Pergamon

Bioorganic & Medicinal Chemistry 7 (1999) 2139–2144

BIOORGANIC & MEDICINAL CHEMISTRY

Incremental Truncation as a Strategy in the Engineering of Novel Biocatalysts

Marc Ostermeier, Andrew E. Nixon* and Stephen J. Benkovic†

Department of Chemistry, The Pennsylvania State University, University Park, PA 16802, USA

Received 15 October 1998

Abstract—The application and success of combinatorial approaches to protein engineering problems have increased dramatically. However, current directed evolution strategies lack a combinatorial methodology for creating libraries of hybrid enzymes which lack high homology or for creating libraries of highly homologous genes with fusions at regions of non-identity. To create such hybrid enzyme libraries, we have developed a series of combinatorial approaches that utilize the incremental truncation of genes, gene fragments or gene libraries. For incremental truncation, Exonuclease III is used to create a library of all possible single base-pair deletions of a given piece of DNA. Incremental truncation libraries (ITLs) have applications in protein engineering as well as protein folding, enzyme evolution, and the chemical synthesis of proteins. In addition, we are developing a methodology of DNA shuffling which is independent of DNA sequence homology.

Introduction

Protein mutagenesis has long been used as a tool for structure/function studies of proteins. With the advent of modern DNA manipulation techniques and advancements in protein structure determination, large numbers of protein sequences and structures are available which can be sorted into groups or superfamilies based on structural similarity.[1] Such alignments show that enzymes that are structurally similar often catalyze similar reactions and have active sites with shared amino acids. This allows identification of side chain residues important in binding and catalysis and their modification to yield enzymes with altered properties.

Such structure-based rational approaches to protein engineering have, through introduction of point mutations, exchange of secondary structural elements, and exchange of whole domains or subunits, given rise to enzymes that have altered substrate specificities, catalytic properties and oligomeric states.[2] Although few protein-engineering failures have been published, the difficulty in rationally engineering an enzyme to have a specific function is widely appreciated. Any alteration introduced into a wildtype protein can, in ways that often cannot be predicted, disrupt the fine balance that nature has achieved, consequently giving rise to proteins that are unstable, fail to fold properly and lack catalytic activity. As a result of the difficulties encountered using strict rational design approaches, there is an increasing trend towards the use of molecular biology strategies that mimic evolutionary processes. These strategies are known as 'directed evolution'.[3]

Most directed evolution strategies incorporate some method of introducing random mutations into a gene followed by screening or selection for a desired property. The cycle is then repeated several times until the desired property is achieved or until further cycling produces no improvement in the desired property. Early methodologies utilized point mutations generated by error prone PCR, chemical mutagenesis or mutator strains of *E. coli*. This type of approach is something akin to an asexual evolutionary process with non-beneficial and beneficial mutations becoming fixed. The advent of DNA shuffling,[4,5] which more closely approximates the natural evolutionary process, has had an enormous impact on directed evolution. In this technique, parental genes are fragmented and subsequently reassembled by PCR to reconstitute the full-length genes. During this reassembly process, novel combinations of the parental genes arise along with new point mutations. The result of DNA shuffling is a large library of mutant genes from which acquisition of a

Key words: Incremental truncation; hybrid enzymes; directed evolution; protein engineering.

† Corresponding author. Tel.: +1-814-865-2882 fax: +1-814-865-2973; e-mail: sjb1@psu.edu

* Present address: Dyax Corporation, One Kendall Square, Building 600, Cambridge, MA 02139, USA.

0968-0896/99/$ - see front matter
PII: S0968-0896(99)00143-1

desired function is selected for using an appropriate selection or screening system.

Such strategies have been particularly successful in achieving improvements in thermostability,[6] altering substrate specificity and improving activity in organic solvents (for a review, see Kuchner and Arnold[3]). However, since directed evolution is a stepwise process, only relatively small steps in sequence space can occur. Thus, the utility of current directed evolution methodologies to evolve novel catalytic sites, which presumably require large excursions in sequence space, is limited. While it is true that DNA shuffling of families of genes with DNA homology can create hybrid enzymes with new properties,[7] such molecular breeding is only feasible for genes with high genetic homology and, for this reason, is unlikely to evolve an entirely novel function. It is important to realize that the primary rationale for success in the shuffling of families of genes is the similarity of the three-dimensional structures of the proteins they encode, not the degree of DNA homology. Indeed, it is an interesting question whether successful directed evolution on homologous families might be equally or better served by the creation of genes with crossovers between family members at regions of little or no genetic homology. However, current DNA shuffling methodologies only produce crossovers within regions of high homology and within significant stretches of identity. Furthermore, crossovers are biased towards those regions of highest identity.

