CHAPTER 8

Changing genes: site-directed mutagenesis and protein engineering

Introduction

The generation and characterization of mutants is an essential component of any study on structure–function relationships. Knowledge of the three-dimensional structure of a protein, RNA species, or DNA regulatory element (e.g. a promoter) can provide clues to the way in which they function but proof that the correct mechanism has been elucidated requires the analysis of mutants that have amino acid or nucleotide changes at key residues (see Box 8.2).

Classically, mutants are generated by treating the test organism with chemical or physical agents that modify DNA (mutagens). This method of mutagenesis has been extremely successful, as witnessed by the growth of molecular biology and functional genomics, but suffers from a number of disadvantages. First, any gene in the organism can be mutated and the frequency with which mutants occur in the gene of interest can be very low. This means that selection strategies have to be developed. Second, even when mutants with the desired phenotype are isolated, there is no guarantee that the mutation has occurred in the gene of interest. Third, prior to the development of gene-cloning and sequencing techniques, there was no way of knowing where in the gene the mutation had occurred and whether it arose by a single base change, an insertion of DNA, or a deletion.

As techniques in molecular biology have developed, so that the isolation and study of a single gene is not just possible but routine, so mutagenesis has also been refined. Instead of crudely mutagenizing many cells or organisms and then analyzing many thousands or millions of offspring to isolate a desired mutant, it is now possible to change specifically any given base in a cloned DNA sequence. This technique is known as *site-directed mutagenesis*. It has become a basic tool of gene manipulation, for it simplifies DNA manipulations that in the past required a great deal of ingenuity and hard work, e.g. the creation or elimination of cleavage sites for restriction endonucleases. The importance of site-directed mutagenesis goes beyond gene structure–function relationships for the technique enables mutant proteins with novel properties of value to be created (protein engineering). Such mutant proteins may have only minor changes but it is not uncommon for entire domains to be deleted or new domains added.

Primer extension (the single-primer method) is a simple method for site-directed mutation

The first method of site-directed mutagenesis to be developed was the single-primer method (Gillam *et al.* 1980, Zoller & Smith 1983). As originally described the method involves *in vitro* DNA synthesis with a chemically synthesized oligonucleotide (7–20 nucleotides long) that carries a base mismatch with the complementary sequence. As shown in Fig. 8.1, the method requires that the DNA to be mutated is available in single-stranded form, and cloning the gene in M13-based vectors makes this easy. However, DNA cloned in a plasmid and obtained in duplex form can also be converted to a partially single-stranded molecule that is suitable (Dalbadie-McFarland *et al.* 1982).

The synthetic oligonucleotide primes DNA synthesis and is itself incorporated into the resulting heteroduplex molecule. After transformation of the host *E. coli*, this heteroduplex gives rise to homoduplexes whose sequences are either that of the original wild-type DNA or that containing the mutated base. The frequency with which mutated clones arise, compared with wild-type clones, may be low. In order to pick out mutants, the clones can be screened by nucleic acid hybridization with ^{32}P-labeled oligonucleotide as probe. Under suitable conditions of stringency, i.e. temperature and cation concentration, a positive signal will be obtained only with mutant clones. This allows ready detection of the

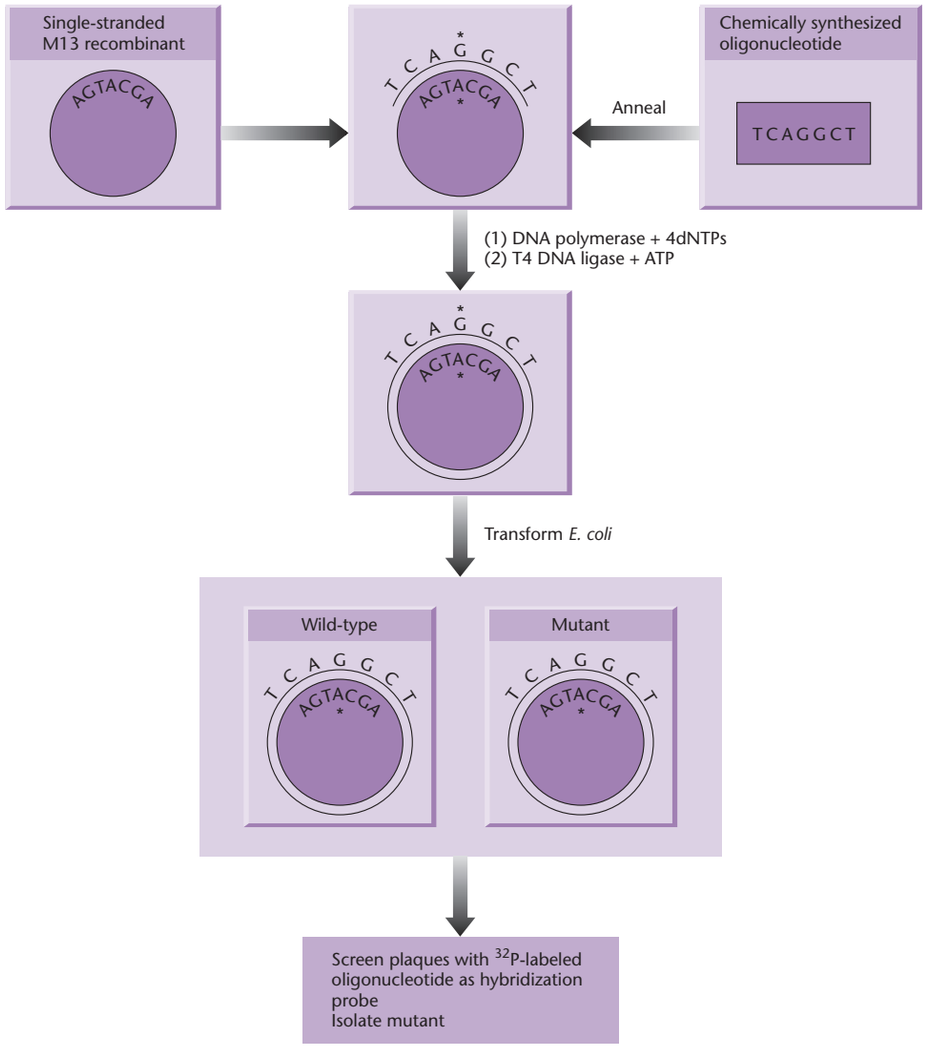


Fig. 8.1
Oligonucleotide-directed mutagenesis. Asterisks indicate mismatched bases. Originally the Klenow fragment of DNA polymerase was used, but now this has been largely replaced with T7 polymerase.

desired mutant (Wallace *et al.* 1981, Traboni *et al.* 1983). It is prudent to check the sequence of the mutant directly by DNA sequencing, in order to check that the procedure has not introduced other adventitious changes. This was a particular necessity with early versions of the technique which made use of *E. coli* DNA polymerase. The more recent use of the high-fidelity DNA polymerases has minimized the problem of extraneous mutations as well as shortening the time for copying the second strand. Also, these polymerases do not "strand-displace" the oligomer, a process which would eliminate the original mutant oligonucleotide.

A variation of the procedure (Fig. 8.2) outlined above involves oligonucleotides containing inserted or deleted sequences. As long as stable hybrids are formed with single-stranded wild-type DNA, priming of *in vitro* DNA synthesis can occur, ultimately giving rise to clones corresponding to the inserted or deleted sequence (Wallace *et al.* 1980, Norrander *et al.* 1983).

