Contents

2 Making variant proteins

Summary. Protein engineering is the modification of proteins to improve their properties, usually by replacing amino acid residues with other amino acids. Proteins evolved in nature for their biological role, but when they are used in ways that differ from this role, they usually require improvements. The three type of improvements are changes in protein stability, binding to a target, reactivity with specific substrates. These improvement factors can be expressed as changes in Gibbs energies. Biotechnology uses engineered proteins in three main application areas: medicine, agriculture, and industry. Protein engineering in medicine has tuned the bioavailability of insulin and increased the potency and plasma half-life of monoclonal antibodies. Protein engineering in agriculture has increased the insect resistance of crops and stabilized enzymes for use as feed additives. Protein engineering for industrial applications include stabilizing detergent enzymes and engineering enzymes for pharmaceutical synthesis. Replacing chemical methods for pharmaceutical manufacture with biocatalysis usually yields a greener, more environmentally friendly process.

Key learning goals

- The main experimental technique in protein engineering is site-directed mutagenesis, which replaces existing amino acids in proteins. The challenge of protein engineering is identifying which replacement amino acids will improve the protein.

2.1 Protein engineering is a key tool of biotechnology

The challenge of protein engineering is to identify which amino acid changes could improve the enzyme. Nature's proteins are polymers of the twenty proteinogenic α-amino acids, typically containing 300 amino acids in a specific sequence. Proteins fold into specific three-dimensional shapes, Figure 1.2. The number of possible proteins is astronomical. For a tripeptide, there are 20 choices for the first amino acid, 20 for the second and 20 for the third. The number of possibilities is 203 or 8000. In general, the number of possibilities is given by the number of choices at each location raised to the power of the number of locations, equation 1.1. (1.1) In this case,

For a protein of 300 amino acids, the number of possibilities is 20300 or ~10390, which is vastly larger than the number of stars in the universe (~1020) or even the number of particles in the universe (~1086). Some proteins create biological structures like muscles

and virus particle coats; other proteins are biochemical signals like insulin or vaccines; other proteins are enzymes that catalyze reactions. The number of physically possible proteins is much larger than the number of biologically relevant proteins. Protein engineering creates some of these physically possible, but nonexistent, proteins. Figure 1.2. Surface representation of the protease subtilisin where the twenty amino acids are colored in one of twelve different colors according to traditional amino acid properties. Protein engineering is the modification of proteins to improve them for medical, agricultural and industrial applications. The challenge is to identify where to make replacements and what the replacement amino acids should be. The stability of this protease was improved by replacing methionine 222 (labelled yellow patch near the center) with alanine. Three categories of improvements. Although proteins can be engineered to be better for many different types of applications, the improvements themselves fall into three categories: protein stability, protein binding to a target and enzyme reactivity. Improvements in protein stability allow proteins to tolerate harsh conditions, improvements in binding increase the affinity of a protein for its target and improvements in enzyme reactivity increase reaction rates. In some case, the selectivity of an enzyme is also important. Selectivity is the ability of an enzyme to distinguish two targets or substrates or to form different products. Selectivity improvements are a type of binding or reactivity improvement where the binding or reactivity of toward one molecule increase and toward the other decreases. Identifying the goal of the protein engineering - the type and amount of improvement needed - is a critical first step of protein engineering. Identifying this goal reveals possible approaches to use and the size of the improvement needed suggests how difficult it will be to achieve the goal.

Applications that differ from the natural function usually require protein engineering Proteins have evolved in nature for specific natural functions. When proteins are used for another function or in a different way, they will likely require engineering. Natural selection chooses proteins that improve fitness in nature, not for various applications that humans identify. Natural proteins may be accidentally suitable for these applications, but won't be optimized for them. For example, nature has not evolved proteins tolerant to organics solvents since no organisms live in organic solvents, but some proteins evolved to tolerate high temperatures often also tolerate organic solvents. In some cases, the different requirements are obvious. For example, an enzyme-catalyzed manufacture of a pharmaceutical intermediate may involve an unnatural substrate, organic cosolvents and elevated temperatures. In these cases, protein engineering can adjust the active site to fit the new substrate to increase reaction rate and stabilize the protein to organic solvents and elevated temperatures. In other cases, the differences between the natural protein function and application may be subtle. For example, human insulin regulates the glucose levels in the blood. Patients with type I diabetes cannot make insulin and require injections of human insulin. Although the goal is the same - to maintain glucose levels, the dosing method differs. Patients inject insulin several times per day in contrast to the pancreas continuously secreting varying amounts of insulin. This difference in dosing required engineering fast-acting variants of human insulin for injection before meals and slow-acting variants to maintain glucose levels overnight. Evolution does not maximize the properties of a protein, but only maintains the minimum necessary. Natural selection cannot distinguish between a protein that is improved just