The increasing numbers of protein structures available and the study of enzyme structural families[1] has shown that many enzymes with little or no DNA homology can have high protein structural homology. Constructing hybrids of such structural homologues may well be an important strategy for engineering of novel activities; however, no combinatorial approach for the construction of such hybrids has been reported. Our interest in developing such a combinatorial approach stemmed from our work in the inter-conversion of formyltetrahydrofolate utilizing enzymes. We recently demonstrated the feasibility of creating active hybrids between such proteins by engineering a functional hybrid enzyme through fusing domains from two enzymes that overall had very little genetic homology.[8] We made discreet domain fusions between the glycinamide ribonucleotide (GAR) binding domain of the *E. coli purN* gene (GAR transformylase) and the formyl-tetrahydrofolate binding and catalytic domain of the *E. coli purU* gene (formyltetrahydrofolate hydrolase). Although we created a hybrid enzyme that had the desired property (GAR transformylase activity), this activity was low prompting us to search for a combinatorial approach to this engineering problem. Using GAR transformylase as a model system we have developed a number of potential solutions to the problem which are detailed in the following sections.

Knowing where to make the fusions is a central problem in the creation of such hybrids. Since current methodologies for genes lacking high homology were limited to 'try it and see if it works,' we developed a combinatorial approach to this problem termed incremental truncation. Through incremental truncation we can create fusion libraries of many (or all) different combinations of lengths of two genes. This approach, described herein, is thus a combinatorial solution to the questions 'where can enzymes or enzyme fragments be fused to produce active hybrids' as well as 'where are the points at which an enzyme can be bisected'. In addition, we outline a method that should circumvent homology limitations to DNA shuffling by allowing shuffling of genes independent of sequence homology.

Incremental Truncation

For the average size gene, the separate construction of all possible one-codon truncations would require the assembly of hundreds of plasmids, a labor intensive and time consuming task. Incremental truncation of DNA, on the other hand, allows the construction of a library containing all possible truncations of a gene, gene fragment or DNA library in a single experiment (Fig. 1).