The single-primer method has a number of deficiencies

The efficiency with which the single-primer method yields mutants is dependent upon several factors. The double-stranded heteroduplex molecules that are generated will be contaminated both by any single-stranded non-mutant template DNA that has remained uncopied and by partially double-stranded molecules. The presence of these species considerably reduces the proportion of mutant progeny. They can be removed by sucrose gradient centrifugation

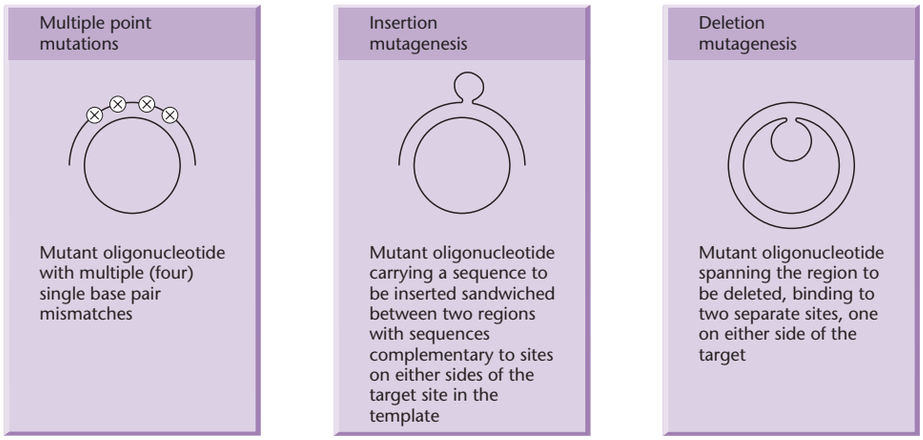


Fig. 8.2
Oligonucleotide-directed mutagenesis used for multiple point mutation, insertion mutagenesis, and deletion mutagenesis.

or by agarose gel electrophoresis, but this is time-consuming and inconvenient.

Following transformation and *in vivo* DNA synthesis, segregation of the two strands of the heteroduplex molecule can occur, yielding a mixed population of mutant and non-mutant progeny. Mutant progeny have to be purified away from parental molecules, and this process is complicated by the cell's mismatch repair system. In theory, the mismatch repair system should yield equal numbers of mutant and non-mutant progeny, but in practice mutants are counterselected. The major reason for this low yield of mutant progeny is that the methyl-directed mismatch repair system of *E. coli* favors the repair of non-methylated DNA. In the cell, newly synthesized DNA strands that have not yet been methylated are preferentially repaired at the position of the mismatch, thereby eliminating a mutation. In a similar way, the non-methylated *in vitro*-generated mutant strand is repaired by the cell so that the majority of progeny are wild type (Kramer *et al.* 1984). The problems associated with the mismatch repair system can be overcome by using host strains carrying the *mut*L, *mut*S, or *mut*H mutations, which prevent the methyl-directed repair of mismatches.

A heteroduplex molecule with one mutant and one non-mutant strand must inevitably give rise to both mutant and non-mutant progeny upon replication. It would be desirable to suppress the growth of non-mutants, and various strategies have been developed with this in mind (Kramer, B. 1984, Carter *et al.* 1985, Kunkel 1985, Sayers & Eckstein 1991).

Another disadvantage of all of the primer extension methods is that they require a single-stranded template. In contrast, with PCR-based mutagenesis (see below) the template can be single-stranded or double-stranded, circular or linear. In comparison with single-stranded DNAs, double-stranded DNAs are much easier to prepare. Also, gene inserts are in general more stable with double-stranded DNAs.

The issues raised above account for the fact that most of the mutagenesis kits that are available commercially make use of multiple primers and double-stranded templates. For example, in the GeneEditor™ system (Fig. 8.3), two primers are used. One of these primers encodes the mutation to be inserted into the target gene. The second encodes a mutation that enhances the antibiotic resistance properties of the ampicillin-resistance determinant on the vector by conferring resistance to ceftazidime as well. After extending the two primers to yield an intact circular DNA molecule, the mutated plasmid is transformed into *E. coli* and selection made for the enhanced antibiotic resistance. Plasmids encoding the enhanced antibiotic resistance also should carry the mutated target gene. In a variant of this procedure, the vector has two antibiotic resistance determinants (ampicillin and tetracycline) but one of these (Amp^R) carries a mutation. Again, two primers are used: one carrying the mutation to be introduced to the target gene and the other restores ampicillin resistance. After the *in vitro* mutagenesis steps, the plasmid is transformed into *E. coli* and selection made for ampicillin resistance.

Methods have been developed that simplify the process of making all possible amino acid substitutions at a selected site

Using site-directed mutagenesis it is possible to change two or three adjacent nucleotides so that

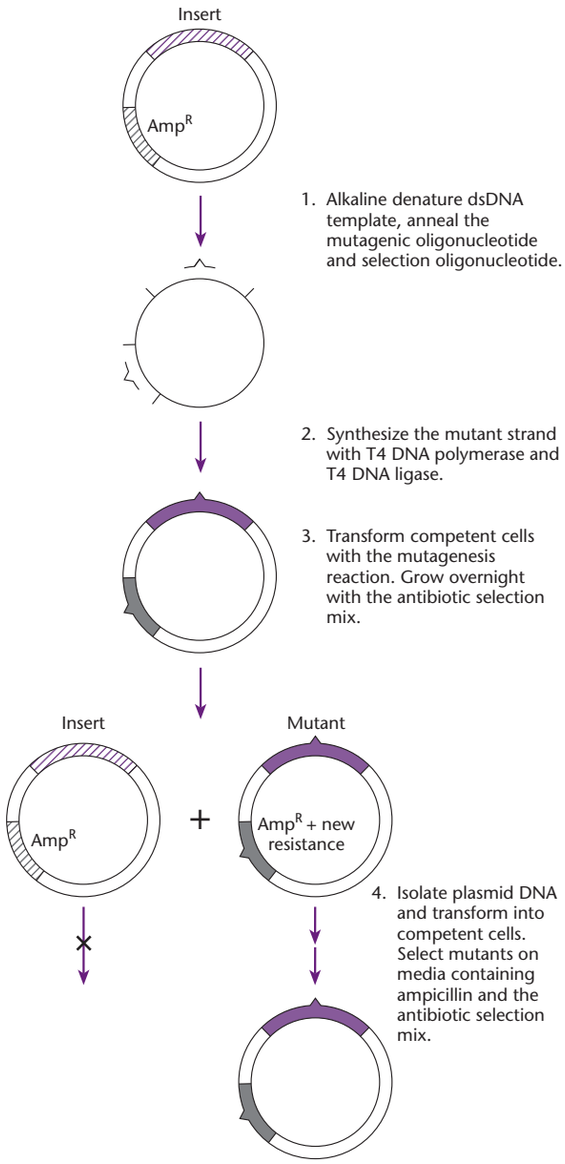


Fig. 8.3 The GeneEditor™ system for generating a high frequency of mutations using site-directed mutagenesis.

every possible amino acid substitution is made at a site of interest. This generates a requirement for 19 different mutagenic oligonucleotides assuming only one codon will be used for each substitution. An alternative way of changing one amino acid to all the alternatives is cassette mutagenesis. This involves replacing a fragment of the gene with different fragments containing the desired codon changes. It is a simple method for which the efficiency of mutagenesis is close to 100%. However, if it is desired to change the amino acids at two sites to all the possible alternatives then 400 different oligos or fragments would be required and the practicality of the method becomes questionable. One solution to this problem is to use doped oligonucleotides (Fig. 8.4). Many different variations of this technique have been developed and the interested reader should consult the review of Neylon (2004).

The PCR can be used for site-directed mutagenesis

Early work on the development of the PCR method of DNA amplification showed its potential for mutagenesis (Scharf *et al.* 1986). Single bases mismatched between the amplification primer and the template become incorporated into the template sequence as a result of amplification (Fig. 8.5). Higuchi *et al.* (1988) have described a variation of the basic method which enables a mutation in a PCR-produced DNA fragment to be introduced anywhere along its length. Two primary PCR reactions produce two overlapping DNA fragments, both bearing the same mutation in the overlap region. The overlap in sequence allows the fragments to hybridize (Fig. 8.5). One of the two possible hybrids is extended by DNA polymerase to produce a duplex fragment. The other hybrid has recessed 5′ ends and, since it is not a substrate for the polymerase, is effectively lost from the reaction

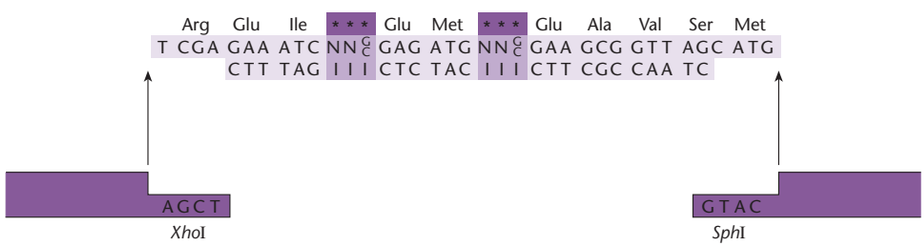


Fig. 8.4 Mutagenesis by means of doped oligonucleotides. During synthesis of the upper strand of the oligonucleotide, a mixture of all four nucleotides is used at the positions indicated by the letter N. When the lower strand is synthesized, inosine (I) is inserted at the positions shown. The double-stranded oligonucleotide is inserted into the relevant position of the vector.

