stranded circular DNA that can be cleaved with appropriate restriction sites to produce the gene library in a form ready for cloning. Adapted from Murakami *et al*. (26) with permission from Nature Publishing Group (http://www.nature.com/nbt/).

deleted. The method (Fig. 2) is based on ligating an insertion cassette (or deletion cassette) at random locations within the gene. The key elements of the RID method are that the identity of any inserted element can be absolutely controlled by the design of the cassette, and the use of Ce(IV) as an oxidative cleavage agent makes the position of insertion/deletion reasonably random [although there is still some bias; see Fig. 5A of (26)]. Another major advantage of the RID method is the ability to perform deletions *and* insertions concurrently. The authors show this by deleting three nucleotides and replacing them with a mixture of 20 codons, one for each amino acid. Thus as well as providing a means for generating insertions and deletions, the RID method can also be used to generate libraries of single amino acid mutations without any codon bias.

The disadvantage of the RID method is that it is a complex multistep procedure requiring a significant investment in resources, particularly in the time to insure that each step is working correctly. This is therefore a method for those who are prepared to deploy the extra resources required. If it is demonstrated that the introduction of insertions and deletions is valuable in improving protein function, then it will become an important tool in the set of library construction techniques. It is unlikely that the ability to avoid codon bias will appeal sufficiently on its own to most users to justify using RID.

Summary

Error-prone PCR methods remain one of the most popular approaches for generating libraries for directed evolution experiments. The ease with which mutations can be generated by modifying PCR conditions (addition of Mn^{2+}, biasing of dNTP concentrations or addition of dNTP analogues) makes these methods appealing for any laboratory that is approaching directed evolution as a means to an end. The use of mutator strains is somewhat less popular but may be particularly useful for laboratories with less experience in molecular biology. Most of these methods produce libraries with a bias in the type of nucleotide mutations (error bias), a bias in the types of amino acid changes seen in the protein (codon bias) and a bias in the distribution of specific sequences in the library (amplification bias). The error bias can be overcome to a certain extent by combining libraries constructed via Taq-based PCR with those constructed using the Stratagene GeneMorph kit which has a different bias. Most reported directed evolution experiments have used either error-prone PCR, some form of shuffling (below) or a combination of both. It remains to be seen whether more difficult directed evolution experiments will require an elimination of these forms of bias. For those investigating this issue, or those who require less bias for other experimental reasons, the RID methodology offers a general method to reduce this bias as well as accessing insertions and deletions at random positions within the amino acid sequence. This new dimension of diversity provided by amino acid insertions and deletions remains largely unexplored.

DIRECTING DIVERSITY: OLIGONUCLEOTIDE-BASED METHODS

The techniques described above all, at least ideally, generate diversity along the length of a DNA sequence. The techniques discussed in this section are at the opposite extreme and at their simplest randomize a single position in the target gene. All of these techniques are based on the incorporation of a synthetic DNA sequence within the coding sequence. The synthetic DNA is randomized at specific positions and this randomization is incorporated directly into the target gene. There are therefore two elements to all of these techniques:

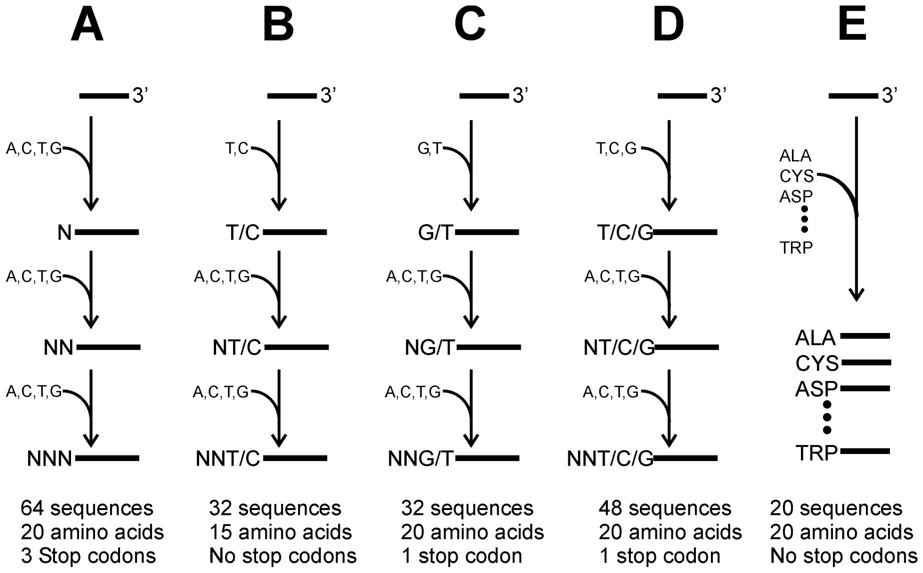

Figure 3. Approaches to randomizing synthetic DNA. Examples show randomization of one codon with mixed nucleotides (NNN, NNT/C, NNG/T or NNT/G/C) and with trinucleotide phosphoramidites. Synthesis in all three cases commences conventionally 3′ of the randomized codon. At the 3′-end of the randomized codon (**A**) all four nucleotides, (**B**) a mixture of T and C, (**C**) a mixture of G and T or (**D**) a mixture of T, G and C can be added. In each case a mixture of all four nucleotides is added at each of the remaining two positions. Having a mixture of G and C at the 3′-end of the codon will provide 32 codons, all 20 amino acids and one stop codon. (**E**) Conversely, the codon can be synthesized by the direct addition of a mixture of 20 trinucleotide phosphoramidites in one step. ALA–TRP represent 20 presynthesized 3-nt codons, one to code for each amino acid.

first the means by which the DNA itself is randomized during synthesis, and secondly the methodology for incorporating the synthetic oligonucleotide. These two issues will be discussed separately, although some issues raised by one can be dealt with by the other and vice versa.

The synthesis of randomized oligonucleotides

The value of oligonucleotide-based mutagenesis is that control over the chemistry of DNA synthesis allows complete control over the level, identity and position of randomization. Thus, if an oligonucleotide can be synthesized as a mixture, or if a number of synthetic oligonucleotides can be mixed, then this can be incorporated directly into a complete gene sequence. There are a wide range of techniques from the field of combinatorial chemistry that are available to a combinatorial biologist. Indeed, the biologist has an advantage over the chemist as a mixture of genes can be readily separated for analysis by transformation into bacterial cells and isolation of single transformed colonies.