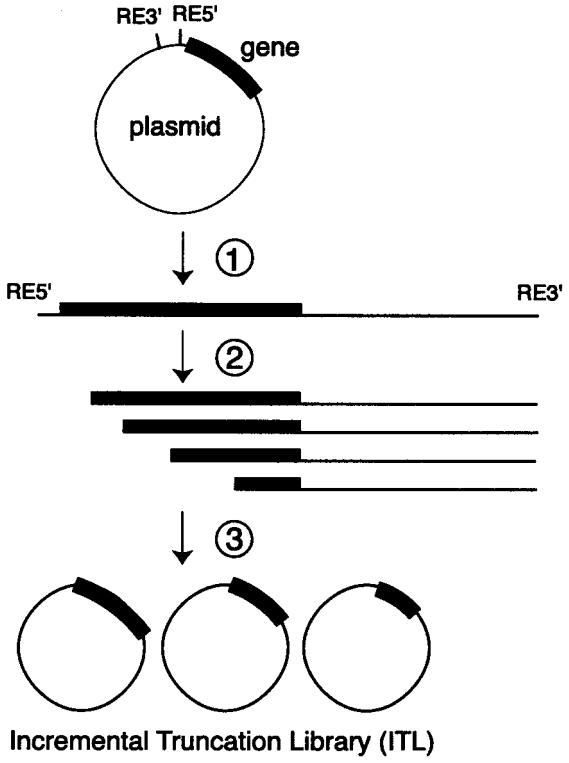


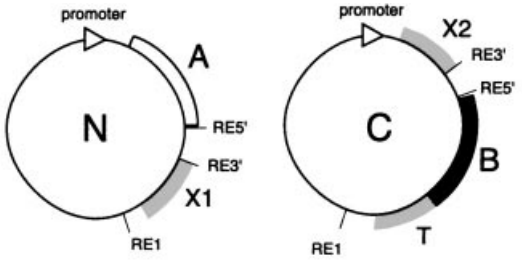
Figure 1. Generalized incremental truncation. Incremental truncation is performed on a linear piece of DNA containing a gene, gene fragment or DNA library has one end protected from digestion and the other end susceptible to digestion. This is easily accomplished, for example, by (1) digestion of plasmid DNA with two restriction enzymes: one that produces a 3′ overhang (RE3′; which is resistant to Exo III digestion) and the other which produces a 5′ overhang (RE5′; which is susceptible to Exo III digestion). (2) Digestion with Exonuclease III proceeds under conditions such that the digestion rate is slow enough that the removal of aliquots at frequent intervals results in a DNA library with every one codon (or base pair) deletion. (3) The ends of the DNA can be blunted by treatment with S1 nuclease and Klenow so that unimolecular ligation results in the desired incremental truncation library. For some applications, additional DNA manipulations are required before recircularizing the vector.

Incremental truncation is achieved by utilizing the slow, directional, controlled digestion of DNA. During this digestion, small aliquots are frequently removed and the digestion quenched. Thus by taking multiple samples over a given time period we can create a library of all possible single base-pair deletions of a given piece of DNA.

We have been using Exonuclease III (Exo III) which exhibits such properties. Exo III has been previously shown to be useful in the creation of large truncations of linear DNA[9] and for techniques in the sequencing of large genes.[10] The digestion rate of Exo III at 37°C (~500 bases/min)[11] is much too fast for purposes of incremental truncation where every one-codon deletion is desired. However, the digestion rate of the exonuclease can be affected by a variety of methods such as lowering the incubation temperature, altering the digestion buffer composition, inclusion of a nuclease inhibitor or lowering the ratio of enzyme to DNA.

Applications

Figure 2 shows the two-vector system utilized for the applications of incremental truncation. We will refer to this figure in the following examples of applications of the technique.



Application	X1	X2	T
Protein Frag. Compl. with dimer fusion	3-frame stop dimer domain	start codon dimer domain	stop stop
Seamed ITCHY	RE2	RE2	stop
Seamless ITCHY	--------	--------	stop
Trans Inteins	intein (I_N)	intein (I_C)	stop
SCRATCHY	--------	--------	stop or in-frame reporter

Figure 2. Vectors for incremental truncation applications. Plasmids N and C are two compatible vectors with origins of replication belonging to different compatibility groups and bearing genes coding for different antibiotic resistances. For some applications, it is advantageous that the two vectors are phagemids (e.g. that they also contain a phage origin of replication) for packaging into phage particles. The gene, gene fragment or gene library to be truncated (A and B) is positioned downstream from a promoter. The identity of other features of the vectors is shown in the table and depends on the application. The X1 and X2 segments (when used) represent the piece of DNA that the ITLs of A or B are fused to in the unimolecular ligation step. The use of 'RE' designates a unique restriction enzyme site. RE5′ and RE3′ indicate that digestion with the restriction enzyme produces a 5′ or 3′ overhang respectively. A 5′ overhang is susceptible to Exo III digestion whereas a 3′ overhang is not.

Protein fragment complementation

Reverse evolution (monomer to heterodimer conversion). One theory of evolution of enzymes has catalytic function arising from the interaction of protein fragments, which eventually become condensed to a single gene product. The reverse of this process would be to convert an existing monomeric enzyme into its functional heterodimer, also referred to as protein fragment complementation. The use of incremental truncation libraries (ITL), along with a suitable screen or selection, allows one to determine all points in the backbone polypeptide chain that can be broken, yet the two resulting fragments still retain the ability to fold and associate into an active heterodimer.

For example, the gene for a protein (P) is first divided into two non-active, overlapping fragments: A (containing the N-terminus of P) and B (containing the C-terminus of P) which are cloned into vectors suitable for incremental truncation (Fig. 2). For this experiment, X1 is a series of stop codons in all three frames, X2 is the start codon ATG, and T is a stop codon in frame with B. After linearizing the vector with restriction enzymes RE3′ and RE5′ and subsequent incremental truncation, unimolecular ligation results in the 3′ end of the ITL of A being fused to a series of stop codons in all three frames and the 5′ end of the ITL library of B being fused to a start codon. Although two-thirds of the ITL library of A will have 1–3 foreign amino acids on the end and two-thirds of the ITL library of B will be out of frame, one-third of each library will be in-frame and not code for any foreign amino acids. Crossing the ITL libraries of A and B, obtained by transforming both libraries into the same *E. coli* cells, will have each cell producing a different combination of a N-terminal fragment and a C-terminal fragment of P. Active members of this crossed ITL library can be identified by screening or selection. We have recently successfully applied this methodology to *E. coli* glycinamide ribonucleotide transformylase.[12]