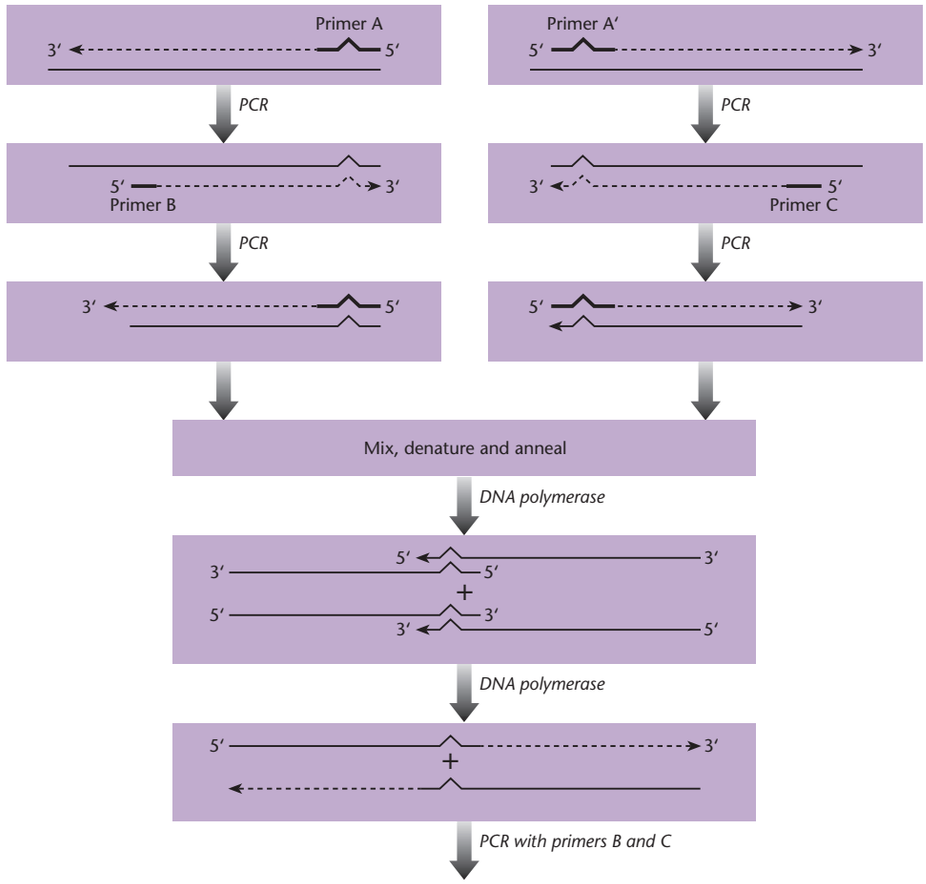


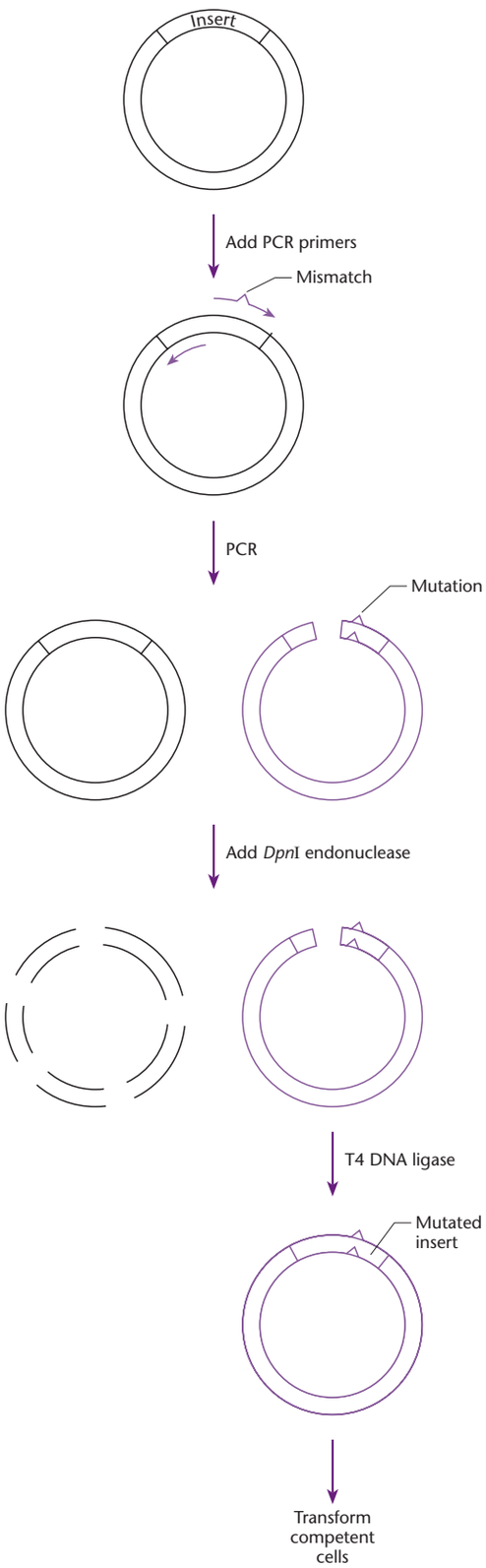
Fig. 8.5 Site-directed mutagenesis by means of the PCR. The steps shown in the top-left corner of the diagram show the basic PCR method of mutagenesis. The bottom half of the figure shows how the mutation can be moved to the middle of a DNA molecule. Primers are shown in bold and primers A and A′ are complementary.

mixture. As with conventional primer-extension mutagenesis, deletions and insertions can also be created.

The method of Higuchi *et al.* (1988) is rather complicated in that it requires four primers and three PCRs (a pair of PCRs to amplify the overlapping segments and a third PCR to fuse the two segments). Commercial suppliers of reagents have developed simpler methods and two of these methods are described below. Two features of PCR mutagenesis should be noted. First, the procedure is not restricted to single base changes: by selecting appropriate primers it is possible to make insertions and deletions as well. Second, *Taq* polymerase copies DNA with low fidelity (see p. 29) and there is a significant risk of extraneous mutations being introduced during the amplification reaction. This problem can be minimized by using a high fidelity thermostable polymerase, and a high template concentration, and fewer than 10 cycles of amplification.

In the Exsite™ method (Fig. 8.6), both strands of the vector carrying the target gene are amplified using the PCR but one of the primers carries the desired mutation. This results in the production of a population of linear duplexes carrying the mutated gene that is contaminated with a low level of the original circular template DNA. If the template DNA was derived from an *E. coli* cell with an intact restriction modification system then it will be methylated and will be sensitive to restriction by the *Dpn*I endonuclease. The linear DNA produced by amplification will be resistant to *Dpn*I cleavage and after circularization by blunt-end ligation can be recovered by transformation into *E. coli*. Any hybrid molecules consisting of a single strand of the methylated template DNA and unmethylated amplicon also will be destroyed by the endonuclease.

In the GeneTailor™ method (Fig. 8.7), the target DNA is methylated *in vitro* before the mutagenesis step and overlapping primers are used. Once again, linear amplicons are produced that carry the desired mutation but in this case they are transformed directly into *E. coli*. The host-cell repair enzymes circularize the linear mutated DNA while the *Mcr*BC endonuclease digests the methylated template DNA leaving only unmethylated, mutated product.



Methods are available to enable mutations to be introduced randomly throughout a target gene

The methods described above enable defined mutations to be introduced at defined locations within a gene and are of particular value in determining structure–activity relationships. However, if the objective of a study is to select mutants with altered and/or improved characteristics then a better approach is to mutate the gene at random and then positively select those with the desired properties. Methods for the random mutagenesis of cloned genes are described in this section and the next while selection methods are described later (p. 148).

It is well known that the polymerase chain reaction is error prone and that there is a high probability of base changes in amplicons. However, even the relatively low fidelity *Taq* polymerase is too accurate to be of value in generating mutant libraries. Nevertheless, increases in error rates can be obtained in a number of ways. One of the commonest ways of achieving this is to introduce a small amount of Mn^{2+}, in place of the normal Mg^{2+}, and to include an excess of dGTP and dTTP relative to the other two nucleotide triphosphates. With this protocol it is possible to achieve error rates of one nucleotide per kilobase (Caldwell & Joyce 1994, Cirino *et al.* 2003). Even higher rates of mutagenesis can be achieved by using nucleoside triphosphate analogs (Zaccolo *et al.* 1996).