enough so that another protein limits function and a protein that is much better. Selection favors both equally since the much better property does not contribute to fitness. A hammer made from extra-hard steel does not drive nails any better than a hammer with steel just hard enough. An example of 'just enough' is that most proteins are just stable enough for their natural function (Pace, 1975; Taverna & Goldstein, 2002). This marginal stability is not an inherent limitation of proteins since single amino acid substitutions can stabilize most proteins without compromising activity (Fersht & Serrano, 1993). The reason for the marginal stability is evolution cannot select for more than the minimum needed stability (Bloom et al., 2007). One can also generalize this 'just enough' quality to other protein properties. Enzymes are just fast enough for their function, they bind their substrates just tightly enough and discriminate between substrates just enough for them to function. Improvement factors and Gibbs energy changes. Improvements achieved by protein engineering can be expressed as Gibbs energy changes, eq. 1.2. The Gibbs energy change, , corresponds to differences in rates or equilibria associated with stability, binding or reactivity of the original protein. The corresponds to the change between the original protein and the variant. Here R represent the gas constant, which is 8.314 J/mol· °K to express the Gibbs energy change in J/mol or 1.987 cal/mol·°K to express it in cal/mol. T is the temperature in °K and ln is the natural logarithm function. Improvement factor must be a ratio of rates or equilibrium constants. For example, a ten-fold improvement in reaction rate corresponds to a Gibbs energy change of –1.36 kcal/mol (5.70 kJ/mol) at room temperature (25 °C). Because of the logarithmic relationship between improvement factor and Gibbs energy change, a hundred-fold improvement in reaction rate corresponds to twice the Gibbs energy change and a thousand-fold improvement corresponds to three times the Gibbs energy change of the ten-fold improvement.

(1.2)

The sign of for an improvement may be positive or negative depending on the property being improved. In the example above, the improvement factor was >1 since the reaction rate is faster for the improved enzyme. The logarithm of a number >1 is positive, so the Gibbs energy change is negative. If the improvement was in protein stability, then the unfolding equilibrium constant is larger for the improved, more stable, protein (Kunfold-variant < Kunfold-original) and the improvement factor is less than one. The logarithm of a value less than one is negative, so is positive. These calculated Gibbs energy changes will help predict that the number and extent of structure changes needed in the molecular interactions to achieve this degree of improvement. The improvement factor must be expressed as a ratio of equilibrium constants or rates. For example, it may be a protein unfolding equilibrium, Kunfold or a dissociation equilibrium between a protein and its target, Kdiss or a rate constant such as the turnover number of an enzyme kcat or the specificity constant, kcat/KM. Equation 1.2 cannot be used for ratios of other quantities such as yield of a reaction or melting temperature of a protein.

1.2 Engineering proteins for medicine Biopharmaceuticals are biomolecules used as drugs. They can be proteins such as enzymes or monoclonal antibodies, nucleic acids such as micro RNAs or microbes such as attenuated live vaccines. This text considers only protein biopharmaceuticals. The first protein biopharmaceuticals were recombi-

