2 Making Protein Variants

Summary. Protein expression converts DNA into protein. It involves two steps: the transcription of the coding region of the DNA into mRNA by RNA polymerase followed by the translation of the mRNA into protein by the ribosome. Proteins are manufactured by modification of the natural protein synthesis pathways in microbes like *Escherichia coli*. The modifications specify the protein to make and increase the amount of protein made. The DNA encoding the target proteins is carried on plasmids which are circular, independently replicating genetic elements. Protein synthesis also requires DNA regions outside the coding sequence to start and stop transcription of the DNA to mRNA and translation of the mRNA to protein. Experimental techniques for construction of plasmids include template-independent DNA synthesis, cutting of DNA with restrictions enzymes and assembly of DNA fragments using the polymerase chain reaction.

Key learning goals

- Protein overexpression plasmids encode the target proteins, the control elements needed to make large amounts of protein, and other regions required for plasmids to replicate in microbes.
- The main control of protein expression is at transcription of DNA to mRNA step. The *lac* operator (a region of DNA) and lac repressor (a protein that binds to the *lac* operator) serve as on/off switches for transcription of DNA to mRNA. Addition of a non-hydrolyzable lactose analog, IPTG, turns on transcription. Other control points are the plasmid copy number and the strength of the ribosome binding site.
- The pET21 plasmid contains the gene encoding the target protein, T7 RNA polymerase promoter & terminator, *lac* operator, *lac*I gene and an ampicillin resistance gene. The pET plasmids uses T7 RNA polymerase, which is 5-10-fold more efficient than endogenous *E. coli* RNA polymerase.
- Site-directed mutagenesis uses the polymerase chain reaction with chemically or

enzymatically synthesized DNA fragments (primers) to alter the DNA sequence to encode protein variants.

2.1 Protein expression (DNA → mRNA → protein)

Protein expression is the conversion of a DNA sequence into the corresponding protein. It consists of two steps. First, the DNA encoding the target protein is transcribed into messenger RNA (mRNA) by RNA polymerase. Second, this mRNA is translated into protein by the ribosome.

In biotechnology, proteins are manufactured by microbes, typically bacteria or yeast, that have been modified to make large amounts of the target protein. In the laboratory, researchers grow these microbes in 2-liter shake flasks containing approximately 400 mL of culture media (a mixture of amino acids, sugars, and other nutrients) in a warm incubator. Industrial manufacture uses large fermenters similar to those in a brewery. The details below refer to protein expression in prokaryotes like *Escherichia coli*.

Protein expression requires control elements outside the region that encodes the target protein. One set of control elements bind and release the RNA polymerase that transcribes the DNA, while another set of control elements bind and release the ribosome that translates the genetic code into protein. Thus, the DNA surrounding the gene that encodes the target protein also contains four additional regions (two start signals, two stop signals) outside the DNA sequence that encodes the target protein.

Transcription starts by binding the RNA polymerase to the double-stranded DNA upstream of the coding sequence, Figure 2.1. The recognition site, called the promoter, includes of two regions each 6 bases long. One occurs approximately 10 base pairs upstream of the start codon and the other occurs 35 base pairs upstream of the start codon. The RNA polymerase is larger than this region so that it covers a larger region when it binds - about 40 base pairs upstream and 20 base pairs downstream of the start codon.

Upon binding, the RNA polymerase unwinds approximately 16 bases of DNA around the start codon to expose single-stranded DNA as a template for transcription. The RNA polymerase starts the synthesis of an RNA strand that is complementary to the DNA. As the synthesis proceeds the polymerase unwinds the template DNA ahead of it and rewinds the DNA behind it, maintaining an unwound region of about 16 base pairs in the region of transcription.

RNA synthesis continues until the polymerase encounters a termination signal. Here the transcription stops, the RNA is released from the polymerase, and the polymerase dissociates from its DNA template. The termination signal in *E. coli* consists of a symmetrical inverted repeat of a GC-rich sequence followed by four or more A residues. Transcription of the GC-rich inverted repeat results in the formation of a segment of RNA that can form a stable stem-loop structure by complementary base pairing. This structure in the RNA disrupts its base pairing with the DNA template, which causes the termination. Termination occurs *after* the terminator region because it is the RNA, not the DNA that forms the stem-loop structure. The DNA must first be transcribed into RNA to form the stem-loop structure that causes termination.