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 or a lack of proofreading ability on the part of the polymerase. It might be expected that they generate a completely random set of mutants but in reality the mutant libraries produced are heavily biased. There are three sources of bias. First, the inherent characteristics of the DNA polymerase used mean that some types of errors are more common than others (Cirino *et al.* 2003). The second source of bias arises because of the nature of the genetic code. For example, a single point mutation in a valine codon can change it to one encoding phenylalanine, leucine, isoleucine, alanine, aspartate, or glycine but

Fig. 8.6 (*left*) The Exsite™ method for generating mutants using the PCR. The parental plasmid (shown in blue) carrying the target gene is derived from a restriction-proficient strain of *E. coli* and so is methylated. This makes it sensitive to the *Dpn*I endonuclease and hence it can be eliminated selectively from the final PCR mixture.

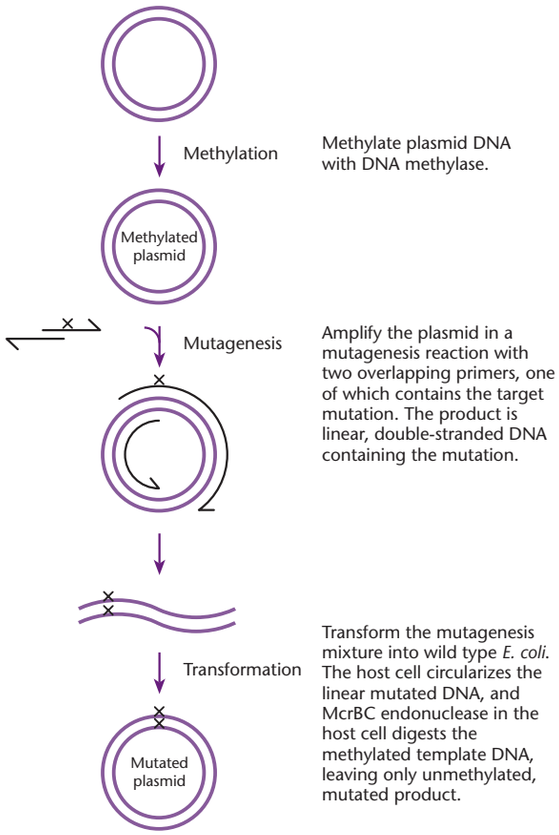


Fig. 8.7 The GeneTailor™ method for generating mutants using the PCR.

two or three adjacent point mutations are required to change it to one encoding all the other amino acids. The final source of bias arises from the process of amplification. A mutant that is generated early in the amplification process will be over-represented in the final library compared to one that arises in later rounds of amplification.

Error-prone PCR protocols are effective as a means of randomly changing one amino acid into another in the final protein. However, sometimes it might be desirable to explore the effect of randomly deleting or inserting amino acids and this is possible using the random insertion/deletion (RID) process devised by Murakami *et al.* (2002, 2003). The method is based on ligating an insertion or deletion cassette at nearly random locations within the gene.

Altered proteins can be produced by inserting unusual amino acids during protein synthesis

All the mutation methods described above result in the replacement of one or more amino acid residues in a protein with other *natural* amino acids, e.g. the replacement of a phenylalanine residue with tyrosine, tryptophan, histidine, etc. The ability to incorporate unnatural amino acids into proteins *in vivo* would permit the production of large quantities of proteins with novel properties. For example, the replacement of methionine with selenomethionine facilitates the determination of the three-dimensional structure of proteins (Hendrickson *et al.* 1990). While it is possible to "force" bacteria to incorporate unnatural amino acids into proteins (for review, see Link *et al.* 2003) a better method is to engineer the translational apparatus. This is achieved by generating an aminoacyl-tRNA synthetase and tRNA pair that function independently of the synthetases and tRNAs endogenous to *E. coli* (Wang *et al.* 2001a, Santoro *et al.* 2003). Such a pair are said to be *orthogonal* and satisfy a number of criteria:

- The tRNA is not a substrate for any of the endogenous *E. coli* synthetases but functions efficiently in protein translation.
- The orthogonal synthetase efficiently aminoacylates the orthogonal tRNA whose anticodon has been modified to recognize an amber (UAG) or opal (UGA) stop codon.
- The synthetase does not aminoacylate any of the endogenous *E. coli* tRNAs.

Archaebacteria appear to be an especially good source of orthogonal pairs for use in *E. coli*.

Modifying the anticodon on the tRNA such that it recognizes amber and opal codons is relatively easy. However, the synthetase also needs to be modified such that it charges the cognate tRNA with unusual amino acids more efficiently than the normal amino acid. To do this a library of synthetase mutants is generated and subjected to positive selection based on suppression of an amber codon located in a plasmid-borne gene encoding chloramphenicol acetyltransferase (Wang *et al.* 2001a). Using this approach the tyrosyl-tRNA synthetase of *Methanococcus jannaschii* was modified to permit the site-specific incorporation into proteins of phenylalanine and tyrosine derivatives such as *O*-allyltyrosine, *p*-acetyl-phenylalanine, and *p*-benzoyl-phenylalanine. These modified amino acids can be used as sites for chemical modification of the protein *in vitro* after purification, e.g. the attachment of fluorescent labels (Chin *et al.* 2003, Link *et al.* 2003).

There have been two significant developments of the above technique. In the first of these, Zhang

et al. (2003) have shown that chemical modification of proteins can occur *in vivo* as well as *in vitro*. For example, *m*-acetylphenylalanine was substituted for Lys7 of the cytoplasmic domain of protein Z and for Arg200 of the outer membrane protein LamB. On addition of a membrane-permeable dye (fluorescein hydrazide) to intact cells, these modified proteins were selectively labeled. In the case of cells expressing the modified LamB derivative, labeling was possible with a range of fluorescein derivatives that are not membrane permeable. The second development is the ability to charge the orthogonal tRNA with glycosylated amino acids. For example, Zhang *et al.* (2004) were able to synthesize in *E. coli* a myoglobin derivative containing β–N-acetylglucosamine (GlcNAc) at a defined position. This GlcNAc moiety was recognized by a saccharide-binding protein and could be modified by a galactosyltransferase.

Phage display can be used to facilitate the selection of mutant peptides

In phage display, a segment of foreign DNA is inserted into either a phagemid or an infectious filamentous phage genome and expressed as a fusion product with a phage coat protein. It is a very powerful technique for selecting and engineering polypeptides with novel functions. The technique was developed first for the *E. coli* phage M13 (Parmley & Smith 1988), but has since been extended to other phages such as T4 and λ (Ren & Black 1998, Santini *et al.* 1998).

The M13 phage particle consists of a single-stranded DNA molecule surrounded by a coat consisting of several thousand copies of the major coat protein, P8. At one end of the particle are five copies each of the two minor coat proteins P9 and P7 and at the other end five copies each of P3 and P6. In early examples of phage display, a random DNA cassette (see above) was inserted into either the P3 or the P8 gene at the junction between the signal sequence and the native peptide. *E. coli* transfected with the recombinant DNA molecules secreted phage particles that displayed on their surface the amino acids encoded by the foreign DNA. Particular phage displaying peptide motifs with, for example, antibody-binding properties were isolated by affinity chromatography (Fig. 8.8). Several rounds of affinity chromatography and phage propagation can be