The synthesis of degenerate oligonucleotides is well established; synthetic primers incorporating mixtures of any combination of the four natural bases at any position can be ordered directly from most suppliers. Such pieces of synthetic DNA can be used to completely randomize a specific position within a gene. The synthesis of ‘doped’ oligonucleotides, where a small proportion have a mutation at a specific position or positions, is a slightly more specialist process, but oligonucleotides of this type can be ordered from most suppliers. These are used to generate libraries where the randomization is spread out but still targets those positions that are doped in the primers. Any synthetic process where a number of reagents are used as mixtures is susceptible to bias arising from greater incorporation of one reagent than another. Quantitative studies indicate that where synthesis is carefully controlled and/or uses optimized reagents (e.g. Transgenomic’s ‘Precision Nucleotide Mix’), this bias is small in synthetic DNA libraries (30,31). It should be noted that this relative lack of bias is not maintained when these libraries are cloned, although the reason for this is not clear (31).

Another bias problem arises due to the mismatch between the base-by-base synthesis of the oligonucleotide and the triplet nature of the genetic code. To randomize a codon so that it can encode all 20 amino acids, a mixture of all four bases is required at the first two positions and at least three bases in the third position. This in turn leads to a form of codon bias as there are six times as many codons for some amino acids, such as serine, than others such as tryptophan and methionine. In addition, there is the potential for the introduction of stop codons. This can be avoided by limiting the mixture of bases at the third position of the codon to T and C, but this means that codons for a range of amino acids will not be present (Fig. 3). A compromise is to randomize the codon with T, C or G in the final position, giving only one stop codon in every 48 primers, and encoding all 20 amino acids or NNG/T or NNG/C which provide all amino acids with slightly more common stop codons. Another result of this form of codon bias is that it is difficult to insert codons for a subset of amino acids if this is desirable.

A number of solutions have been developed to this problem. The simplest solution is to synthesize the DNA for each desired mutation separately. For relatively small libraries the falling cost of oligonucleotide synthesis makes this possible with the size of the library limited by the size of the budget and not by technical considerations. The oligonucleotides can then either be mixed or used separately to construct the gene library. A second solution is to use trinucleotide phosphoramidites in the synthesis of the oligonucleotides. This solves the problem of the codon bias by synthesizing the DNA one codon at a time. If it is desired to completely randomize one amino acid, then a mixture of 20 codons can be added (Fig. 3). If a low level of mutagenesis is required then the mixture will be present at a lower concentration than the wild-type codon, and if a subset of amino acids is desired then this is easily accommodated by including only the desired codons. However the trimer phosphoramidites are not straightforward (or cheap) to prepare. A number of syntheses are described in the literature (32–34) with probably the most appealing being the large-scale solid phase synthesis described by Kayushin *et al.* (35). These reagents have recently become commercially available from Glen Research making the strategy more accessible to the general user. Twenty specific trinucleotide phosphoramidites are available, one for each amino acid, as well as a mixture of all 20 prepared for direct use in oligonucleotide synthesis.

The difficulty involved in synthesizing and using trinucleotide phosphoramidites led Gaytan *et al.* (36) to develop a strategy based on orthogonal protecting groups. In this strategy the wild-type sequence is synthesized using standard acid labile trityl protecting groups, but at each point where mutagenesis is desired the penultimate phosphoramidite is spiked with a small proportion of Fmoc protected monomer. The synthesis of the wild-type codon continues with standard trityl chemistry. Once the wild-type codon is complete the base labile Fmoc groups are removed and the mutagenic codon is synthesized with Fmoc chemistry. This can also be used to remove stop codons while still maintaining access to codons for all 20 amino acids. The Fmoc protected phosphoramidites are relatively straightforward to synthesize (37) in comparison with trinucleotide phosphoramidites, making this strategy more accessible. However, it is still limited to laboratories with access to a DNA synthesizer and synthetic experience.

Another strategy, which is logically similar, is derived from classical split and mix approaches. In this case, instead of being differentiated by protecting groups a proportion of oligonucleotides destined for mutagenic codons are physically separated from the wild-type sequences (38,39). Again this methodology requires access to a DNA synthesizer and, as originally reported, requires extensive manipulations to allow for the removal and recombining of the solid support. Using this type of approach it is possible to prepare a library of oligonucleotides that target multiple positions with no codon bias, and with only one codon being randomized in each DNA sequence, without requiring any reagents beyond those required for standard DNA synthesis.

Overall, while modern chemistry makes a high level of control over the make-up of a library of oligonucleotides possible, most of these sophisticated approaches require access to a DNA synthesizer at a minimum and can require considerable expertise and resources for synthesis as well. These techniques are likely to remain the preserve of specialist laboratories. Most users are restricted to the choice between NNN, NNT/C, NNT/G, NNG/C and NNT/G/C codons in their oligonucleotides or the option of ordering a large number of individual mutagenic sequences. In most cases this is not a serious restriction as the most common use of oligonucleotide-directed mutagenesis is to randomize a limited number of positions. Most commonly a primer-directed method is used to completely randomize specific positions that have been identified by screening libraries constructed by error-prone PCR (25,40–42). As the library sizes are relatively small, an inefficiently constructed library is not a serious drawback. The problem of stop codons and codon bias only becomes a serious issue when the libraries are very large and efficiency is crucial, or where statistical analysis is required.

Incorporating synthetic DNA into full-length genes

Regardless of how an oligonucleotide is synthesized, whether a single codon or several are randomized and whether the level of mutagenesis is high or low, it is necessary to incorporate the synthetic DNA sequence into a full-length gene. A wide range of methods are available based on conventional site-directed mutagenesis techniques and these will not be reviewed in detail here. The basic requirement of incorporation is that the level of wild-type sequence contamination should be as low as possible. For this reason PCR-based techniques such as strand overlap extension (SOE) and megaprimer-based procedures are usually the method of choice [for examples and protocols see (43–47)]. Some groups have used mutagenic plasmid amplification (MPA) (marketed in kit form by Stratagene as QuikChange system) and related methods successfully. The QuikChange system can also be used with megaprimers (48,49), meaning only one mutagenic primer is required. Synthetic primers can also be incorporated via a number of recombination strategies and these will be discussed below.

The problem of bias arises again with primer incorporation procedures. As discussed above, with any procedure that includes an exponential amplification there is the potential problem of amplification bias. If a great effort has been expended on removing bias from the oligonucleotide library, then reducing it at later stages of the construction process is clearly desirable. An additional problem with primer incorporation is that those sequences with greatest similarity to the wild-type DNA sequence will be incorporated more efficiently than those that diverge more. In a PCR-based strategy primers mutagenized near the 5′-end will be more efficiently incorporated than those modified near the 3′-end. Careful design of primers and the provision of a reasonable length of fully annealing sequence at the 3′-end can reduce the risk of this occurring. Reduction in amplification bias is again best achieved by performing a number of separate amplifications with the smallest possible number of cycles.