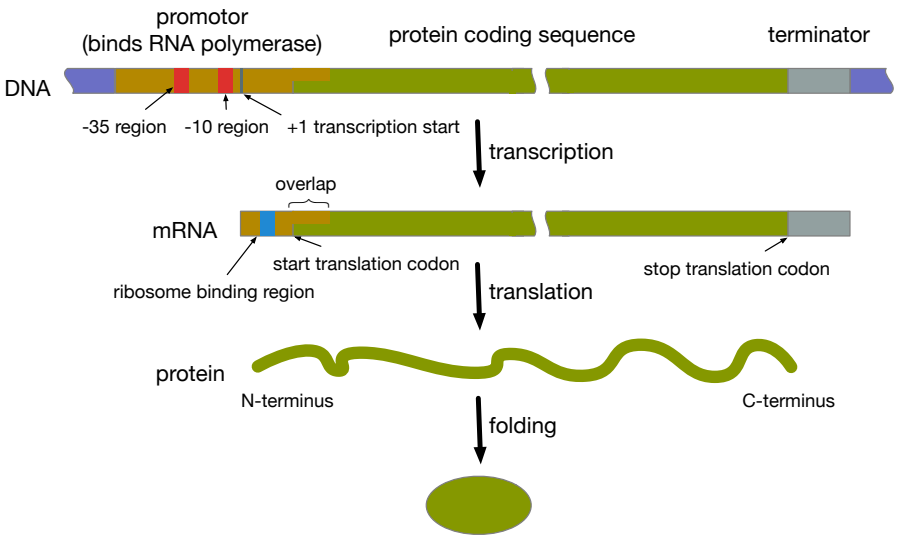

Figure 2.1. Protein expression requires control elements outside the protein coding region. The transcription of DNA into RNA requires an upstream region to bind the RNA polymerase (promoter region) and a downstream region to release the RNA polymerase (terminator region). The promoter region includes two six-base regions at 10 bp upstream and 35 bp upstream that recognize the RNA polymerase. The translation of RNA to protein requires a ribosome binding site, a start codon and a stop codon. Both the RNA polymerase and the ribosome are large molecular machines. Although they recognize and bind to the regions marked, they also cover a much larger region upon binding including part of the protein coding region marked overlap.

Translation starts when the ribosome binds to the mRNA. The ribosome binding site (Shine-Dalgarno sequence) is approximately 8 bases upstream of the start codon (methionine codon), but the ribosome covers a much larger region including the start codon and the beginning of the coding region. Protein synthesis occurs as the ribosome moves along the RNA sequence in groups of three nucleotides and attaches the next amino acid to the growing protein chain. Termination occurs when the ribosome encounters a stop codon and detaches from the RNA and releases the completed protein. Thus, translation requires a ribosome binding site before the coding sequence and a stop codon after the coding region. The start codon, also required, is within the coding region.

The DNA sequence encoding the control elements for expression of a β-lactamase protein, Figure 2.2, shows a specific example of these elements.

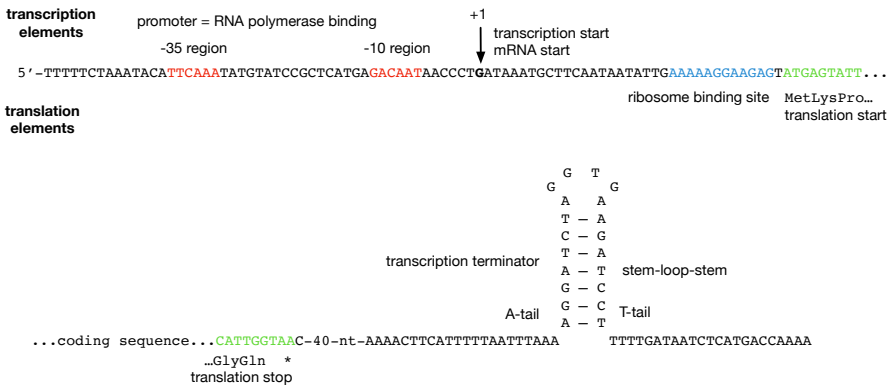

Figure 2.2. DNA sequence for the expression of a β-lactamase protein showing the transcription and translation control elements. The -35 and -10 regions (red text) in the promoter bind the endogenous RNA polymerase, but the polymerase extends beyond these regions. The ribosome binds to the ribosome binding site (blue text), but also extends beyond this region. During transcription, the terminator region of the RNA forms the stem-loop-stem structure shown, which releases the mRNA from the polymerase and stops transcription. The DNA does not for the stem-loop-stem structure.

2.2 Protein overexpression

Cells typically make small amounts of each protein, but to manufacture proteins for biotechnology applications one wants to make as much protein as possible. This overproduction of the target protein is called protein overexpression. The two common strategies to increase protein expression are to use a more efficient RNA polymerase and to include multiple copies of the DNA encoding the target protein. These approaches are very effective; up to 50% of the protein in an *E. coli* culture can be the target protein.

2.2.1 A better RNA polymerase

E. coli bacteria contain an endogenous RNA polymerase, but RNA polymerases from bacteriophage are often faster. Bacteriophage RNA polymerases hijack the cellular machinery of the host bacteria to make phage proteins instead of bacterial proteins. The most common replacement polymerase for increased protein expression is the T7 RNA polymerase, which originates from bacteriophage T7. The T7 RNA polymerase catalyzes RNA synthesis 5-10-fold faster than the *E. coli* RNA polymerase, so more than the normal amount of mRNA will be produced, which will lead to more protein.

Overexpressing the target protein with T7 RNA polymerase requires a gene that encodes and expresses the T7 RNA polymerase protein. Some strains of *E. coli* (such as BL21(DE3) and BL21(DE3)pLysS) have been engineered to contain the T7 RNA polymerase gene in their genomic DNA. These strains should be used for protein overexpression that requires the T7 RNA polymerase.

Overexpressing the target protein with T7 RNA polymerase also requires a different promoter upstream of the target gene so that it binds the T7, not the *E. coli*, RNA polymerase. The DNA sequence of the T7 promoter differs from the endogenous promoters in *E. coli*.

2.2.2 Multiple copies of the target protein gene

A second way to increase the amount of protein is to have many copies of the DNA that encodes the target protein and its control elements. Protein overexpression uses plasmids, which are small (5-10 kb), independently replicating, circular DNA molecules. The pET plasmid, which will be discussed later in this chapter, makes 15-20 copies of itself in *E. coli*. These multiple copies of the target protein DNA increase the amount of protein that is produced. pET plasmids are classified as low copy number plasmids; some high copy number plasmids make 500-700 copies of themselves in each cell.