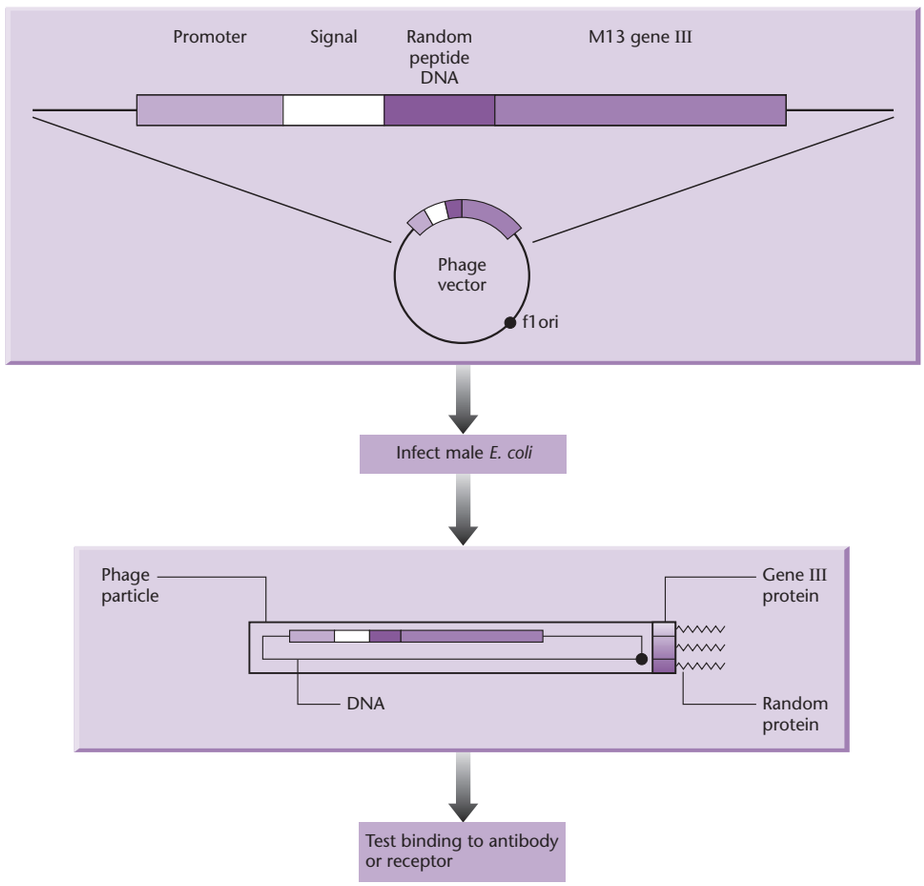


Fig. 8.8 The principle of phage display of random peptides.

used to further enrich for phage with the desired binding characteristics. In this way, millions of random peptides have been screened for their ability to bind to an anti-peptide antibody or to streptavidin (Cwirla *et al.* 1990, Devlin *et al.* 1990, Scott & Smith 1990), and variants of human growth hormone with improved affinity and receptor specificity have been isolated (Lowman *et al.* 1991).

One disadvantage of the original method of phage display is that polypeptide inserts greater than 10 residues compromise coat-protein function and so cannot be efficiently displayed. This problem can be solved by the use of phagemid display (Bass *et al.* 1990). In this system, the starting-point is a plasmid carrying a single copy of the P3 or P8 gene from M13 plus the M13 *ori* sequence (i.e. a phagemid, see p. 75). As before, the random DNA sequence is inserted into the P3 or P8 gene downstream from the signal peptide-cleavage site and the construct transformed into *E. coli*. Phage particles displaying the amino acid sequences encoded by the DNA insert are obtained by superinfecting the transformed cells with helper phage. The resulting phage particles are phenotypically mixed and their surfaces are a mosaic of normal coat protein and fusion protein.

Specialized phagemid display vectors have been developed for particular purposes. For example, phagemids have been constructed that have an amber (chain-terminating) codon immediately downstream from the foreign DNA insert and upstream from the body of P3 or P8. When the recombinant phagemid is transformed into non-suppressing strains of *E. coli*, the protein encoded by the foreign DNA terminates at the amber codon and is secreted into the medium. However, if the phagemid is transformed into cells carrying an amber suppressor, the entire fusion protein is synthesized and displayed on the surface of the secreted phage particles (Winter *et al.* 1994). Other studies (Jespers *et al.* 1995, Fuh & Sidhu 2000, Fuh *et al.* 2000) have shown that proteins can be displayed as fusions to the carboxy terminus of P3, P6, and P8. Although amino-terminal display formats are likely to dominate established applications, carboxy-terminal display permits constructs that are unsuited to amino-terminal display.

For a detailed review of phage and phagemid display, the reader should consult Sidhu (2000) and Sidhu *et al.* (2000).

Cell-surface display is a more versatile alternative to phage display

As noted in the previous section, the size of foreign protein that can be expressed by phage display is rather limited. Microbial cell-surface display systems were developed to solve this problem (for review, see Lee *et al.* 2003) and these systems also have far more applications (Box 8.1). These display systems

Box 8.1 Applications of cell-surface display

There are many different biotechnological and industrial applications of the cell-surface display technology (Fig. B8.1). For example, key proteins from microbial pathogens can be displayed on the surface of bacteria and their ability to elicit antigen-specific responses determined as a major step towards the development of live vaccines. Proteins that bind heavy metals or specific organic pollutants can be expressed on the surface of cells and these cells can be used as specific bioadsorbents for environmental remediation. Alternatively, new enzyme activities can be expressed on the cell surface to promote environmental degradation of pollutants or for use in industrial biocatalysis. Finally, by anchoring enzymes, receptors, or other signal-sensitive components to the cell's surface new biosensors could be developed. For a review of this topic the reader should consult Benhar (2001).

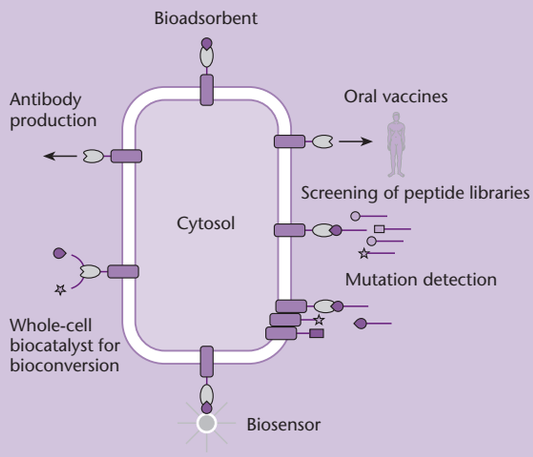


Fig. B8.1 Applications of microbial cell-surface display. Reproduced from Lee *et al.* (2003), with permission from Elsevier.

involve expressing a heterologous peptide or protein of interest (the passenger or target protein) as a fusion protein with various cell-surface proteins (carrier proteins). Depending on the properties of the passenger and carrier proteins, the passenger protein is expressed as an N-terminal, a C-terminal or a sandwich fusion.

For a cell-surface protein to be a successful carrier it should satisfy four requirements. First, it should have an efficient signal peptide to permit the fusion protein to pass through the inner membrane. Second, it should have a strong anchoring structure to keep fusion proteins on the cell surface without detachment. Third, it should be compatible with the passenger protein such that the fusion is not unstable. Finally, it should be resistant to attack by proteases present in the periplasmic space or the growth medium. In Gram-negative bacteria such as *E. coli* many different proteins have been subjugated as carriers. Basically, these proteins fall into two classes: outer membrane proteins (e.g. the adhesin protein, peptidiglycan-associated lipoprotein, and the OmpC and TraT proteins) and protein components of appendages such as pili and flagella. Where outer membrane proteins are used as the carrier it is important to know which part of them is exposed on the outer surface of the cell since this needs to be the site of insertion of the passenger protein.

The passenger protein to be displayed is selected by the required application but its properties influence the translocation process and the effectiveness of the display procedure. For example, the formation of disulfide bridges at the periplasmic side of the outer membrane can affect the efficiency of translocation. Also, the presence of many charged or hydrophobic residues can result in inefficient secretion. Thus, if display technology is used to screen variants produced by random mutagenesis, there may be negative or positive selection for those mutants that affect the efficiency of translocation.

Protein engineering

One of the most exciting aspects of recombinant DNA technology is that it permits the design, development, and isolation of proteins with improved operating characteristics and even completely novel proteins (Table 8.1). The principle of the methods described so far in this chapter is that the gene is mutated, either at a discrete site or at random, and then selection made for a protein variant with the desired property. The improved variant can be subjected to further rounds of mutagenesis and selection, a process known as *directed evolution*. The paradigm for this approach is the enzyme subtilisin. Every

Table 8.1 Some examples of protein engineering.