The identification of points for functional bisection of an enzyme has applications in enzyme evolution and protein folding, since such bisection points potentially identify ancestral fusion points as well as independent folding units. Such dissection of enzymes into smaller fragments also subverts impediments in the chemical synthesis of proteins: proteins too large to be chemically synthesized as a monomer can be synthesized as fragments, thus allowing the introduction of unique side chain functions. Moreover, the identification of functional structural motifs, subdomains, or domains will facilitate the construction of hybrid enzymes and the creation of enzymes with novel activities. The construction of crossed ITLs of protein structural homologues illustrates one combinatorial approach to domain swapping made feasible by incremental truncation.

Use of a dimerization motif. Bisection of a protein in the manner described above could potentially lead to problems with association of the two fragments, particularly between structural homologues. The two protein

fragments may be unable or have little tendency to associate. The addition of tight binding dimerization domains circumvents this issue. This has recently been described for dihydrofolate reductase (DHFR).[13] DHFR was bisected and fused to dimerization domains that resulted in association of the protein fragments and restoration of catalytic activity as shown by an in vivo activity screen. The two fragments of DHFR without the dimerization domains demonstrated no activity in the same screen.

Such facilitated association of protein fragments allows for the creation of structural-homologue heterodimers. One could imagine, for example, creating hybrid enzymes such that an ITL of the catalytic machinery of one enzyme (A) is fused to one dimerization domain (X1) and a ITL of a substrate binding domain (B) is fused to a second dimerization domain (X2). Such A-X1 and B-X2 fusion libraries would then be crossed into *E. coli* cells as above and the functional association of the two subunits A and B would be facilitated by the dimerization of X1 and X2. Although it is not absolutely necessary, X1 and X2 preferably should be different (e.g. they form a heterodimer) so as to avoid homodimerization of A-X1:X1-A and B-X2:X2-B in lieu of heterodimerization (A-X1:X2-B). Structures such as anti-parallel helixes or parallel helix-turn-helixes may be preferable to avoid the necessity of long linkers. This type of approach would allow scanning for novel activities across families of proteins in one experiment, as A and B need not be discrete genes but could be a library of family members.

One advantage to this method would be the ability to access very large libraries ($\sim 10^{11}$) if vectors N and C are phagemids and can be packaged into phage particles. Since phage infection is a very efficient method of introducing vectors into *E. coli*, the library size is limited primarily by the number of *E. coli* cells in the culture. For example, if each individual A-X1 and B-X2 library has a library size of 2×10^6, then the crossed library of these two has a maximum library size of 4×10^{12}. If a liter of 10^{11} *E. coli* cells is infected with phagemid containing each of the ITL-dimer libraries, and 30% of the cells become infected with both vectors, then the crossed library size is 3×10^{10}. Although the ability to use selection on such large libraries can be problematic, such methodology still makes facile the creation of smaller, manageable libraries.

Hybrid enzymes

For domain swapping, it can be difficult to predict exactly which fusion-points will produce an active hybrid enzyme. The use of incremental truncation in the creation of hybrid enzyme libraries solves this problem by a stochastic method. This method does not require any homology on the DNA level or any knowledge of the structure of either enzyme. Theoretically, all possible combinations of two genes can be created and, with the use of a suitable screen or selection, active hybrids can be identified. Three variations of this methodology, which we have termed Incremental Truncation for the Creation of Hybrid enzYmes (ITCHY), are outlined below.

Seamed ITCHY libraries. In this case X1 and X2 are identical restriction sites (RE2) and T is a stop codon in frame with B. The individual ITLs of A and B are constructed as in protein fragment complementation above (e.g. linearization of the plasmid DNA with RE3′ and RE5′ followed by incremental truncation and recircularization). Next, the ITL of B is cloned into plasmid N bearing the ITL of A between the RE2 and RE1 sites using identical restriction sites on plasmid C. The resulting ITCHY library is seamed since it will contain the restriction enzyme site RE2 at the junction of the two gene fragments and thus code for foreign amino acids. One third of the library will have B in frame with A. If a linker is desired between the two genes, it can be included in either X1 or X2 such that it is between RE2 and the truncated gene.