nant equivalents of the natural protein where the biotechnology advantage was the ability to produce large amounts of protein. For example, blood factors that were isolated from donated blood could now be produced in mammalian cell cultures due to genetic engineering. The subsequent generations of blood factors were improved engineered variants. For example, engineering increased the potency of the blood coagulation protein, Factor VII used to treat hemophilia, so that each dose required less protein (Harvey et al., 2003), Table 1.1. The action of this blood coagulation protein depends on its binding to a membrane surface containing acidic phospholipids, which is characteristic of damaged vascular cells or of platelets that adhere to the damage as the clot forms. Mutagenesis of the membrane-binding domain of Factor VII yielded variants that bound more tightly to the membrane and therefore were more potent. The best variant contained five substitutions and was 149-296 fold better than wild-type Factor VII. Other researchers increased its potency with three amino acid substitutions that increased its catalytic activity 30-fold (Persson et al., 2012). Proteolytic activity of Factor VII activates the next factor in the coagulation cascade. Table 1.1 Examples of engineered protein biopharmaceuticals Insulin variants engineered for altered bioavailability are even better than human insulin at controlling blood sugar levels. The engineering relied on the fact that monomeric insulin is the active form, while multimeric forms are inactive. Insulin lispro is a fast-acting insulin analog for injections before meals because it favors the monomeric form (Bolli et al., 1999). The amino acid sequence of lispro differs only in the reversal of the penultimate lysine and proline residues on the C-terminal end of the B-chain. (Insulin consists of 51 amino acids arranged as a dimer of an A-chain and a B-chain linked by disulfide bonds.) This residue change blocks the formation of insulin dimers and hexamers and increases the amount of monomeric insulin (the active form), thereby creating a faster-acting variant. In contrast, substitutions to create a slow-acting insulin analog, glargine, promote association of the monomers into an insoluble precipitate, which then slowly releases monomeric insulin to maintain basal levels of insulin for an extended time. An essential class of therapeutic proteins is monoclonal antibodies and monoclonal antibody drug conjugates. Common goals of antibody engineering are increasing binding the target and minimizing immunogenicity (Ryu et al., 2012). One example of a therapeutic monoclonal antibody is trastuzumab (Herceptin®) used to treat breast cancer. This antibody interferes with the HER2 receptor (human ErbB2 receptor tyrosine kinase), which signals cell proliferation. Cancer cells, usually breast cancer, overexpress the HER2 receptor causing the cancer cells to grow uncontrollably. Interfering with the receptor stops the uncontrolled growth. Trastuzumab is a humanized version of a mouse antibody, Figure 1.3. Humanization prevents an allergic reaction to a mouse antibody. This humanization involved first, the transplantation of the complementarity-determining regions from the mouse antibody into a human IgG antibody and second, an optimization of the surrounding regions, called framework regions (Carter et al., 1992). The transplantation of the complementarity-determining regions changed 24 amino acids in the light chain and 32 in the heavy chain. This substitution did not yield a useful antibody. The engineered version bound less tightly than the mouse antibody and did not block cancer cell proliferation upon binding. Next, an additional seven substitutions in the surrounding (framework) regions increased the binding to the target 100-fold making it three-fold tighter than the mouse antibody. This tighter binding restored the ability of the antibody to block cell proliferation. Trastuzumab

was approved in the USA in 1998 to treat metastatic breast cancer. Figure 1.3 To engineer trastuzumab, researchers combined the complementarity-determining regions from a mouse antibody (orange; 56 amino acids) with a human IgG antibody (lavender). Full activity required seven additional substitutions (green) outside the binding region. Growth factors and cytokines, such as vascular endothelial growth factor, epidermal growth factor and granulocyte colony-stimulating factor, stimulate cell recruitment, proliferation, morphogenesis, and differentiation. Inhibitors of these growth factors act as anticancer drugs, while enhanced growth factors promote wound repair. Natural growth factors lack stability and specificity; they are prone to degradation in serum and have multiple activities. In one example of protein engineering of a cytokine, Sarkar et al. (2002) increased the half-life and potency of granulocyte colony-stimulating factor by modifications that increased its recycling back to the bloodstream. Cantor and coworkers (2011) engineered asparaginase, an enzyme for leukemia treatment, to be less immunogenic. Treatment of childhood acute lymphoblastic leukemia involves injections of asparaginase, which cleaves asparagine to aspartate, thereby depriving the cancerous cells of asparagine. Healthy cells can make asparagine, but these cancer cells cannot. Humans do not have an asparaginase enzyme, so the treatment uses recombinant enzyme from Escherichia coli. One limitation is that some patients' immune systems recognize this enzyme as a non-human protein and inhibit it by binding it with an antibody or even induce a severe allergic reaction. Cantor and coworkers replaced the amino acid sequences on the surface of E. coli asparaginase that are likely to be recognized by human T-cells with non-recognized sequences. The replacements maintained the catalytic activity of the asparaginase. Vaccines against pertussis (whooping cough) contain engineered proteins. The bacteria that causes pertussis, Bordetella pertussis, secrete a five-subunit protein, known as pertussis toxin. Immunization with the pertussis toxin protein or just its S1 subunit protects against disease. However, the S1 subunit is an enzyme that catalyzes the ADP-ribosylation of GTP-binding proteins, thereby causing toxicity. Two amino acid substitutions in the S1 subunit - Glu129Gly, Arg9Lys - deactivated its catalytic activity, and therefore toxicity, while retaining its immunogenic potential (Loosmore et al. 1990). The mRNA vaccines against the SARS-CoV-2 virus encode the spike protein, which sits on the virus surface and binds to cell surface proteins during infection. Upon binding to the target cell, the spike protein changes conformation. Researchers reasoned that spike protein in the pre-binding conformation would make a better vaccine that the spike protein in the post-binding conformation. Binding of antibodies to the spike protein before it binds to the cell could prevent infection. To stabilize the pre-binding conformation, researchers replaced six residues in the spike protein with proline (Hsieh et al., 2020). Proline limits the flexibility at those sites due to its ring structure and keeps the spike protein in the pre-binding conformation.