2.2.3 Ribosome binding site

The slow step of translating mRNA into protein is the initiation of protein synthesis, which depends on binding of the ribosome to the ribosome binding site (Shine-Dalgarno sequence). Optimization of the initiation step can include optimizing the distance between the Shine-Dalgarno sequence and the start codon, eliminating any secondary structure in the mRNA that could hinder binding of the ribosome, and adding additional binding elements upstream of the Shine-Dalgarno sequence.[1]

2.3 Turning protein overexpression on or off

Strong overexpression of the target protein hinders the growth of microbes. Diverting most of the cell's energy and resources to synthesis of the target protein can even prevent growth entirely. The solution is to initially turn off the overepression of the target protein to allow the bacteria to grow, then turn it on to produce the protein. During overexpression phase the bacteria no longer multiply, but instead make the target protein. The most commonly used on/off switch is the *lac* operator.

The *lac* operator originates from the lactose operon or *lac* operon. An operon is a set of genes whose expression is regulated together. The *lac* operon consists of three structural genes and a *lac* operator, Figure 2.3. The *lac* operator is a section of DNA that controls the expression of the three genes by reversibly binding the *lac* repressor protein, which is encoded by the *lac*I gene elsewhere in the genome. The *lac* operator is in the promoter region where the RNA polymerase must bind, so when the *lac* repressor protein binds to the *lac* operator, it blocks the RNA polymerase from binding. By default, the expression of the three genes in the lactose operon is off. Binding of allolactose to the *lac* repressor protein changes its shape so that it dissociates from the *lac* operator. Expression is now turned on because the RNA polymerase can bind to the promoter.

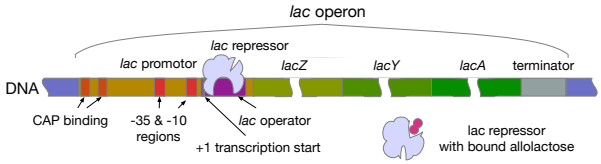

Figure 2.3. The *lac* operon consists of three structural genes and control elements that turn on expression of these genes only when lactose is present in the cell. By default the *lac* repressor protein binds to the *lac* operator and prevents the RNA polymerase from binding in this region. When allolactose is present, it binds to the *lac* repressor protein and changes it shape so that it no longer binds to the *lac* operator. Upon release of the *lac* repressor protein, the RNA polymerase can bind and start transcription of the three structural genes.

The off state of the lactose operator is not perfect. The *lac* repressor protein sometimes dissociates from the *lac* operator even with no lactose present leading to leakiness, that is, constant low-level expression. Given the strong expression from the T7 promoter, the imperfect off state can lead to difficulties in cell growth.

To improve the off state, researchers use a low copy number plasmid and two copies of the *lac* operator. The low copy number of the plasmid reduces the probability that the *lac* repressor protein spontaneously dissociates from the *lac* operator to turn on protein synthesis.

The two *lac* operator switches are used to turn off expression of both the target gene and the T7 RNA polymerase gene. Overexpression of the target protein first requires expression of the T7 RNA polymerase. Placing the *lac* operator in the promoter region blocks this first step. (Expression of the T7 RNA polymerase uses the endogenous *E. coli* promoter.) The second location for the *lac* operator is in the T7 promoter region upstream of the target gene. In the no lactose state, the *lac* repressor protein should prevent expression of the T7 RNA polymerase and of the target protein thus decreasing the leakiness of the off state.

To further ensure that protein overexpression is turned off, some researcher add a third switch - an inhibitor of the T7 RNA polymerase. This inhibitor is a protein, T7 lysozyme. It requires an additional plasmid, pLysS, that encodes synthesis of the T7 lysozyme protein.

In the presence of lactose or an analog, the switches are on. Lactose binds to the lactose repressor protein, which changes its shape in a way that releases it from the two *lac* operators. Upon release of the lactose repressor protein, the endogenous *E. coli* polymerase binds and starts to transcribe the gene encoding the T7 RNA polymerase. After it is expressed, the T7 RNA polymerase can bind to the promoter upstream of the target gene to initiate protein expression.

Lactose induces the *lac* promoter indirectly. Lactose can be isomerized into allolactose, which induces the *lac* promoter.[2] The natural inducer of the *lac* promoter is likely galactosyl glycerol found in plant galactolipids and common in the diet of ruminants and herbivores,[3] Figure 2.4.

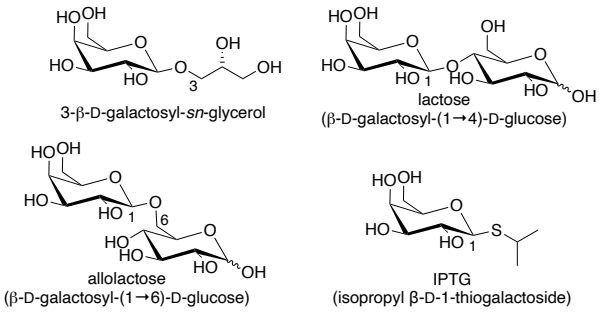

Figure 2.4. Inducers of the *lac* promoter. The natural inducer is likely galactosyl glycerol found in plant galactolipids. Lactose itself is not an inducer. IPTG is an non-hydrolyzable analog of galactosyl glycerol and is used to induce the *lac* promoter in biotechnology.

For biotechnology applications, researchers use a non-hydrolyzable analog of galactosyl glycerol to induce the *lac* promoter. Isopropyl β-D-thiogalactopyranoside, IPTG, contains a sulfur atom at the glycosidic link, which prevent hydrolysis. Induction starts upon addition of IPTG and continues indefinitely since the compound does not hydrolyze.