Seamless ITCHY libraries. It is possible to avoid the seam at the interface between the two genes. This method, however, depends on the cloning of a fragments with one blunt end, so the library size may be less than in a seamed ITCHY. Incremental truncation proceeds as in protein fragment complementation above, except that before the vector is recircularized, plasmids N and C are digested with RE1 (Fig. 3). Vector N (containing the ITL of A) is isolated away from fragment X1 and the ITL of B is isolated from the rest of the vector C. The ITL of B is then ligated into vector N (containing the ITL of A) by a sticky/blunt ligation. The blunt end ligation is what produces the seamless fusion of the two genes. The sticky end ligation (at RE1) provides directionality and improved cloning efficiency (compared to a blunt end ligation). As in a seamed ITCHY, one-third of the library will have B in frame with A. Unlike a seamed ITCHY, a seamless ITCHY is

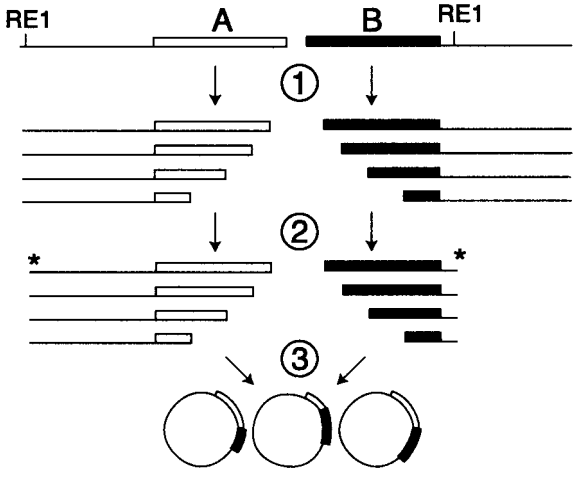


Figure 3. Seamless ITCHY. Linearized versions of vectors N and C from Fig. 2 are prepared by digestion with RE3′ and RE5′. (1) Incremental truncation proceeds as in Figure 1. (2) The linear ITLs are digested with RE1 and the indicated fragments are isolated. (3) Ligation of the fragments containing the ITL of B into the vector containing the ITL of A proceeds by a sticky end ligation at the asterix and a blunt end ligation between the truncated genes.

not easily amenable to linker incorporation. Our initial efforts to create seamless ITCHY libraries have created libraries of up to 129,000 fusions (43,000 in-frame fusions) between the incremental truncation libraries of two genes. This library size is the minimum necessary to theoretically have all possible fusions between two ITLs whose members contain between 0 and 359 deleted bases.

Use of trans-inteins. Protein splicing is a post-translational event involving precise excision of an intein fragment from precursor protein sequences. While most inteins described to date have been *cis*-inteins (encoded on one polypeptide), recently engineered[14–18] and naturally occurring[19] *trans*-inteins have been described. The ability of *trans*-inteins to fuse potentially any two polypeptides is well suited for the creation of hybrid enzyme libraries (Fig. 4).

For this application, incremental truncation is performed as in protein fragment complementation above, resulting in a fusion of an ITL of A to one half of the *trans*-intein (I_N) and an ITL of B to the other half of the *trans*-intein (I_C). If desired, a linker could be incorporated so that either A or B or both are fused to a linker after incremental truncation. Both vectors (containing an ITL fused to an intein or linker-intein) could then be introduced into the same cell and hybrid enzymes created in vivo as a result of the intein's activity. All the hybrid enzyme products produced using *trans*-inteins will necessarily have a cysteine residue at the fusion point.

As in the use of dimerization domains for protein fragment complementation above, one advantage in the use of *trans*-inteins is that very large hybrid enzyme libraries are possible. These libraries would theoretically be much larger than those made by genetic fusions above (ITCHY libraries). However, one caveat to this approach is that it is unknown to what extent *trans*-inteins retain activity with different sequences fused to them. That is, it is unknown what fraction of the library will actually produce fused protein fragments.