Example	Method	Reference
Increased rate and extent of biodesulfurization of diesel by modification of dibenzothiophene mono-oxygenase	RACHITT	Coco *et al.* 2001
Generation of a subtilisin with a half-life at 65°C that is 50 times greater than wild type by recombining segments from five different subtilisin variants	StEP	Zhao *et al.* 1998
Conversion of a galactosidase into a fucosidase	Shuffling	Zhang *et al.* 1997
Enhanced activity of amylosucrase	Random mutagenesis plus shuffling	Van der Veen *et al.* 2004
Generation of novel DNA polymerases from a combination of rat DNA polymerase beta and African swine fever virus DNA polymerase X	SCOPE	O'Maille *et al.* 2002
Generation of novel β-lactamase by recombining two genes with 40% amino acid identity and 49% nucleotide sequence identity	SISDC	Hiraga & Arnold 2003

property of this serine protease has been altered including its rate of catalysis, substrate specificity, pH-rate profile, and stability to oxidative, thermal, and alkaline inactivation (for review, see Bryan 2000). Variants also have been produced that favor aminolysis (synthesis) over hydrolysis in aqueous solvents (see Box 8.2).

An alternative approach to directed evolution is *gene shuffling*. The principle of this method is that many protein variants with desirable characteristics

Box 8.2 Improving enzymes

Oxidation-resistant variants of α_1-antitrypsin (AAT)

Cumulative damage to lung tissue is thought to be responsible for the development of emphysema, an irreversible disease characterized by loss of lung elasticity. The primary defense against elastase damage is AAT, a glycosylated serum protein of 394 amino acids. The function of AAT is known because its genetic deficiency leads to a premature breakdown of connective tissue. In healthy individuals there is an association between AAT and neutrophil elastase followed by cleavage of AAT between methionine residue 358 and serine residue 359 (see Fig. B8.2). After cleavage, there is negligible dissociation of the complex.

Smokers are more prone to emphysema, because smoking results in an increased concentration of leucocytes in the lung and consequently increased exposure to neutrophil elastase. In addition, leucocytes liberate oxygen free radicals and these can oxidize methionine-358 to methionine sulfoxide. Since methionine sulfoxide is much bulkier than methionine, it does not fit into the active site of elastase. Hence oxidized AAT is a poor inhibitor. By means of site-directed mutagenesis, an oxidation-resistant mutant of AAT has been constructed by replacing methionine-358 with valine (Courtney *et al.* 1985). In a laboratory model of inflammation, the modified AAT was an effective inhibitor of elastase and was not inactivated by oxidation. Clinically, this could be important, since intravenous replacement therapy with plasma concentrates of AAT is used with patients with a genetic deficiency in AAT production.

Improving the performance of subtilisin

Proof of the power of gene manipulation coupled with the techniques of *in vitro* (random and site-directed) mutagenesis as a means of generating improved enzymes is provided by the work done on subtilisin over the past 15 years (for review, see Bryan 2000). Every property of this serine protease has been altered, including its rate of catalysis, substrate specificity, pH-rate profile, and stability to oxidative, thermal, and alkaline inactivation. In the process, well over 50% of the 275 amino acids of subtilisin have been changed. At some positions in the molecule, the effects of replacing the usual amino acid with all the other 19 natural amino acids have been evaluated.

Many of the changes described above were made to improve the ability of subtilisin to hydrolyze protein when incorporated into detergents. However, serine proteases can be used to synthesize peptides and this approach has a number of advantages over conventional methods (Abrahmsen *et al.* 1991). A problem

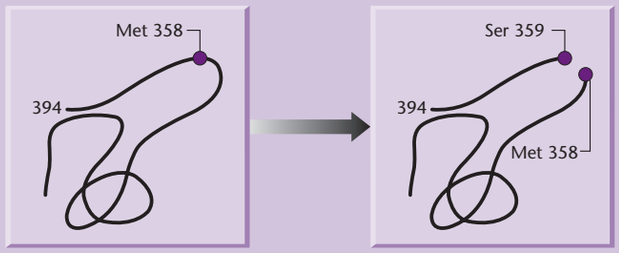


Fig. B8.2 The cleavage of α_1-antitrypsin on binding to neutrophil elastase.

continued

Box 8.2 *continued*

with the use of subtilisin for peptide synthesis is that hydrolysis is strongly favored over aminolysis, unless the reaction is undertaken in organic solvents. Solvents, in turn, reduce the half-life of subtilisin. Using site-directed mutagenesis, a number of variants of subtilisin have been isolated with greatly enhanced solvent stability (Wong *et al.* 1990, Zhong *et al.* 1991). Changes introduced included the minimization of surface changes to reduce solvation energy, the enhancement of internal polar and hydrophobic interactions, and the introduction of conformational restrictions to reduce the tendency of the protein to denature. Designing these changes requires an extensive knowledge of the enzyme's structure and function. Chen and Arnold (1991, 1993) have provided an alternative solution. They utilized random mutagenesis combined with screening for enhanced proteolysis in the presence of solvent (dimethyl formamide) and substrate (casein).

The engineering of subtilisin has now gone one step further, in that it has been modified such that aminolysis (synthesis) is favored over hydrolysis, even in aqueous solvents. This was achieved by changing a serine residue in the active site to cysteine (Abrahmsen *et al.* 1991). The reasons for this enhancement derive mainly from the increased affinity and reactivity of the acyl intermediate for the amino nucleophile (Fig. B8.3). These engineered "peptide ligases" are in turn being used to synthesize novel glycopeptides. A glycosyl amino acid is used in peptide synthesis to form a glycosyl peptide ester, which will react with another C-protected peptide in the presence of the peptide ligase to form a larger glycosyl peptide.

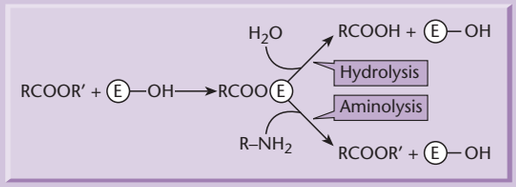


Fig. B8.3 The aminolysis (synthetic) and hydrolysis reactions mediated by an acylated protease.

already exist in nature and novel combinations of these variants may have even more desirable properties (Fig. 8.9). There are three sources of variants for gene shuffling. First, different polymorphisms of the gene of interest might exist naturally in a single organism or might have been created by random *in vitro* mutagenesis (as described on p. 146). Second, the same protein with the same activity may be found in other organisms but the gene and protein sequences will be different. Third, the protein of interest might belong to a protein family where the different members have different but related activities.

A good example of gene shuffling is work done on subtilisin by Ness *et al.* (1999). They started with the genes for 26 members of the subtilisin family and created a library of chimeric proteases. When this library was screened for four distinct enzyme properties, variants were found that were significantly improved over any of the parental enzymes for each individual property. Similarly, Lehmann *et al.* (2000) started with a family of mesophilic phytases whose amino acid sequence had been determined. Using these data they constructed a "consensus"

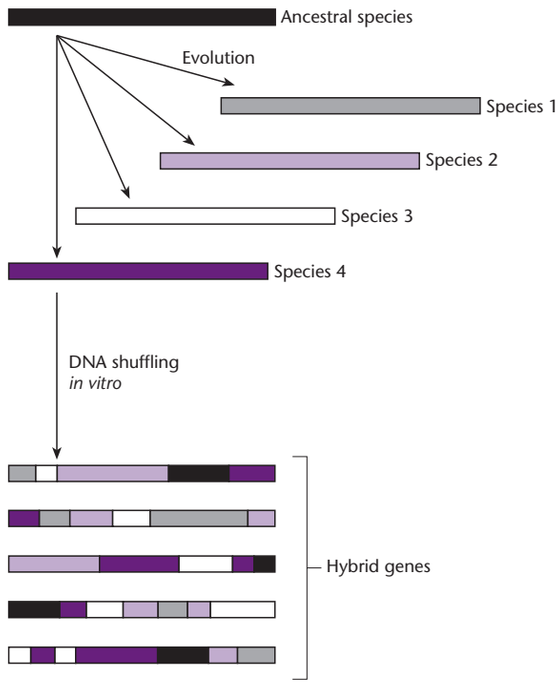


Fig. 8.9 Schematic representation of gene shuffling.