Simple methodologies are capable of randomizing a gene in a single region, but the length of oligonucleotides that can be reliably synthesized limits the size of the region that can be randomized. Randomizing multiple regions requires either multiple rounds of mutagenesis or more complex methods. The assembly of designed oligonucleotides method (50) and synthetic shuffling (51) utilize a shuffling type approach with synthetic oligonucleotides, and the QuikChange system can be used with multiple mutagenic primers to randomize multiple positions (52). Another similar approach is to construct overlapping gene segments by PCR with mutagenic primers and then reconstruct these in an overlap extension reaction. If a small number of gene fragments (up to four) are used, then the reconstruction can be performed by strand extension rather than PCR amplification (i.e. without external primers). This

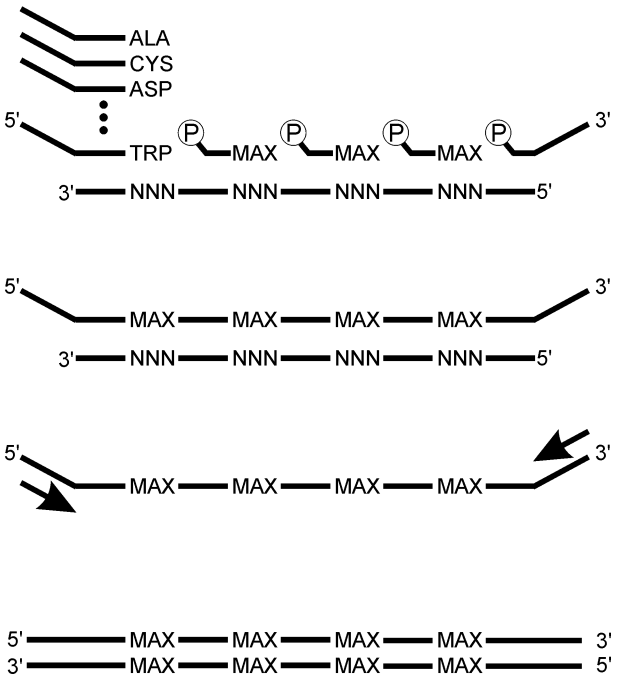

Figure 4. The MAX method of library generation. A template sequence contains the codons designated for randomization in the form of NNN triplets. A set of selection primers for each position contains those specific triplets that are desired in the final library. The selection primers are annealed to the template. Completely annealed selection primers can then be ligated to each other to form a full-length DNA fragment. Finally, the ligated selection primers are converted to double-stranded DNA for further manipulations.

has the advantage of reducing the risk of amplification bias. There is a linear rather than exponential amplification bias in the construction of the fragments as the randomized region is in the primers. Using PCR rather than strand extension in the second step will lead to exponential bias towards those full-length fragments copied early in the amplification process. However, PCR is required to reconstruct genes from larger numbers of fragment as the yield of full-length product from the strand extension reaction drops rapidly as the number of fragments increases.

MAX randomization: beating the bias problem with ordinary oligonucleotides

The appeal of a trinucleotide-phosphoramidite-based synthesis, or the type of split-and-mix strategy pursued by Lahr *et al.* (38), is the removal of codon bias and the ability to include codons for any subset of amino acids at a given position. A recently described oligonucleotide incorporation approach provides many of the same advantages while only requiring simple primers (53). The MAX system described by Hine and coworkers (53) relies on the annealing of specific oligonucleotides to select the specific subset of codons desired from a template that is completely randomized at the target codons (Fig. 4). A template oligonucleotide is synthesized covering the region to be mutagenized, with each target codon completely randomized (NNN). Specific primers are then synthesized that cover the region 5′ to each target codon and terminate with each specific codon that is required for inclusion in the library. Thus if codons for all 20 amino acids are required, then 20 ‘selection primers’, one with each codon, are synthesized. A set of primers is synthesized for each codon to be randomized. These primers are then annealed to the template and ligated. The selection primers will only ligate to the primer immediately 5′ when that primer is completely annealed to the template. The selection primers therefore select a subset from the 64 codons in the template. The ligated single strand containing the selection primers is then converted to dsDNA for incorporation into the full-length gene.

The advantage of the MAX system is that, although a sizable number of primers are required, the maximum number will be 20 times the number of codons to be randomized. If three codons are randomized then 60 primers are required, but these 60 can be used to construct a library containing 8000 mutants with no codon bias. Amplification bias is a potential issue if PCR is used to construct the second strand. This can be reduced by the usual methods or by using a simple second-strand synthesis rather than PCR. The MAX system is therefore an excellent means of randomizing multiple codons within a single region. The reconstructed double-stranded DNA can then be either ligated directly into an expression construct or used in strand extension reactions to reconstruct the full-length gene. MAX is not advantageous if a single codon is to be randomized and cannot be used if more than two adjacent codons are to be independently randomized. Again, it is a more complex technique and is therefore less likely to appeal to the general user. However, it allows the construction of unbiased libraries by users without access to a DNA synthesizer and will therefore be extremely valuable where efficient screening of medium to large (10^3–10^6 variants) libraries is required.

Summary

Oligonucleotide-directed methods offer a very powerful route to randomizing specific chosen positions and regions within protein encoding DNA sequences. The essence of oligonucleotide-directed methods is that a synthetic DNA sequence is incorporated into the full-length gene. This means that any form of randomization that can be achieved in a synthetic DNA fragment can be replicated in the full-length gene. A wide range of synthetic strategies are available that allow highly precise and controlled randomization within oligonucleotides. However, the majority of these techniques require access to a DNA synthesizer and some require the synthesis of reagents that are not, as yet, commercially available. In most cases randomization of codons to NNN or NNT/G/C, either at saturation levels or at some lower level, will be sufficient.

A range of straightforward techniques is available for incorporating synthetic DNA sequences into full-length genes. Overlap extension and megaprimer protocols are simple to use and are the most popular methods. The incorporation of multiple primers is more complex but can be achieved by a number of methods that are reasonably straightforward to apply. The MAX technique recently described by Hine and coworkers (53) offers an elegant and accessible approach to efficiently constructing libraries where multiple codons are randomized.

TECHNIQUES FOR RECOMBINATION OF DNA SEQUENCES

The methods described above all produce sequence diversity, either along the length of a sequence or at specific positions. Natural evolution, however, also exploits recombination to bring together advantageous mutations and separate out deleterious mutations. Until 1993 there were no random recombination methods available for directed evolution. The original DNA shuffling technique (9,10,54) allowed a step change in what was possible with directed evolution and is still one of the most popular tools in any optimization strategy. It was now possible to recombine a range of similar genes from different sources, or to combine selected point mutations in novel combinations. A number of other techniques are now available each with their own characteristics and uses including the staggered extension process (StEP) (11,55), random chimeragenesis on transient templates (RACHITT) (56,57) and the various techniques based on iterative truncation for the creation of hybrid enzymes (ITCHY) (12,58–60). All of these methods are based on linking gene fragments together. In the case of DNA shuffling, RACHITT and ITCHY these fragments are physically generated by cleavage of the source DNAs and then recombined, whereas in StEP the fragments are added to the growing end of a DNA strand in rapid rounds of melting, strand annealing and extension. The sum result for all these techniques is to bring fragments from different source genes into one DNA molecule, recombining the source DNA in new ways to form novel sequences (Fig. 5).