In summary, there are two or three off switches for protein expression using pET plasmids. Two of the off switches are lac operators in the promoter region: one in the T7lac promoter upstream of the coding region on the plasmid and one in the endogenous promoter upstream of the gene for the T7 RNA polymerase on chromosome. Both of these switches control the synthesis of mRNA. The addition of IPTG turns these switches on. The final switch is the low level expression of T7 lysozyme by the pLysS plasmid to inhibit the T7 RNA polymerase. The BL21(DES)pLysS strain of *Escherichia coli* includes this plasmid.

Figure 2.5 shows an example gene and its associated control elements for expression in *Escherichia coli*. This coding region encodes a poly(ethyleneterephthalate)-degrading enzyme, a PETase.[4] The gene originates from the bacterium *Ideonella sakaiensis*, but the DNA sequence has been optimized for expression in *Escherichia coli*. This optimization replaces codons that are rarely used in *E. coli* with a synonymous, more com-

monly used codon. The start and stop elements for the RNA polymerase are the T7 promoter and the T7 terminator. The T7 promoter contains the *lac* operator, which serves as an on/off switch. The start elements for translation are the ribosome binding site (T7_trans_en_RBS) and the start codon (ATG) and the stop element for translation is the stop codon (TGA).

The PETase gene contains eight additional amino acid residues at the C-terminus: a linker (not labelled) and a His-tag. The linker (CTCGAG) encodes the dipeptide Leu-Glu, while the His-tag (CACCACCATCACCACCAC) encodes six histidine residues. The addition of six histidine residues to the protein simplifies purification. After protein expression, the mixture of proteins is added to a column containing bound nickel or cobalt ions, which have a high affinity for the imidazole ring of histidine. Most proteins bind poorly to the column and elute quickly, but the His-tagged protein binds tightly. After washing away the unwanted proteins, eluting with a buffer containing imidazole releases the His-tagged protein.

2.4 pET plasmids

The pET family of plasmids are often used for protein overexpression. The pET family of plasmids all use the T7 RNA polymerase (pET stands for plasmid for Expression by T7 RNA polymerase), but differ in the some details. The expression plasmid used to make the PETase protein mentioned above is derived from the pET21 plasmid, Figure 2.6. The pET21 plasmid contains an origin of replication, which determines its copy number, and genes to express the lactose repressor protein (*lac*I), and an ampicillin resistance protein (AmpR). The lactose repressor protein binds to the *lac* operator to turn off protein expression. The expression of *lac*I and AmpR are constitutive (always on) at a normal level (the promoters bind the *E. coli* RNA polymerase).

The ampicillin resistance protein or β-lactamase is a selectable marker that ensures that only bacteria containing the pET21 plasmid grow in the culture. Researchers add ampicillin to the growth media, which prevents bacterial growth. Bacteria containing the pET21 plasmid express the β-lactamase, so they grow in the presence of ampicillin. This expression of the AmpR gene is constitutive and at a normal level. Figure 2.2 above shows the sequences of control elements for the AmpR gene on a pET21 plasmid.

The different pET plasmids (about fifty) differ in mainly in the purification tag. The His-tag may be at the N- or C-terminus; it may be removable with a protease if the linker contains a protease-sensitive site. Some plasmids encode a glutathione-S-transferase (GST) purification tag. GST is a 26 kDa protein. It has a high affinity for glutathione and yields higher purity protein than the His-tag. Another family of plasmids used for protein expression are the pBAD plasmids, which use an arabinose promoter, Table 2.1. Protein expression from these plasmids is more tightly controlled than for pET plasmids and can be turned off more completely or tuned to different levels. Protein synthesis is turned on by the addition of different amounts of arabinose.