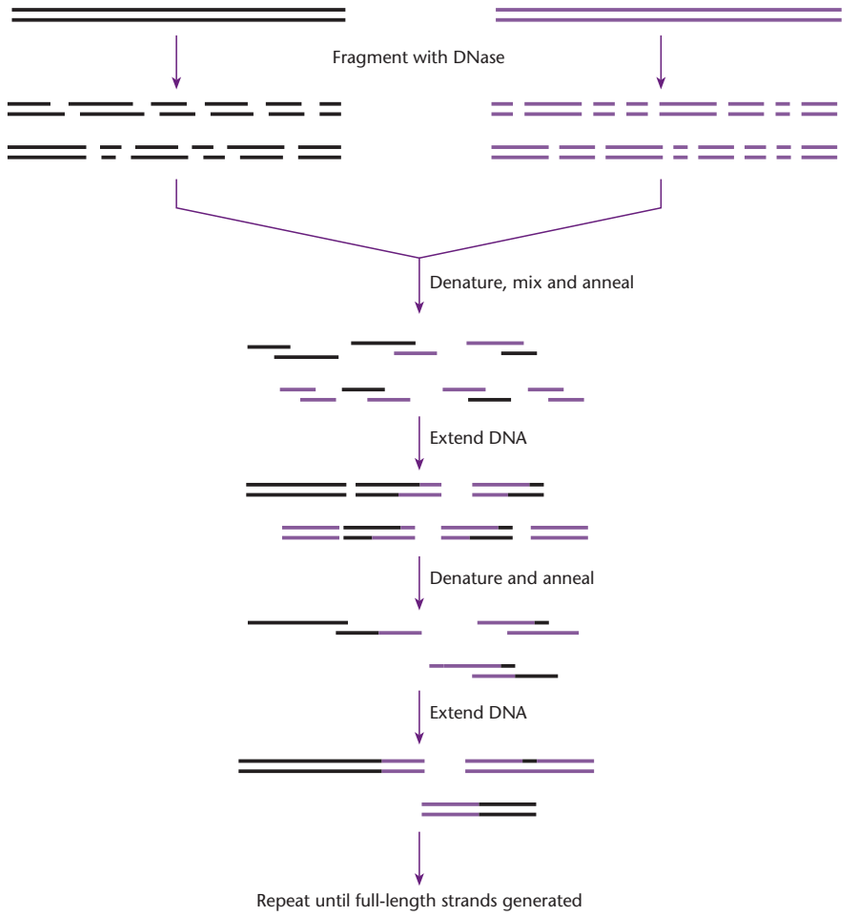


Fig. 8.10 The original method of gene shuffling. After fragmentation of the two homologous genes, the cycles of denaturation, annealing, and extension are continued until full-length genes can be detected by gel electrophoresis.

phytase sequence and found that an enzyme with this sequence was much more thermostable than any of the parent enzymes.

A number of different methods of gene shuffling have been developed

In the original method of gene shuffling (Stemmer 1993, 2004), one starts by purifying the different genes that will provide the source of variation. These genes are digested with DNase to generate the fragments that will be recombined. The fragments from the different sources are mixed together and subjected to repeated rounds of melting, annealing, and extension (Fig. 8.10). Eventually a full-length gene should be synthesized and this can be amplified by the PCR and cloned. The smaller the fragments that are produced in the initial step the greater the number of single site variations that can be incorporated in the final product. However, the smaller the fragments the greater the number of cycles needed to reassemble a complete gene.

An alternative method is the staggered extension process (StEP, Zhao *et al.* 1998). This also relies on repeated cycles of melting, annealing, and extension to build the variant genes. However, in the StEP process one starts with a mixture of full-length genes, denatures them, and then primes the synthesis of complementary strands (Fig. 8.11). After a short period of primer extension, the DNA is subjected to a round of melting, annealing, and extension. Some of the extended primers will anneal to templates with a different base sequence and on further extension will generate chimeras. The more cycles of extension, melting, and annealing the greater the variability that can be produced.

RACHITT (random chimeragenesis on transient templates) is conceptually similar to the original DNA-shuffling method but is designed to produce chimeras with a much larger number of crossovers (Coco *et al.* 2001, Coco 2003). In this method the gene fragments are generated from one strand of all but one of the parental DNAs (Fig. 8.12). These fragments then are reassembled on the full-length

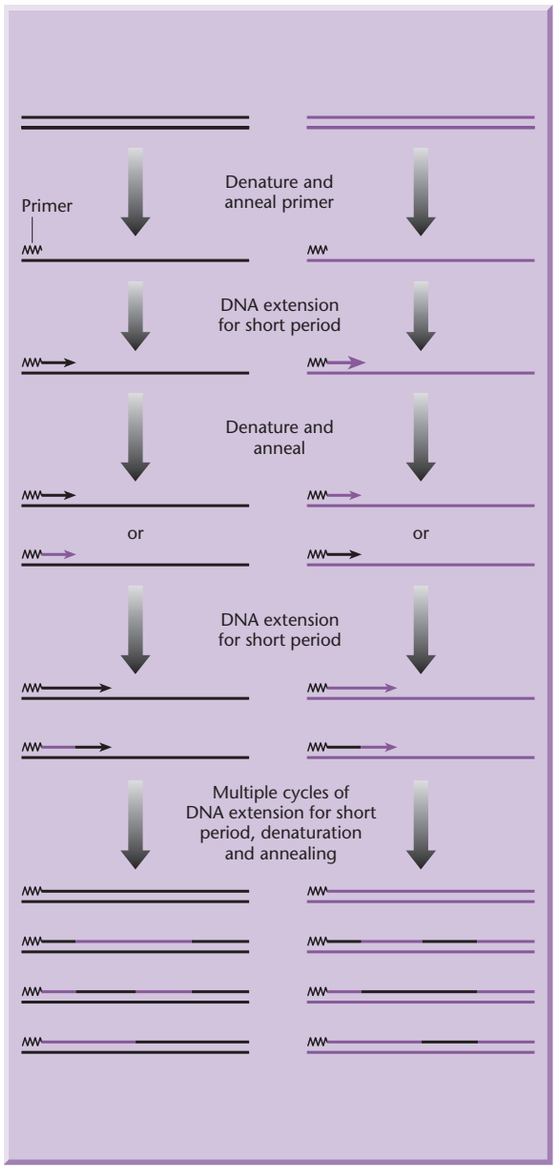


Fig. 8.11 The StEP method for generating hybrid proteins. In the example shown, a hybrid gene will be constructed from two homologous genes (shown in purple and black). Cloning of the hybrid gene will result in the production of a hybrid protein. For clarity, only one strand of each gene is shown after the initial denaturation step.

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.

Each of the methods described above has its advantages and disadvantages and all of them rely to a greater or lesser extent on the annealing of mismatched DNA sequences. Thus there is always a chance that the parental molecules will be recreated preferentially or that the degree of variation generated will not be as great as expected. However, methods for "forcing" the generation of recombinants have been developed (for review, see Neylon 2004).

Chimeric proteins can be produced in the absence of gene homology

The gene-shuffling methods described above have an absolute requirement for significant homology between the parental sequences. However, there may be a wish to create hybrids between proteins with functional similarities but whose sequence homology is less than 50%. Achieving this requires methods for combining non-homologous sequences and the first one to be developed (Ostermeier *et al.* 1999) was ITCHY (incremental truncation for the creation of hybrid enzymes). This method is based on the direct ligation of libraries of fragments generated by the truncation of two template sequences, each template being truncated from opposite ends (Fig. 8.13). This ligation procedure removes any need for homology at the point of crossover but the downside is that the DNA fragments may be reconnected in a way that is not at all analogous to their position in the template gene.

In the original ITCHY process the incremental truncation was performed using timed exonuclease digestions. In practice, these digestions are difficult to control. An improved process was developed where the initial templates are generated with phosphorothioate linkages incorporated at random along the length of the gene (Lutz *et al.* 2001a). Complete exonuclease digestion then generates fragments with lengths determined by the position of the nuclease-resistant phosphorothioate linkage. This method is known as thio-ITCHY and is much simpler to perform. One drawback of ITCHY libraries is that they contain only one crossover per gene. However, by combining ITCHY libraries with DNA-shuffling methods, a process known as SCRATCHY, it is possible to generate additional variation (Lutz *et al.* 2001b).