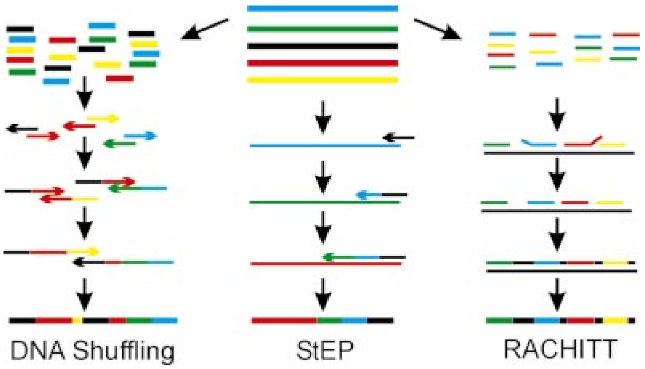

Figure 5. Homology based methods for recombining DNA sequences. All methods commence with a series of parental DNA sequences. In DNA shuffling this parental DNA is cleaved into random small fragments by DNAse digestion. The fragments are then used in a self-priming reaction to reconstruct the full-length DNA. In StEP the DNA is not fragmented. Instead, small segments are added to the end of a growing DNA strand in a series of very short extension steps. When the strand is removed from an initial template it can reanneal to another to generate a crossover. In RACHITT one parental DNA is used as a template. One strand of this template containing dUTP is generated. Fragments of the opposite strand of the other parental DNAs are then produced and annealed to the template. Non-annealed flaps are then removed by exonuclease digestion and remaining gaps filled in with a DNA polymerase. These fragments are then ligated together and the template strand removed by endonuclease V digestion. The single strand is then converted to double-stranded DNA for further manipulations.

DNA shuffling is the most popular of recombination techniques [see (20,44,61,62) for recent examples with experimental details] because it is straightforward to perform. Gene fragments are generated by digesting the source DNA molecules with DNAse. The size of the fragments can be selected to gain some control over the frequency of crossover between source sequences. Other methods for fragment generation, such as the use of endonuclease V treatment of source DNA with incorporated dUTP (63), have also been reported. The mixture of fragments is then subjected to repeated cycles of melting, annealing and extension. The quantity of full-length reconstructed sequence produced is very small, so the production of a reasonable quantity of full-length DNA requires PCR amplification. As the final PCR amplification is performed on a sample that originally contained only gene fragments, a successful amplification generally indicates a successful reconstruction. This makes it easy to tell whether the conditions are correct for reconstruction. However, it does not confirm that conditions are optimal for recombination.

StEP (11,55) also relies on repeated cycles of melting, annealing and extension to build up the full-length gene. However, in StEP fragments are added in steps to the end of a growing strand. The growing strand is prevented from reaching its full length by keeping the extension time very short. This results in only partial elongation of a strand in any one extension step. The strand is then melted from its template and may anneal in the next step to a different template leading to a crossover. StEP can be harder for an inexperienced user to set up than DNA shuffling as full-length templates are included in the StEP reaction. This means that the production of full-length template may indicate simple amplification rather than recombination. Balancing the need for yield and recombination can be challenging, as it is not always straightforward to determine whether recombination has occurred. Careful selection, or preparation, of templates to include convenient restriction endonuclease sites will aid optimization by making it easier to quantify the degree of recombination. Once optimized for a specific thermal cycler, primers and template, StEP can be easier to perform than DNA shuffling as fewer steps are involved [for examples see (55,64,65)].

RACHITT (56,57) is a technique that is conceptually similar to StEP and DNA shuffling but is designed to produce chimeras with a much larger number of crossovers. In this case, the fragments are generated from one strand of all but one of the parental DNAs. These fragments are then reassembled on the full-length opposite strand of the remaining parent (the transient template). The fragments are cut back to remove mismatched sections, extended and then ligated to generate full-length genes. Finally, the template strand is destroyed to leave only the ligated gene fragments to be converted to double-stranded DNA. The advantage of this assembly process is that it creates a greater number of crossovers. The disadvantage lies in the additional steps required; the generation of single strands, the removal of 'flaps' of non-annealing DNA and the removal of the template strand. Care is required in each of these steps as contamination and carry-over can lead to contamination of the library with parental sequence. None the less, RACHITT is the method of choice if a large number of crossovers at random positions is required.

The major difficulty with both DNA shuffling and StEP is that they rely on the annealing of a growing DNA strand to a template. This annealing is most likely to occur to a template sequence that is very similar to the 3′-end of the growing strand. Sequences can therefore only be recombined when they are similar enough to allow annealing, and crossovers will occur preferentially where the template sequences are most similar. It is common for a new user to sequence a number of products from a DNA shuffling reaction only to find that the full-length sequences are identical to the templates. RACHITT, although it does not rely on priming, is still limited to the incorporation of sequence elements that are similar to the template. There is now a significant quantity of literature available on computational methods to predict where crossovers will occur, optimization of sequences to increase the number of crossovers and prediction of optimal conditions for the recombination reaction (23,66–68). These computational methods can provide a good guide to whether given recombination experiments will work and what the optimal conditions are likely to be.

There is also a wide range of experimental techniques that can improve the chance of recombining sequences with limited homology or increase the distribution of crossover events. Recombination can be forced to occur by eliminating a segment of each template gene from the recombination reaction. For the full-length gene to be reconstructed it must combine portions from at least two templates. Restriction digest of each template can be used to force recombination in this manner (69). An alternative strategy is to use single-stranded DNA templates to prevent the formation of homoduplexes (70,71). However, the crossover events are still usually restricted to the region of highest homology, generating libraries with limited diversity. A complementary approach to increasing the number and distribution of crossover events is therefore to increase the homology of the template genes. This can be achieved by optimizing the template gene sequences directly, either by mutagenesis or by complete gene synthesis. The nucleotide homology of two genes can often be significantly increased without changing the encoded protein sequence, particularly if the template genes are from species with different codon preferences (67). Again, it is very valuable to design or modify source sequences to contain restriction endonuclease sites that will allow a rapid analysis of the degree of recombination that has occurred.