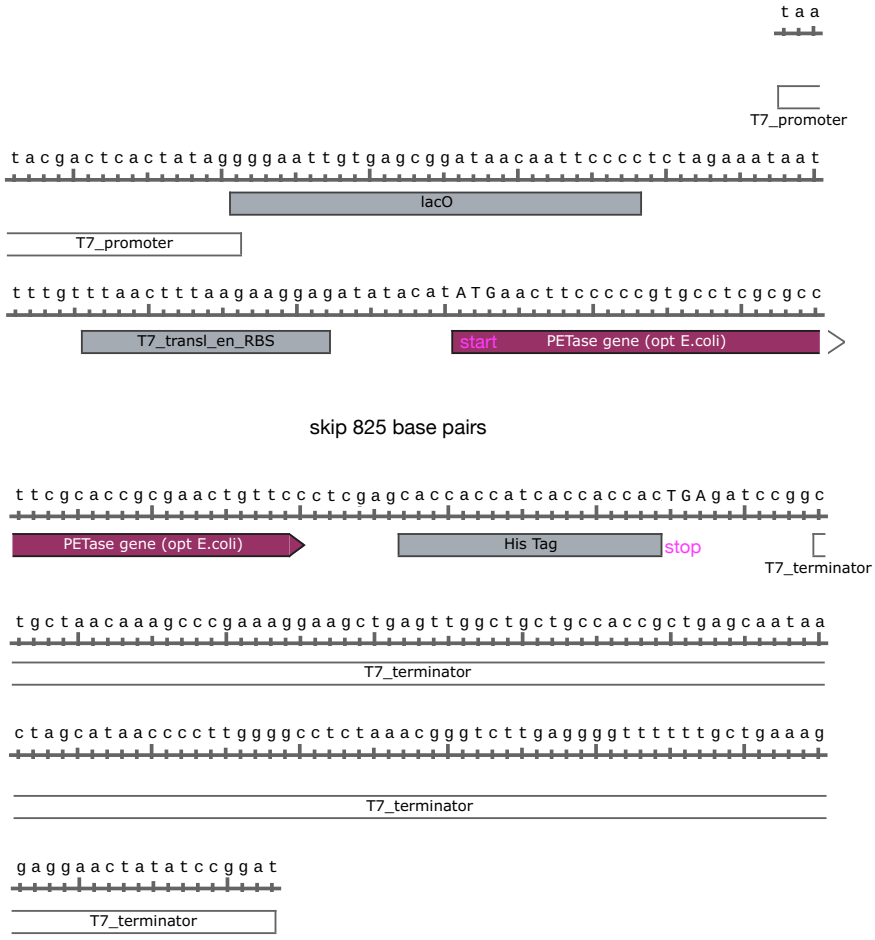

Figure 2.5. Control elements before and after the DNA encoding a PETase gene (maroon fill). The T7 promoter region binds the T7 RNA polymerase. The *lac* operator (lacO) is an on/off switch to control expression. The ribosome binding site is labelled T7_trans_en_RBS. The PETase gene is 870 bp long not including the added His-tag; the central 825 bp are not shown to save space. The His-tag encodes a six-histidine C-terminal tag added to the PETase. The T7_terminator forms secondary structures in the RNA that stop the RNA polymerase.

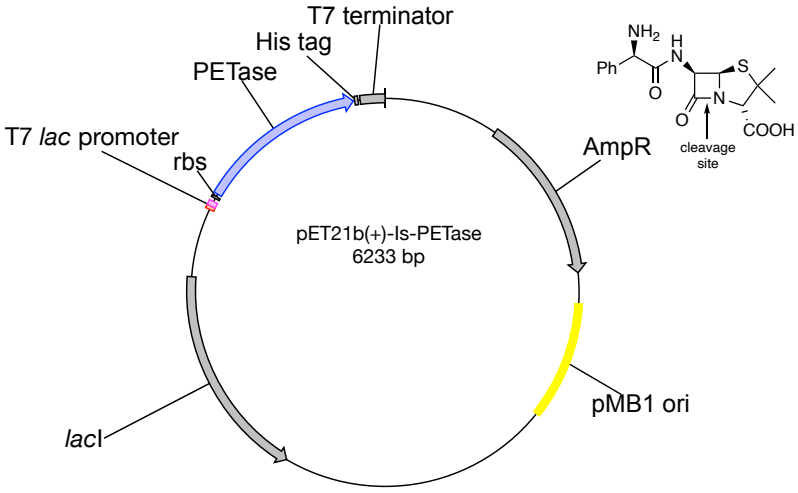

Figure 2.6. A plasmid derived from pET21 to overexpress the PETase protein in *Escherichia coli*. This plasmid contains an origin of replication (yellow section) that controls the copy number of the plasmid. The plasmid also expresses the lactose repressor protein (*lac*I) and the β-lactamase (AmpR) at normal levels. The structure of the antibiotic ampicillin is at the upper right.

Table 2.1. Plasmid families used for protein overexpression.

name	**antibiotic resistance**	**origin of replication**	**promoter, inducer**
pET	ampicillin or kanamycin	pMB1 ori, 15-20 copies per cell	strong, T7 or T7*lac*, IPTG
pBAD	ampicillin or kanamycin	p15A ori, 10-12 copies per cell	tunable, arabinose

2.5 Site-directed mutagenesis

The most common method of protein engineering is site-directed mutagenesis to replace an existing amino acid residue in a protein with a different amino acid that is expected to improve the protein. Site-directed mutagenesis changes the DNA sequence encoding the target protein to encode a different amino acid at the selected location. Although amino acid replacements are the most common change made to improve proteins, one can also insert additional amino acids or delete amino acids using the same methods.

As an example of site-directed mutagenesis, consider the replacement of methionine 222 with alanine to stabilize subtilisin for use in laundry detergent,[5] Fig 2.7. This replacement required replacing the ATG codon for methionine with GCG to encode alanine. In contrast to methionine, alanine lacks a sulfur atom and is unaffected by bleach. Alanine is also smaller than methionine so it does not hinder access of the substrate to the active site.

```
... 218 219 220 221 222 223 224 225 ... amino acid numbering
... AAC GGT ACG TCA ATG GCA TCT CCG ... original DNA sequence (forward strand)
... Asn Gly Thr Ser Met Ala Ser Pro ... original protein encoded

... AAC GGT ACG TCA gcG GCA TCT CCG ... modified DNA sequence
... Asn Gly Thr Ser Ala Ala Ser Pro ... variant protein encoded
```

Figure 2.7. Site-directed mutagenesis to replace methionine at position 222 with alanine involves replacement of two nucleotides in the codon. The original DNA sequence encoding subtilisin (top) specifies methionine at position 222 using the ATG codon. Replacement of this codon with GCG encodes a variant of subtilisin where position 222 is an alanine (bottom). The two substituted nucleotides are shown in lower case.

Smith and coworkers first reported site-directed mutagenesis in 1978,[6] but advances in molecular biology methods have yielded improved procedures. The section below describes using an inverse polymerase chain reaction (inverse PCR) to replace methionine 222 in subtilisin with alanine. In the original work the researchers at Genentech and Genencor used another method since their experiments were completed before the invention of the polymerase chain reaction.[5]

This site-directed mutagenesis method uses an inverse polymerase chain reaction (inverse PCR) to copy the whole plasmid using back-to-back primers,[7] Figure 2.8. The primers are designed to bind back-to-back with the DNA synthesis proceeding outward. Since the plasmid is circular, the result is a linear double stranded copy of the plasmid. Copying the whole plasmid avoids the cut and paste steps that complicated older site-directed mutagenesis methods.