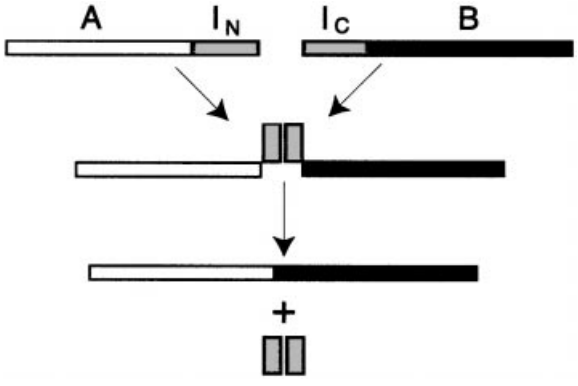


Figure 4. Protein splicing with *trans*-inteins. Fusion proteins of an ITL of A and the N-intein (I_N) and of an ITL of B and the C-intein (I_C) associate in solution via the interaction of I_N and I_C. The intein heterodimer (I_N:I_C) directs the splicing reaction resulting in the joining of A to B with a native peptide bond and the release of I_N:I_C.

Shuffling of non-homologous genes

One requirement for successful DNA shuffling between two or more genes is that they must have a high degree of identity on the DNA level. Furthermore, DNA shuffling crossover points are limited precisely to regions of identity. Many interspecies homologues have sequence homology below that which traditional DNA shuffling can be performed. Furthermore, proteins with little or no sequence identity can have strong structural homology. It is reasonable to assume that shuffling of such genes, for example within a fold superfamily, could result in hybrid enzymes with interesting and useful activities.

We are currently developing a technique for shuffling of genes that does not require any sequence identity. Such shuffling uses, as its starting point, either seamed or (preferably) seamless ITCHY libraries as outlined above. Whereas crossover points between genes in traditional DNA shuffling are defined and confined by the regions of identity, shuffled ITCHY library crossover points are defined by the fusion-points. An ITCHY library theoretically will have all possible crossover points thus there is no limitation on the location of crossover points in the resulting hybrid enzyme library. It follows then, that shuffled ITCHY libraries (which we call SCRATCHY libraries) of genes of high identity will create more diverse libraries than traditional DNA shuffling.

A SCRATCHY library is created by making two ITCHY libraries: one with gene A on the N-terminus (A–B), and one with gene B on the N-terminus (B–A) (Fig. 5). Next, DNA fragments of the A–B and B–A fusions are isolated that are approximately the same size as the original genes. This can be done by gel electrophoresis after restriction enzyme digestion (and judicious location of restriction sites) or after PCR with primers near or just outside the ends of fused genes. This step is to ensure that the pool of DNA to be shuffled contains fusions at points on the primary and three-dimensional structure which are near each other (i.e. we are limiting crossover points to 'intelligent' locations). Thus, SCRATCHY will work best with genes A and B being roughly the same size. This DNA with 'intelligent' crossover points may then be amplified by PCR to obtain enough sample to perform DNA shuffling. The two libraries (A–B and B–A PCR products of about the same size as the original genes) are mixed, digested with DNase I and DNA shuffling is performed.[5]

One potential problem is that the number of hybrids in frame will decrease exponentially with total number of crossovers. For example, the original ITCHY libraries will only have one-third of the hybrids in-frame. A resulting member of the SCRATCHY library with two crossovers will only have a 1 in 9 chance of being completely in-frame; with three crossovers only 1 in 27 will be completely in-frame. This problem is potentially circumvented by a selection on the original ITCHY libraries for those in frame. For example, if gene B is fused in frame to a reporter gene T with a selectable