Recombination can also be increased by including synthetic oligonucleotides that combine sequence elements from two different templates in either a separate amplification step (72) or the shuffling reaction itself. This is, in a sense, a method of *directing* crossover events that bears the same relation to the generation of random crossover events as oligonucleotide-directed mutagenesis does to random mutagenesis. Thus each oligonucleotide will direct one specific crossover event and each desired crossover requires an oligonucleotide. This strategy is effective in combination with DNA shuffling in generating a much broader selection of crossover events. The methodology is taken to its logical extreme with the synthetic shuffling method described by Ness *et al.* (51) and the assembly of designed oligonucleotides (ADO) method described by Reetz and coworkers (50). In both approaches synthetic oligonucleotides are designed based on template sequences to generate recombined full-length genes entirely from synthetic DNA. The advantage of using entirely synthetic oligonucleotides is that absolute control over the synthesis procedure gives absolute control over the template sequence and position of crossovers, and also allows the introduction of specific point mutations.

Recombination without homology

The difficulty of generating recombination events where there is little sequence homology between template genes has led to the development of a number of techniques that do not require strand extension or annealing to a template. These methods are not appropriate for the recombination of point mutations, the most common use of DNA shuffling, but are useful where it is desirable to generate hybrids of genes that share little DNA sequence similarity. For instance Ostermeier and coworkers (58), created hybrids of human and bacterial glycinamide ribonucleotide transformylase enzymes. These enzymes have functional similarities and it was therefore of interest to examine hybrid enzymes. However, the genes share only 50% sequence homology.

The first description of a general technique for recombining non-homologous sequences was by Benkovic and coworkers. The method, termed incremental truncation for the creation of hybrid enzymes (ITCHY), is based on the direct ligation of libraries of fragments generated by the truncation of two template sequences, with each template being truncated from the opposite end. Fragments of one template that have been digested with exonuclease III and S1 nuclease from the 5′-end of the gene are ligated to fragments of the second template that have been digested from the 3′-end. This ligation process removes any need for homology at the point of crossover, but the result of this is usually that the connection is made at random. Thus the DNA fragments may not be connected in a way that is at all analogous to their position in the template gene and may be ligated out of frame, generating a nonsense product. The potential for generating out-of-frame products restricts the number of crossover events to one or perhaps two.

In the initial version of ITCHY this incremental truncation was performed via timed exonuclease digestions. This proved difficult to control and optimize, so an improved procedure was developed where initial templates are generated with phosphorothioate linkages incorporated at random along the length of the gene (58). Complete exonuclease digestion then generates fragments with lengths determined by the position of the nuclease resistant phosphorothioate linkage. This method, named thio-ITCHY, is much more straightforward to perform. Two variants are described, differing mainly in the way in which the phosphorothioates are incorporated, but for most users PCR-based incorporation is probably the most straightforward. Plasmid constructs to facilitate the use of these methods and for removing out-of-frame ligation products are also described (58,73). Libraries generated using ITCHY can be used as templates for further recombination by DNA shuffling to generate a hybrid with more than one crossover (60,74,75).

Recently, a number of papers have described general methods for the defined recombination of parental sequences. These methods generate crossovers at specific positions and do not rely on any sequence homology between the parents. O'Maille *et al.* (76) designed primers for the amplification of specific gene fragments from each parent that could then be reassembled by an overlap extension approach. The overlaps here were designed manually based on the structural and sequence homology of the parents. Hiraga and Arnold (77) have developed a method based on the insertion of tag sequences containing type II restriction enzyme sites into the parent genes. The tag sequences are designed to provide for the easy generation of specific fragments by the use of a single restriction enzyme and the correct reassembly of fragments via the different overhangs yielded by the type II restriction enzyme. The issue with such directed methods is the choice of where crossovers between parental sequences should be placed. Arnold and coworkers have reported and validated a general approach to identifying optimal crossover sites based on a computational analysis of the structures of the parental proteins to identify regions with minimal interactions with the rest of the protein (78,79), whereas O'Maille *et al.* (76) design the crossover points manually. In both cases structural information is important, and in many cases this will either be available or can be inferred from structure and sequence alignments.

Summary

A range of different techniques are available for recombining diverse sequences. DNA shuffling remains the most popular technique. It is an effective way of recombining sequences with high homology and is easy to set up and perform. A whole toolkit of techniques has grown up around DNA shuffling with methods to increase recombination between less related sequences and to optimize sequences and conditions for optimal recombination. StEP is a broadly similar technique to DNA shuffling although the implementation is very different. Both DNA shuffling and StEP suffer from one major problem: recombination is limited to parental genes with very similar sequences and crossover events are strongly biased towards regions of highest sequence similarity. A particular problem when attempting to recombine a number of point mutants of an original sequence is that the majority of recombination products will be either the original wild-type sequence or the unrecombined point mutants (80). Techniques to overcome this problem generally rely on the inclusion of synthetic oligonucleotides in the shuffling reaction to encourage specific crossover events (72) or the exclusion of specific regions of template genes from the final product (69). The use of synthetic oligonucleotides reaches its logical conclusion in methods where recombination is performed between entirely synthetic sequences, allowing optimization of the template DNA sequence, crossover points and the addition of further point randomization if desired (50,51). Benkovic and coworkers have described a range of related techniques that allow recombination between two unrelated template sequences. These methods rely on truncation of the two templates from opposite ends followed by religation of the remaining fragments together. The method is effective but is limited to products with a single recombination event between two template sequences. Recombination at any set of specific positions can be performed by primer-based methods such as strand extension or by the incorporation of specific restriction enzyme sites.

Overall, DNA shuffling looks set to remain the most popular method for recombination. The combination of error-prone PCR followed by shuffling of selected mutants with improved function is the most commonly followed strategy for directed evolution experiments. DNA shuffling has drawbacks and creates biased libraries but equally has produced good results. The occasional user will find DNA shuffling the most straightforward method to use. Again, specific experiments that require less bias or more efficient libraries and those that require statistical analysis may require either improved or modified techniques. General users will require experimental demonstration that the more complex techniques are required for their specific application.

PATENT AND LICENSING ISSUES

The targets of most directed evolution experiments are usually technologically based and often have some commercial value. It is therefore worth noting which randomization methods are protected by patent. The underlying use of PCR in the majority of these methods is covered by the original patents for the polymerase chain reaction held by Hoffman–La Roche. It is however, worth noting that some commercially available thermostable polymerases, such as Vent from New England Biolabs, are apparently not licensed for PCR. The use of Mn^{2+}, biased nucleotides and dNTP analogues to increase the error rate of PCR amplification is not protected. The DNA polymerase on which the Stratagene GeneMorph mutagenesis system is based is not apparently patented for the purpose of mutagenesis. The standard license agreement for use of the GeneMorph system is not limited according to the instruction manual. Both DNA shuffling and StEP have been protected. Patents for DNA shuffling (US5605793, US5830721 and US6506603) including synthetic shuffling (US6521453) are held by Affymax Technologies and Maxygen, while those for StEP (US6153410 and US6177263) are held by the California Institute of Technology. As these techniques form such a crucial part of virtually all directed evolution experiments it may be useful to consider licensing issues when deciding on which method to use where any commercial outcome is expected. The various ITCHY methods have not been patented.