A major problem with methods such as ITCHY is that they generate large numbers of non-functional sequences due to mutations, insertions, and deletions. Furthermore, when one examines the three-dimensional structure of proteins it is clear that they are organized into domains and motifs. Therefore, a more attractive way of generating chimeric

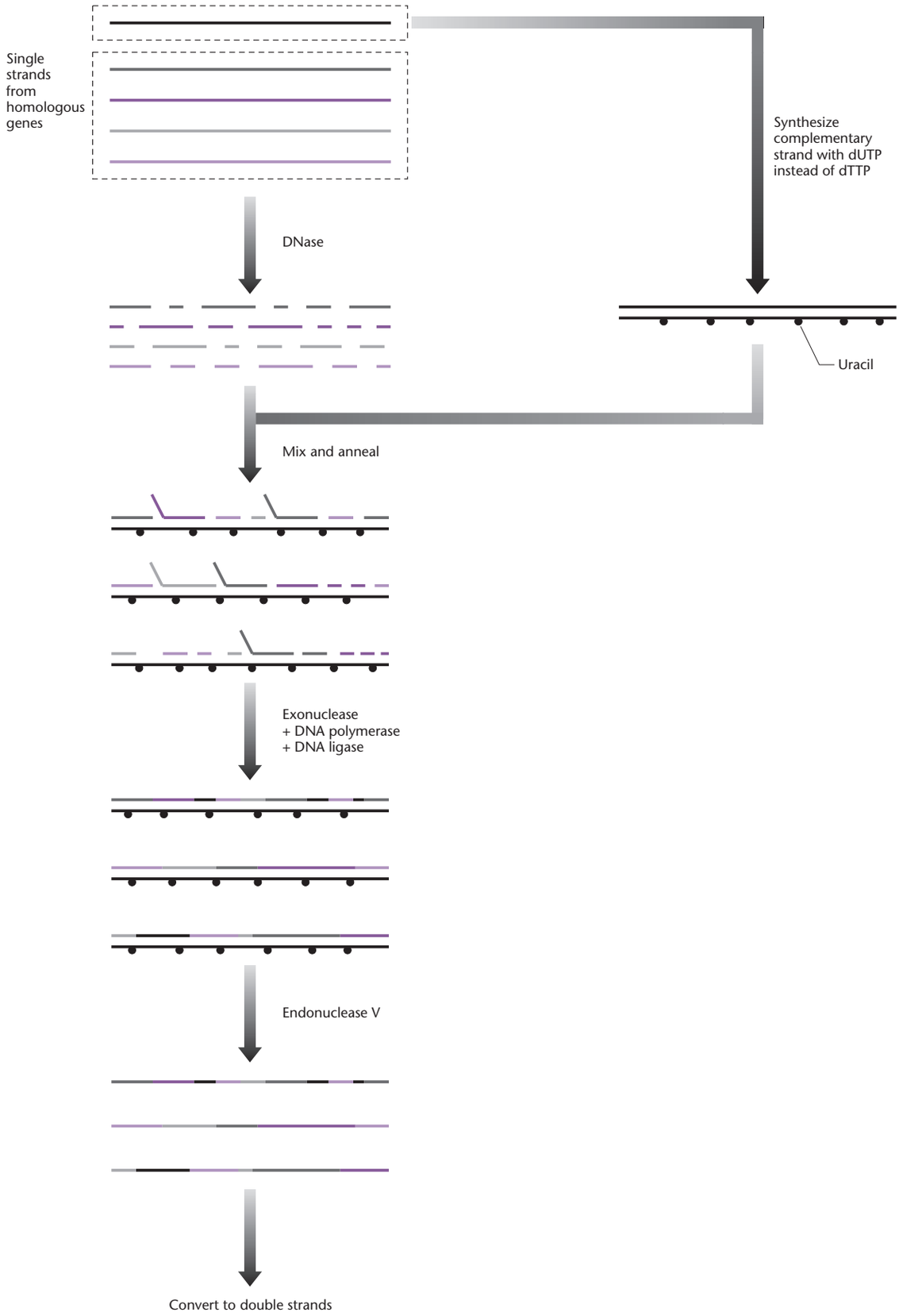


Fig. 8.12 The RACHITT method for creating hybrid proteins.

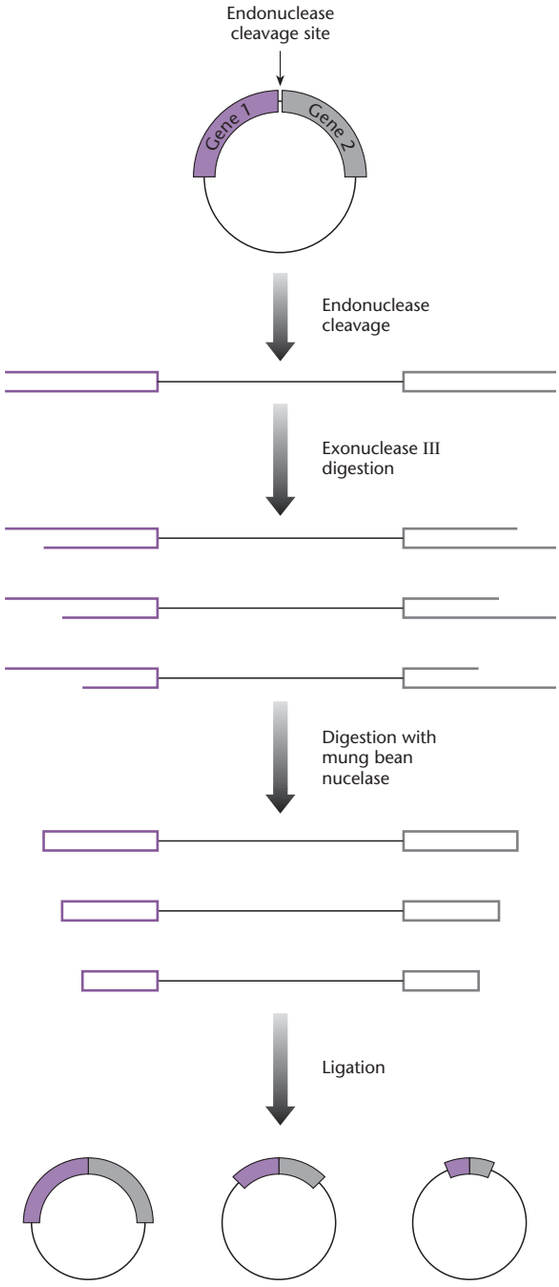


Fig. 8.13 The ITCHY method for creating hybrids of two related proteins. In the figure, the two related proteins are encoded by genes 1 (shown in purple) and 2 (shown in gray). The end result is a hybrid gene comprising the 5′ end of gene 1 and the 3′ end of gene 2.

proteins might be to recombine these domains and motifs in novel ways. Two general methods of doing this have been developed (O'Maille *et al.* 2002, Hiraga & Arnold 2003) and these are SCOPE (structure-based combinatorial protein engineering) and SISDC (sequence-independent site-directed chimeragenesis).

Suggested reading

Brannigan J.A. & Wilkinson A.J. (2002) Protein engineering 20 years on. *Nature Reviews Molecular Cell Biology* **3**, 964–70.

A short but excellent review of the development and pharmaceutical applications of protein engineering.

Collins C.H., Yokobayashi Y., Umeno D. & Arnold F.H. (2003) Engineering proteins that bind, move, make and break DNA. *Current Opinion in Biotechnology* **14**, 371–8.

Another short but excellent review that focuses on what can be achieved with protein engineering rather than on the methods themselves.

Link A.J., Mock M.L. & Tirrell D.A. (2003) Non-canonical amino acids in protein engineering. *Current Opinion in Biotechnology* **14**, 603–9.

Lu Y. (2005) Design and engineering of metalloproteins containing unnatural amino acids as non-native metal-containing cofactors. *Current Opinion in Chemical Biology* **9**, 118–26.

These two papers provide short reviews of the novel chemistries that are possible once unusual amino acids are introduced to proteins.

Lutz S. & Patrick W.M. (2004) Novel methods for directed evolution of enzymes: quality not quantity. *Current Opinion in Biotechnology* **15**, 291–7.

Neylon C. (2004) Chemical and biochemical strategies for the randomisation of protein encoding DNA sequences: library construction methods for directed evolution. *Nucleic Acids Research* **32**, 1448–59.

Each of the methods for generating gene libraries is reviewed in these papers with particular attention being given to the practicality of the methods and the characteristics of the libraries that are produced.

Roodveldt C., Aharoni A. & Tawfik D.S. (2005) Directed evolution of proteins for heterologous expression and stability. *Current Opinion in Structural Biology* **15**, 50–6.

A short review of the application of protein engineering for overproduction of commercial proteins.