The DNA primers are single-stranded 20-30 nucleotide long oligomers. The location of the primers should be such that desired mutation lies approximately in the center of one of the primers, which is the mutagenic primer. The mutagenic primer encodes the new DNA sequence; thus several nucleotides are not complementary to the plasmid DNA. All extensions of this mutagenic primer contain this mutation. In addition, the complementary strands copied from these extensions also contain this mutation. The result of

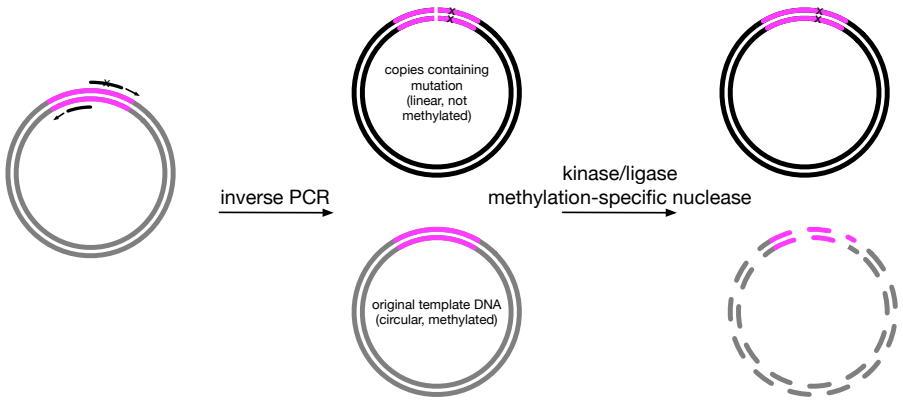

Figure 2.8. Site-directed mutagenesis involves copying the entire plasmid using mutagenic primers followed by a clean-up step. The magenta section represents the target gene for mutagenesis. Extension of two back-to-back primers (black) using the inverse polymerase chain reaction in the direction shown by the arrows yields a linear copy of the plasmid. If one of the primers contains a mismatch (x), then the copies contain this altered DNA sequence. Clean up involves circularization of the linear copies with kinase and ligase enzymes and fragmentation of the original wild type plasmid with a nuclease that only cleaves DNA containing methylation. The wild-type plasmid, prepared by growth in *E. coli*, contains methylation, but the mutated plasmid, created by the polymerase chain reaction, does not contain methylation, so the nucleases does not fragment it.

this copying is linear, double-stranded copies with a mutation at the desired location. One primer is complementary to one DNA strand; the other primer is complementary to the other DNA strand. The primers should have similar melting temperatures and not form dimers with themselves or each other. Figure 2.9 shows an example of primers designed to replace methionine 222 in subtilisin with alanine.

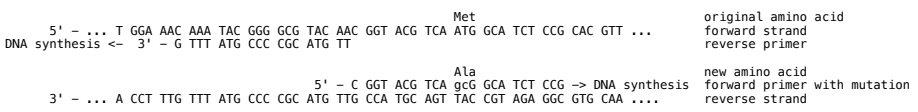

Figure 2.9. Mutagenic back-to-back primers to replace methionine 222 in subtilisin with alanine and a section of the circular plasmid DNA sequence. Top: The reverse primer (18 nt, 50% GC, Tm = 63 °C) binds to the forward strand and contains no mutations. DNA polymerase extends this primer to the left generating the complement to the forward strand. In subsequent PCR cycles, this primer extension will also create complements to the mutated forward strand. Bottom: The forward primer (22 nt, 68% GC, Tm = 61 °C) binds to the reverse strand and contains a two-nucleotide mismatch. DNA polymerase extends this primer to the right generating a complement to the reverse strand. Primers were predicted by NEBaseChanger (https://nebasechanger.neb.com).

An additional incubation with three enzymes circularizes this linear DNA fragment for insertion and replication in bacteria, Figure 2.8. Simultaneous treatment with a kinase to add a phosphoryl group to the 5'-ends of the PCR products and a ligase to join the two ends yield a circular double stranded DNA. Also in the same reaction, methylation-specific nuclease fragments the original template, which does not contain the mutation. The template DNA is the circular wild-type plasmid isolated from *E. coli*. When plasmids replicate in *E. coli*, methyl groups are added to the DNA. The nuclease recognizes and cleaves DNA containing these methyl groups. In contrast, the primers and the DNA synthesized using the polymerase chain reaction do not contain methyl groups and are ignored by this nuclease. The result is a circular copy of the plasmid only containing the altered DNA sequence. Transfer of this plasmid into *E. coli* allows researchers to grow cells, which then make the variant protein encoded.

2.6 Template-independent synthesis of DNA

The site-directed mutagenesis experiment above relies on the ability to synthesize DNA fragments (primers) with a defined sequence. DNA polymerase can synthesize DNA, but only by copying an existing template. To make primers and even entire genes, one needs a template-independent synthesis of DNA.