Of the more recently described and more complex methodologies, the RID method for the generation of random insertions and deletions and the ADO method for recombination have not been patented. The MAX method, useful in the generation of unbiased libraries containing multiple randomized positions, has been protected (WO00/15777, held by Aston University and Amersham Pharmacia Biotech, and WO 03/106679, held by Aston University). The use of trinucleotide (and dinucleotide) phosphoramidites in the synthesis of mutagenic primers is covered by patents held by Maxygen (US6436675). Diversa also claims a patent on the use of a complete set of primers to mutagenize every position in a gene (US6562594).

CONCLUSION

The vast majority of reported directed evolution experiments use a combination of error-prone PCR and DNA shuffling, sometimes combined with primer-based saturation mutagenesis, to construct the initial library and subsequent libraries for each cycle of selection. These techniques are straightforward and have been successfully applied to the optimization of a range of protein activities including binding, stability and enzyme selectivity. The challenge now lies in pushing back the boundaries of what can be achieved using directed evolution. Tackling these challenges may require the construction of new types of libraries and more efficiently constructed libraries of types already available. A large toolkit of methods has recently become available that makes possible the construction of these novel and highly efficient libraries. These methods are necessarily more complex than error-prone PCR, DNA shuffling and oligonucleotide-based mutagenesis, and are therefore unlikely to be the first choice of the general user. However, they are likely to come into their own where the easy methods have failed. They will also be valuable for those groups working to develop optimized and rational approaches to directed evolution. It is not, as yet, clear which of these new techniques will be most valuable or most popular. However, the ability to generate and combine such a wide range of sequence diversity is an important staging post of the route to developing the full promise of the technology of directed evolution.

REFERENCES

1. Farinas,E.T., Bulter,T. and Arnold,F.H. (2001) Directed enzyme evolution. *Curr. Opin. Biotechnol.*, **12**, 545–551.
2. Reetz,M.T., Rentzsch,M., Pletsch,A. and Maywald,M. (2002) Towards the directed evolution of hybrid catalysts. *Chimia*, **56**, 721–723.
3. Taylor,S.V., Kast,P. and Hilvert,D. (2001) Investigating and engineering enzymes by genetic selection. *Angew. Chem. Int. Ed. Engl.*, **40**, 3310–3335.
4. Waldo,G.S. (2003) Genetic screens and directed evolution for protein solubility. *Curr. Opin. Chem. Biol.*, **7**, 33–38.
5. Tao,H. and Cornish,V.W. (2002) Milestones in directed enzyme evolution. *Curr. Opin. Chem. Biol.*, **6**, 858–864.
6. Lin,H. and Cornish,V.W. (2002) Screening and selection methods for large-scale analysis of protein function. *Angew. Chem. Int. Ed. Engl.*, **41**, 4402–4425.
7. Arnold,F.H. and Georgiou,G. (eds) (2003) *Directed Enzyme Evolution: Screening and Selection Methods*. Humana Press, Clifton, NJ.
8. Arnold,F.H. and Georgiou,G. (eds) (2003) *Directed Evolution Library Creation: Methods and Protocols*. Humana Press, Clifton, NJ.
9. Stemmer,W.P.C. (1993) DNA shuffling by random fragmentation and reassembly: *In vitro* recombination for molecular evolution. *Proc. Natl Acad. Sci. USA*, **91**, 10747–10751.
10. Stemmer,W.P.C. (1994) Rapid evolution of a protein *in vitro* by DNA shuffling. *Nature*, **370**, 389–391.
11. Zhao,H., Giver,L., Shao,Z., Affholter,J. and Arnold,F. (1998) Molecular evolution by staggered extension process (StEP) *in vitro* recombination. *Nat. Biotechnol.*, **16**, 258–261.
12. Ostermeier,M., Shim,J.H. and Benkovic,S.J. (1999) A combinatorial approach to hybrid enzymes independent of DNA homology. *Nat. Biotechnol.*, **17**, 1205–1209.
13. Bornscheuer,U.T., Altenbuchner,J. and Meyer,H.H. (1999) Directed evolution of an esterase: screening of enzyme libraries based on pH-indicators and a growth assay. *Bioorg. Med. Chem.*, **7**, 2169–2173.
14. Alexeeva,M., Enright,A., Dawson,M.J., Mahmoudian,M. and Turner,N.J. (2002) Deracemization of alpha-methylbenzylamine using an enzyme obtained by *in vitro* evolution. *Angew. Chem. Int. Ed. Engl.*, **41**, 3177–3180.
15. Nguyen,A.W. and Daugherty,P.S. (2003) Production of randomly mutated plasmid libraries using mutator strains. *Methods Mol. Biol.*, **231**, 39–44.
16. Cadwell,R.C. and Joyce,G.F. (1994) Mutagenic PCR. *PCR Methods Appl.*, **3**, S136–S40.
17. Cirino,P.C., Mayer,K.M. and Umeno,D. (2003) Generating mutant libraries using error-prone PCR. *Methods Mol. Biol.*, **231**, 3–9.