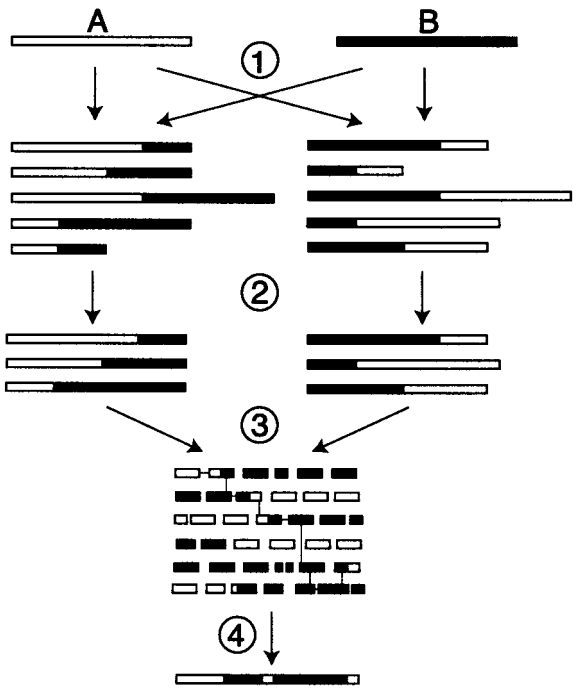


Figure 5. Non-homologous shuffling. (1) Individual A-B and B-A ITCHY libraries are constructed as in Fig. 3. (2) Either through use of outside restriction enzymes or outside PCR primers, those members of the ITCHY libraries which are approximately the same size as the original genes are isolated by gel electrophoresis. (3) These selected ITCHY library members are mixed and fragmented by digestion with DNase I as in traditional DNA shuffling.[5] (4) Reassembly of the random fragments can proceed by template switching which can result in full-length genes with multiple crossovers. One such assembled fragment is shown.

phenotype, then all in frame ITCHY library members with in-frame crossover points can be selected. The T gene need not be a part of the final SCRATCHY library since it can be easily removed in the PCR steps prior to DNase I digestion.

Summary

The current directed evolution toolbox lacks a combinatorial approach to the creation of hybrid enzymes through domain fusions. We have developed several methods that address this need based on the incremental truncation of genes. We are currently testing these strategies using the family of enzymes that utilize formyltetrahydrofolate as a cofactor. We envision that these strategies, particularly when used in concert with existing directed evolution methods, will prove important for the engineering of enzymes with novel activities as well as find applications in addressing problems in protein folding, enzyme evolution and in the chemical synthesis of proteins.

Acknowledgements

This work was supported in part by an NIH grant GM24129 (SJB) and an NIH postdoctoral fellowship GM18560 (MO).

References

1. Babbitt, P. C.; Gerlt, J. A. *J. Biol. Chem.* **1997**, *272*, 30591–30594.
2. Nixon, A. E.; Ostermeier; M. Benkovic, S. J. *Trends Biotechnol.* **1998**, *16*, 258–264.
3. Kuchner, O.; Arnold, F. H. *TIBTECH* **1997**, *15*, 523–530.
4. Stemmer, W. P. C. *Nature* **1994**, *370*, 389–391.
5. Stemmer, W. P. C. *Proc. Natl. Acad. Sci. USA* **1994**, *91*, 10747–10751.
6. Arnold, F. H. *Proc. Natl. Acad. Aci. USA* **1998**, *95*, 2035–2036.
7. Crameri, A.; Raillard, S.; Bermudez, E.; Stemmer, W. P. *Nature* **1998**, *391*, 288–291.
8. Nixon, A. E.; Warren, M. S.; Benkovic, S. J. *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 1069–1073.
9. Maniatis, T.; Fritsch, E. F.; Sambrook, J. *Molecular Cloning: A Laboratory Manual*; Cold Spring Harbor Laboratory: Cold Spring Harbor, NY, 1982.
10. Henikoff, S. *Gene* **1984**, *28*, 351–359.
11. Henikoff, S. *Methods Enzymol.* **1987**, *155*, 156–165.
12. Ostermeier, M.; Nixon, A. E.; Shim, J. H.; Benkovic, S. J. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 3562–3567.
13. Pelletier, J. N.; Campbell-Valois, F.-X.; Michnick, S. W. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 12141–12146.
14. Southworth, M. W.; Adam, E.; Panne, D.; Byer, R.; Kautz, R.; Perler, F. B. *EMBO J.* **1998**, *17*, 918–926.
15. Shingledecker, K.; Jiang, S.-Q.; Paulus, H. *Gene* **1998**, *207*, 187–195.
16. Mills, K. V.; Lew, B. M.; Jiang, S.-Q.; Paulus, H. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 3543–3548.
17. Wu, H.; Xu, M.-Q.; Liu, X.-Q. *Biochim. Biophys. Acta* **1998**, *1387*, 422–432.
18. Lew, B. M.; Mills, K. V.; Paulus, H. *J. Biol. Chem.* **1998**, *273*, 15887–15890.
19. Hu, Z.; Liu, X.-Q. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 9226–9231.