The first approach to template-independent DNA synthesis was chemical synthesis.[8] This synthesis is relatively complex and involves reactive intermediates and toxic solvents. Researchers normally order the desired sequence from a commercial service. The synthesis starts with a solid silica support. Reagents such as protected nucleoside phosphoramidites are washed over the silica to carry out the multiple steps needed to add each base. The harsh reaction conditions slowly damage the DNA, which caps the length

of the DNA to about 200 bases. More recently, researchers have developed enzymatic approaches for the template-independent synthesis of DNA, Figure 2.10. This enzymatic approach relies on an engineered terminal deoxynucleotidyl transferase (Tdt) for the coupling step. The enzymatic approach generates less waste, can make longer oligonucleotides (>1000 bases) and can make repetitive sequences that the chemical methods cannot.[9],[10]

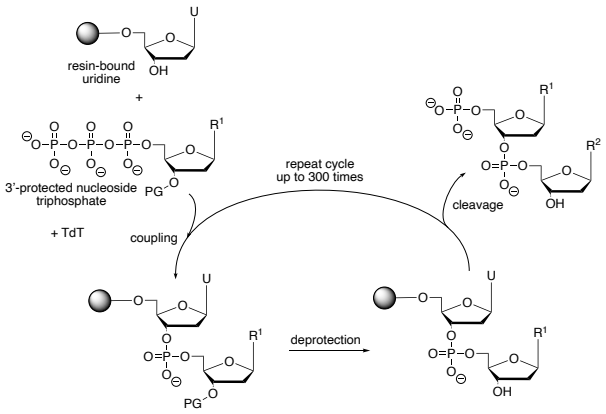

Figure 2.10. Enzymatic synthesis of DNA using 3'-protected nucleoside 5'-triphosphates. Synthesis starts at a uridine attached to a resin bead, which provides a place for terminal deoxynucleotidyl transferase (TdT) to bind and a cleavage site once synthesis is complete. TdT ligates the 3'-protected nucleoside 5'-triphosphate to the 3'-terminus of the growing oligonucleotide chain. Washing removes excess reagents and the pyrophosphate by-product of ligation. Deprotection of the 3'-PG exposed the 3'-hydroxyl for the next synthesis cycle. On completion, the oligonucleotide is cleaved from the resin bead by uracil DNA glycosylase.

Entire genes and longer fragments can be synthesized by chemical or enzymatic synthesis of 300 bp DNA fragments, followed by joining them to create longer oligonucleotides.[11] The experimental procedures for the methods described in this chapter be found online or in molecular biology handbooks.[12]

Glossary

Operator is a section of DNA upstream of the a gene that binds a transcription regulator. For example, the *lac* operator binds the *lac* repressor protein, which inhibits transcription of the downstream gene.

Promoter a region of DNA upstream of a region that encodes one or more proteins. The promoter region binds the RNA polymerase that will transcribe the DNA into mRNA.

Plasmid is a circular, independently replicating, DNA within a microbe separate from the chromosomal DNA. For example, pET plasmids are inserted into *Escherichia coli* to overexpress proteins under the control of a T7 promoter.

Primer is a short (10-50 nucleotides), single-stranded DNA fragment. When this fragment bind to a complementary region on a longer DNA fragment, then DNA polymerase can extend the primer making a complementary copy of the longer DNA fragment. If the primer contains a mutation, then the longer copy will also contain that mutation.

T7 promoter is a promoter that binds the T7 RNA polymerase, which is more efficient than the endogenous RNA polymerase in *Escherichia coli*.

References

1. Salis, H. M., Mirsky, E. A., & Voigt, C. A. (2009). Automated design of synthetic ribosome binding sites to control protein expression. *Nat. Biotechnol.*, *27*(10), 946–950. https://doi.org/10.1038/nbt.1568
2. Jobe, A., & Bourgeois, S. (1972). *lac* Repressor-operator interaction. VI. The natural inducer of the lac operon. *J. Mol. Biol.*, *69*(3), 397–408. https://doi.org/10.1016/0022-2836(72)90253-7
3. Egel, R. (1988). The '*lac*' operon - an irrelevant paradox? *Trends Genet.*, *4*(2), 31–31. https://doi.org/10.1016/0168-9525(88)90062-5
4. Austin, H. P., Allen, M. D., Donohoe, B. S., Rorrer, N. A., Kearns, F. L., Silveira, R. L., Pollard, B. C., Dominick, G., Duman, R., El Omari, K., Mykhaylyk, V., Wagner, A., Michener, W. E., Amore, A., Skaf, M. S., Crowley, M. F., Thorne, A. W., Johnson, C. W., Woodcock, H. L., … Beckham, G. T. (2018). Characterization and engineering of a plastic-degrading aromatic polyesterase. *Proc. Natl. Acad. Sci. U. S. A.*, *115*(19), E4350–E4357. https://doi.org/10.1073/pnas.1718804115
5. Estell, D. A., Graycar, T. P., & Wells, J. A. (1985). Engineering an enzyme by site-directed mutagenesis to be resistant to chemical oxidation. *J. Biol. Chem.*, *260*(11), 6518–6521. https://doi.org/10.1016/S0021-9258(18)88811-1
6. Hutchison 3rd, C. A., Phillips, S., Edgell, M. H., Gillam, S., Jahnke, P., & Smith, M. (1978). Mutagenesis at a specific position in a DNA sequence. *J. Biol. Chem.*, *253*(18), 6551–6560. https://doi.org/10.1016/S0021-9258(19)46967-6
7. Silva, D., Santos, G., Barroca, M., & Collins, T. (2017). Inverse PCR for point mutation introduction. In L. Domingues (Ed.), *PCR: Methods and Protocols* (pp. 87–100). Springer New York. https://doi.org/10.1007/978-1-4939-7060-5_5
8. Roy, S., & Caruthers, M. (2013). Synthesis of DNA/RNA and their analogs via phosphoramidite and H-phosphonate chemistries. *Molecules*, *18*(11), 14268–14284. https://doi.org/10.3390/molecules181114268
9. Hoose, A., Vellacott, R., Storch, M., Freemont, P. S., & Ryadnov, M. G. (2023). DNA synthesis technologies to close the gene writing gap. *Nat. Rev. Chem.*, *7*(3), 144–161. https://doi.org/10.1038/s41570-022-00456-9
10. Seydel, C. (2023). DNA writing technologies moving toward synthetic genomes. *Nat Biotechnol.*, *41*(11), 1504–1509. https://doi.org/10.1038/s41587-023-02006-0
11. Gibson, D. G., Young, L., Chuang, R.-Y., Venter, J. C., Hutchison 3rd, C. A., & Smith, H. O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods*, *6*(5), 343–345. https://doi.org/10.1038/nmeth.1318
12. Green, M. R., & Sambrook, J. (2012). *Molecular Cloning: A Laboratory Manual* (4th ed., 1–3). Cold Spring Harbor Press.