18. Matsumura,I. and Ellington,A.D. (2001) *In vitro* evolution of a beta-glucuronidase into a beta-galactosidase proceeds through non-specific intermediates. *J. Mol. Biol.*, **305**, 331–339.
19. Lingen,B., Grotzinger,J., Kolter,D., Kula,M.R. and Pohl,M. (2002) Improving the carboligase activity of benzoylformate decarboxylase from *Pseudomonas putida* by a combination of directed evolution and site-directed mutagenesis. *Protein Eng.*, **15**, 585–593.
20. Bessler,C., Schmitt,J., Maurer,K.H. and Schmid,R.D. (2003) Directed evolution of a bacterial alpha-amylase: toward enhanced pH-performance and higher specific activity. *Protein Sci.*, **12**, 2141–2149.
21. Zaccolo,M., Williams,D.M., Brown,D.M. and Gheradi,E. (1996) An approach to random mutagenesis of DNA using mixtures of triphosphate derivatives of nucleoside analogues. *J. Mol. Biol.*, **255**, 589–603.
22. Zaccolo,M. and Gherardi,E. (1999) The effect of high-frequency random mutagenesis on *in vitro* protein evolution: a study on TEM-1 β-lactamase. *J. Mol. Biol.*, **285**, 775–783.
23. Patrick,W.M., Firth,A.E. and Blackburn,J.M. (2003) User-friendly algorithms for estimating completeness and diversity in randomized protein-encoding libraries. *Protein Eng.*, **16**, 451–457.
24. Rowe,L.A., Geddie,M.L., Alexander,O.B. and Matsumura,I. (2003) A comparison of directed evolution approaches using the beta-glucuronidase model system. *J. Mol. Biol.*, **332**, 851–860.
25. Miyazaki,K. and Arnold,F.H. (1999) Exploring nonnatural evolutionary pathways by saturation mutagenesis: rapid improvement of protein function. *J. Mol. Evol.*, **49**, 716–720.
26. Murakami,H., Hohsaka,T. and Sisido,M. (2002) Random insertion and deletion of arbitrary number of bases for codon-based random mutation of DNAs. *Nat. Biotechnol.*, **20**, 76–81.
27. Murakami,H., Hohsaka,T. and Sisido,M. (2003) Random insertion and deletion mutagenesis. *Methods Mol. Biol.*, **231**, 53–64.
28. Hayes,F. and Hallet,B. (2000) Pentapeptide scanning mutagenesis: encouraging old proteins to execute unusual tricks. *Trends Microbiol.*, **8**, 571–577.
29. Pikkemaat,M.G. and Janssen,D.B. (2002) Generating segmental mutations in haloalkane dehalogenase: a novel part in the directed evolution toolbox. *Nucleic Acids Res.*, **30**, e35.
30. Ward,B. and Juehne,T. (1998) Combinatorial library diversity: probability assessment of library populations. *Nucleic Acids Res.*, **26**, 879–886.
31. Palfrey,D., Picardo,M. and Hine,A.V. (2000) A new randomization assay reveals unexpected elements of sequence bias in model ‘randomized’ gene libraries: implications for biopanning. *Gene*, **251**, 91–99.
32. Virnekaes,B., Ge,L., Plueckthun,A. Schneider,K.C., Wellnhofer,G. and Moroney,S.E. (1994) Trinucleotide phosphoramidites: ideal reagents for the synthesis of mixed oligonucleotides for random mutagenesis. *Nucleic Acids Res.*, **22**, 5600–5607.
33. Zehl,A., Starke,A., Cech,D., Hartsch,T., Merkl,R. and Fritz,H.J. (1996) Efficient and flexible access to fully protected trinucleotides suitable for DNA synthesis by automated phosphoramidite chemistry. *Chem. Commun. Chem. Soc.*, 2677–2678.
34. Kayushin,A., Korosteleva,M., Miroshnikov,A., Zubov,D., Kosch,W. and Piel,N. (1996) A convenient approach to the synthesis of trinucleotide phosphoramidites–synthons for the generation of oligonucleotide/peptide libraries. *Nucleic Acids Res.*, **24**, 3748–3755.
35. Kayushin,A., Korosteleva,M. and Miroshnikov,A. (2000) Large-scale solid-phase preparation of 3′-unprotected trinucleotide phosphotriesters–precursors for synthesis of trinucleotide phosphoramidites. *Nucleosides Nucleotides Nucleic Acids*, **19**, 1967–1976.
36. Gaytan,P., Yanez,J., Sanchez,F. and Soberon,X. (2001) Orthogonal combinatorial mutagenesis: a codon-level combinatorial mutagenesis method useful for low multiplicity and amino acid-scanning protocols. *Nucleic Acids Res.*, **29**, e9.
37. Gaytan,P., Yanez,J., Sanchez,F., Mackie,H. and Soberon,X. (1998) Combination of DMT-mononucleotide and Fmoc-trinucleotide phosphoramidites in oligonucleotide synthesis affords an automatable codon-level mutagenesis method. *Chem. Biol.*, **5**, 519–527.
38. Lahr,S.J., Broadwater,A., Carter,C.W., Jr, Collier,M.L., Hensley,L., Waldner,J.C., Pielak,G.J. and Edgell,M.H. (1999) Patterned library analysis: a method for the quantitative assessment of hypotheses concerning the determinants of protein structure. *Proc. Natl Acad. Sci. USA*, **96**, 14860–14865.
39. Glaser,S.M., Yelton,D.E. and Huse,W.D. (1992) Antibody engineering by codon-based mutagenesis in a filamentous phage vector system. *J. Immunol.*, **149**, 3903–3913.
40. Liebeton,K., Zonta,A., Schimossek,K., Nardini,M., Lang,D., Dijkstra,B., Reetz,M. and Jaeger,K. (2000) Directed evolution of an enantioselective lipase. *Chem. Biol.*, **7**, 709–718.