Problems

1. Identify whether the items below are proteins, oligonucleotides or small organic molecules. Explain their role in protein overexpression.

*lac*I gene product

T7 RNA polymerase

lac operator

amipicillin

IPTG

ribosome binding site

stop codon

T7 terminator

His-tag

mutagentic PCR primer

transcription start site

2. a) Why does DNA transcription stop after the terminator, but mRNA translation stop before the stop codon? b) What problems could one encounter when using a T7 promoter on a high copy number plasmid. c) Why do researcher put the gene encoding the T7 RNA polymerase on the genome instead of on the protein expression plasmid?

3. *Plain text files.* In computing, plain text files are files that contain only text characters including white spaces and punctuation. For example, a plain text file could contain the following text:

```
>tr|M5B284|M5B284_BACIU Subtilisin OS=Bacillus subtilis OX=1423 GN=aprE
MRNKKLWISLLFALTLIFTMAFSNMSAQAAGKSSTEKKYIVGFKQTMSAMSSAKKKDVIS
EKGGKVQKQFKYVNAATATLDEKAVKELKQDPSVAYVEEDHIAHEYAQSVPYGISQIKAP
ALHSQGYTGSNVKVAVIDSGIDSSHPDLNVRGGASFVPSETNPYQDGSSHGTHVAGTIAA
LNNSIGVLGVAPSASLYAVKVLDSTGSGQYSWIINGIEWAISNNMDVINMSLGGPSGSTA
LKTVVDKAVSSGIVVAAAAGNEGSSGSSSTVGYPAKYPSTIAVGAVNSSNQRASFSSAGS
ELDVMAPGVSIQSTLPGGTYGAYNGTSMATPHVAGAAALILSKHPTWTNAQVRDRLESTA
TYLGSSFYYGKGLINVEAAAQ
```

In contrast, formatted text files include additional characters within the text to indicate font, bold, line spacing, and other information. For example, `<b>` and `</b>` indicate bold in the example below. The program that you use to open this file must know that these formatting characters are not part of the text, but indicate formatting commands. If not, it may misinterpret these characters as data.

```
>tr|M5B284|M5B284_BACIU Subtilisin OS=Bacillus subtilis OX=1423 GN=aprE
<b>MRNKKLWISLLFALTLIFTMAFSNMSAQAAGKSSTEKKYIVGFKQTMSAMSSAKKKDVIS</b>
EKGGKVQKQFKYVNAATATLDEKAVKELKQDPSVAYVEEDHIAHEYAQSVPYGISQIKAP
```

```
ALHSQGYTGSNVKVAVIDSGIDSSHPDLNVRGGASFVPSETNPYQDGSSHGTHVAGTIAA
LNNSIGVLGVAPSASLYAVKVLDSTGSGQYSWIINGIEWAISNNMDVINMSLGGPSGSTA
LKTVVDKAVSSGIVVAAAAGNEGSSGSSSTVGYPAKYPSTIAVGAVNSSNQRASFSSAGS
ELDVMAPGVSIQSTLPGGTYGAYNGTSMATPHVAGAAALILSKHPTWTNAQVRDRLESTA
TYLGSSFYYGKGLINVEAAAQ
```

Bioinformatics and protein engineering often use plain text files to store information and commands. You can open plain text files using word processing programs like Word or Pages, but to keep them as plain text files, that is, not add formatting text to them, be sure to use a plain text format when you save them (usually 'txt' extension) and not a word processing format (e.g., 'pages' or 'docx' extension). More commonly researchers use text editor programs, which are programs intended to manipulate plain text files: Mac: TextEdit,* BBEdit; Windows: Notepad, Textpad; Linux: pico, vim, nano). The next question requires you to work with a plain text file similar to the example above, which is a FASTA format file. FASTA format files are plain text files used to store DNA or amino acid sequences.

Practice: The `<b>` and `</b>` indicate bold in web pages, which use 'html' format files. Copy the second amino acid sequence text above into a text editor, save it with the extension 'html,' then open that file with a web browser. The `<b>` and `</b>` characters should be hidden and the characters in between should be displayed as bold. Next, change the extension of this file from 'html' to 'txt' and open with the web browser again. The browser now displays all the characters in the file, but does not show anything in bold. A web browser can read both plain text files and web pages, which have the html format. The file extension tells the browser whether to interpret the characters as text or as formatting information. Most bioinformatics and computational chemistry programs expect text files as input. If the file contains formatting or other codes, they will change the meaning of the file and the program will not accept them.

*Unfortunately, the default settings use rich text formatting which includes formatting characters. Change the default by selecting Format → Make Plain Text option in the menu bar. This change will also set the extension of files saved in the future to '.txt' from the default '.rtf'.