41. Sakamoto,T., Joern,J.M., Arisawa,A. and Arnold,F.H. (2001) Laboratory evolution of toluene dioxygenase to accept 4-picoline as a substrate. *Appl. Environ. Microbiol.*, **67**, 3882–3887.
42. Horsman,G.P., Liu,A.M., Henke,E., Bornscheuer,U.T. and Kazlauskas,R.J. (2003) Mutations in distant residues moderately increase the enantioselectivity of *Pseudomonas fluorescens* esterase towards methyl 3bromo-2-methylpropanoate and ethyl 3phenylbutyrate. *Chemistry*, **9**, 1933–1939.
43. Leemhuis,H., Rozeboom,H.J., Wilbrink,M., Euverink,G.J., Dijkstra,B.W. and Dijkhuizen,L. (2003) Conversion of cyclodextrin glycosyltransferase into a starch hydrolase by directed evolution: the role of alanine 230 in acceptor subsite +1. *Biochemistry*, **42**, 7518–7526.
44. Wada,M., Hsu,C.C., Franke,D., Mitchell,M., Heine,A., Wilson,I. and Wong,C.H. (2003) Directed evolution of *N*-acetylneuraminic acid aldolase to catalyze enantiomeric aldol reactions. *Bioorg. Med. Chem.*, **11**, 2091–2098.
45. Juillerat,A., Gronemeyer,T., Keppler,A., Gendreizig,S., Pick,H., Vogel,H. and Johnsson,K. (2003) Directed evolution of *O*(6)-alkylguanine-DNA alkyltransferase for efficient labeling of fusion proteins with small molecules *in vivo*. *Chem. Biol.*, **10**, 313–317.
46. Sio,C.F., Riemens,A.M., van der Laan,J.M., Verhaert,R.M. and Quax,W.J. (2002) Directed evolution of a glutaryl acylase into an adipyl acylase. *Eur. J. Biochem.*, **269**, 4495–4504.
47. Georgescu,R., Bandara,G. and Sun,L. (2003) Saturation mutagenesis. *Methods Mol. Biol.*, **231**, 75–83.
48. Miyazaki,K. and Takenouchi,M. (2002) Creating random mutagenesis libraries using megaprimer PCR of whole plasmid. *Biotechniques*, **33**, 1033–1034, 1036–1038.
49. Miyazaki,K. (2003) Creating random mutagenesis libraries by megaprimer PCR of whole plasmid (MEGAWHOP). *Methods Mol. Biol.*, **231**, 23–28.
50. Zha,D., Eipper,A. and Reetz,M.T. (2003) Assembly of designed oligonucleotides as an efficient method for gene recombination: a new tool in directed evolution. *Chembiochemistry*, **4**, 34–39.
51. Ness,J.E., Kim,S., Gottman,A., Pak,R., Krebber,A., Borchert,T.V., Govindarajan,S., Mundorff,E.C. and Minshull,J. (2002) Synthetic shuffling expands functional protein diversity by allowing amino acids to recombine independently. *Nat. Biotechnol.*, **20**, 1251–1255.
52. Hogrefe,H.H., Cline,J., Youngblood,G.L. and Allen,R.M. (2002) Creating randomized amino acid libraries with the QuikChange Multi Site-Directed Mutagenesis Kit. *Biotechniques*, **33**, 1158–1160, 1162, 1164–1165.
53. Hughes,M.D., Nagel,D.A., Santos,A.F., Sutherland,A.J. and Hine,A.V. (2003) Removing the redundancy from randomised gene libraries. *J. Mol. Biol.*, **331**, 973–979.
54. Joern,J.M. (2003) DNA shuffling. *Methods Mol. Biol.*, **231**, 85–89.
55. Aguinaldo,A.M. and Arnold,F.H. (2003) Staggered extension process (StEP) *in vitro* recombination. *Methods Mol. Biol.*, **231**, 105–110.
56. Coco,W.M., Levinson,W.E., Crist,M.J., Hektor,H.J., Darzins,A., Pienkos,P.T., Squires,C.H. and Monticello,D.J. (2001) DNA shuffling method for generating highly recombined genes and evolved enzymes. *Nat. Biotechnol.*, **19**, 354–359.
57. Coco,W.M. (2003) RACHITT: Gene family shuffling by Random Chimeragenesis on Transient Templates. *Methods Mol. Biol.*, **231**, 111–127.
58. Lutz,S., Ostermeier,M. and Benkovic,S.J. (2001) Rapid generation of incremental truncation libraries for protein engineering using alpha-phosphothioate nucleotides. *Nucleic Acids Res.*, **29**, e16.
59. Ostermeier,M. and Lutz,S. (2003) The creation of ITCHY hybrid protein libraries. *Methods Mol. Biol.*, **231**, 129–141.
60. Lutz,S. and Ostermeier,M. (2003) Preparation of SCRATCHY hybrid protein libraries: size- and in-frame selection of nucleic acid sequences. *Methods Mol. Biol.*, **231**, 143–151.
61. Dixon,D.P., McEwen,A.G., Lapthorn,A.J. and Edwards,R. (2003) Forced evolution of a herbicide detoxifying glutathione transferase. *J. Biol. Chem.*, **278**, 23930–23935.
62. Baik,S.H., Ide,T., Yoshida,H., Kagami,O. and Harayama,S. (2003) Significantly enhanced stability of glucose dehydrogenase by directed evolution. *Appl. Microbiol. Biotechnol.*, **61**, 329–335.
63. Miyazaki,K. (2002) Random DNA fragmentation with endonuclease V: application to DNA shuffling. *Nucleic Acids Res.*, **30**, e139.
64. Glieder,A., Farinas,E.T. and Arnold,F.H. (2002) Laboratory evolution of a soluble, self-sufficient, highly active alkane hydroxylase. *Nat. Biotechnol.*, **20**, 1135–1139.
65. Sun,L., Petrounia,I.P., Yagasaki,M., Bandara,G. and Arnold,F.H. (2001) Expression and stabilization of galactose oxidase in *Escherichia coli* by directed evolution. *Protein Eng.*, **14**, 699–704.
66. Moore,G.L., Maranas,C.D., Lutz,S. and Benkovic,S.J. (2001) Predicting crossover generation in DNA shuffling. *Proc. Natl Acad. Sci. USA*, **98**, 3226–3231.
67. Moore,G.L. and Maranas,C.D. (2002) eCodonOpt: a systematic computational framework for optimizing codon usage in directed evolution experiments. *Nucleic Acids Res.*, **30**, 2407–2416.
68. Joern,J.M., Meinhold,P. and Arnold,F.H. (2002) Analysis of shuffled gene libraries. *J. Mol. Biol.*, **316**, 643–656.
69. Kikuchi,M., Ohnishi,K. and Harayama,S. (1999) Novel family shuffling methods for the *in vitro* evolution of enzymes. *Gene*, **236**, 159–167.
70. Kikuchi,M., Ohnishi,K. and Harayama,S. (2000) An effective family shuffling method using single-stranded DNA. *Gene*, **243**, 133–137.
71. Zha,W., Zhu,T. and Zhao,H. (2003) Family shuffling with single-stranded DNA. *Methods Mol. Biol.*, **231**, 91–97.
72. Gibbs,M.D., Nevalainen,K.M. and Bergquist,P.L. (2001) Degenerate oligonucleotide gene shuffling (DOGS): a method for enhancing the frequency of recombination with family shuffling. *Gene*, **271**, 13–20.
73. Lutz,S., Fast,W. and Benkovic,S.J. (2002) A universal, vector-based system for nucleic acid reading-frame selection. *Protein Eng.*, **15**, 1025–1030.
74. Lutz,S., Ostermeier,M., Moore,G.L., Maranas,C.D. and Benkovic,S.J. (2001) Creating multiple-crossover DNA libraries independent of sequence identity. *Proc. Natl Acad. Sci. USA*, **98**, 11248–11253.
75. Kawarasaki,Y., Griswold,K.E., Stevenson,J.D., Selzer,T., Benkovic,S.J., Iverson,B.L. and Georgiou,G. (2003) Enhanced crossover SCRATCHY: construction and high-throughput screening of a combinatorial library containing multiple non-homologous crossovers. *Nucleic Acids Res.*, **31**, e126.
76. O'Maille,P.E., Bakhtina,M. and Tsai,M.D. (2002) Structure-based combinatorial protein engineering (SCOPE). *J. Mol. Biol.*, **321**, 677–691.
77. Hiraga,K. and Arnold,F.H. (2003) General method for sequence-independent site-directed chimeragenesis. *J. Mol. Biol.*, **330**, 287–296.
78. Voigt,C.A., Martinez,C., Wang,Z.G., Mayo,S.L. and Arnold,F.H. (2002) Protein building blocks preserved by recombination. *Nat. Struct. Biol.*, **9**, 553–558.
79. Meyer,M.M., Silberg,J.J., Voigt,C.A., Endelman,J.B., Mayo,S.L., Wang,Z.G. and Arnold,F.H. (2003) Library analysis of SCHEMA-guided protein recombination. *Protein Sci.*, **12**, 1686–1693.
80. Moore,J.C., Jin,H.-M., Kuchner,O. and Arnold,F.H. (1997) Strategies for the *in vitro* evolution of protein function: enzyme evolution by random recombination of improved sequences. *J. Mol. Biol.*, **272**, 336–347.