1.3 Engineering proteins for agriculture Agricultural biotechnology aims mainly to improve productivity of crops and animals. Some products, like bovine somatotrophin (a growth hormone), are copies of the natural protein produced in recombinant E. coli bacteria. The pituitary gland produces small amounts of somatotrophin, but production in bacteria increases the availability and lowers the cost. Treating cows with somatotrophin increases milk production. In other cases, natural proteins are improved by engineering, Table 1.2. Table 1.2. Examples of proteins engineered for agriculture Glyphosate is a

broad-spectrum, systemic herbicide that acts by inhibiting the enzyme EPSP synthase in aromatic amino acid biosynthesis. Farmers can control weeds by spraying glyphosate on fields, but only if the crops are insensitive to glyphosate. Adding a glyphosate-insensitive variant of EPSP synthase from Agrobacterium created the first glyphosate-resistant soybean plants (Padgette et al., 1995). The next generation of glyphosate-resistant crops contain enzymes such as oxidases or acetyl transferases, which inactivate glyphosate. Protein engineering to increase the catalytic activity of these enzymes is an important goal (Castle et al., 2004). Genetic modification of crops such as cotton, corn, and rice to express insecticidal crystal proteins produced by Bacillus thuringiensis protects the crops against insect pests. The toxicity of these proteins varies toward different insects, and some insects develop resistance to these toxins. Enhancing the toxicity of these proteins by increasing the their binding to target proteins in the insect gut is an essential goal in agricultural biotechnology (Bravo & Soberon, 2008). Heat stress hinders plant growth and lowers crop yields mainly due to reduced photosynthesis. This reduction is mainly due to an inactivation of a thermolabile ATPase, Rubisco activase, whose role is to maintain ribulose-1,5-bisphosphate carboxylase/oxygenase (Rubisco) in its active state. Kurek and coworkers (2007) engineered a thermostable Rubisco activase, which improved the growth of Arabidopsis plants at higher temperatures. Whittle and Shanklin (2001) altered the chain length selectivity of a plant fatty acid desaturase to accept 16-carbon instead of 18-carbon substrate. Lin and Lee (2017) engineered algae to produce more medium chain length fatty acids relative to long-chain-length fatty acids for use in biodiesel production. Engineering a more stable feed additive enzyme increases the nutritional value of feed and reduces pollution. Phosphorus in grains and oilseeds often occurs as phytic acid (myo-inositol hexakisphosphate). Non-ruminants such as pigs and chicken digest phytic acid incompletely leading to high phosphorus in wastewater. The enzyme phytase in animal feed aids digestion of phytic acid by catalyzing its hydrolysis. Forming animal feed pellets requires pressing at high temperatures, so the phytases have been engineered for higher thermal stability to withstand this step. (Yao et al., 2012).

1.4 Engineering proteins for industry Industry uses enzymes for manufacturing and similar non-natural applications, Table 1.3. The three main application areas are detergents, food and beverage, and other, which includes enzymes for biofuel and pharmaceutical synthesis. The advantages of using enzymes over chemical reagents or catalysts are that they are faster, greener and more selective. The primary engineering goals are stability and faster catalysis, both of which lower the cost of the enzyme.

Table 1.3 Protein engineering of industrial enzymes Detergent enzymes such as proteases speed removal of food, blood and other stains on clothing. Subtilisin was the first industrial enzyme to be engineered (Estell et al., 1985). Subtilisin tolerated hot water and surfactants, but bleach rapidly inactivated it, which limited its usefulness. Replacement of an oxidation-sensitive methionine 222 in the active site (Figure 1.2 above) by alanine stabilized subtilisin while maintaining high activity. The cost of these industrial proteins typically $100/kg. In contrast, proteins for medical applications can cost-fold more, $10,000/g, because their manufacture and regulatory approval are more complicated. The largest volume application in food and beverages is the conversion of cornstarch to glucose catalyzed by α-amylase and glucoamylase. Glucose isomerase

can isomerize the resulting glucose to high fructose corn syrup, which tastes sweeter than glucose. This process uses high temperatures to prevent microbial growth and high concentrations to minimize reactor size. Both α-amylase (Richardson et al., 2002) and glucose isomerase (Quax et al., 1991) have been engineered to tolerate high temperatures and high glucose concentration. Glucose isomerase has a half-life = 50-100 d at the operating temperature of 55 °C. More than 500 tons of glucose isomerase produce ca. tons of high fructose corn syrup each year, corresponding to productivities of >10,000 kgs corn syrup per kg enzyme. The use of enzymes for chemical synthesis is known as biocatalysis (Bell et al., 2021). In pharmaceutical manufacture, the most significant application is the penicillin G amidase- catalyzed manufacture penicillin and cephalosporin antibiotics. This process yields more than 10,000 tons of antibiotics annually with a productivity of >600 kgs product/kg enzyme (Bruggink, 2001). These antibiotics are natural products, so they are the natural substrates for this enzymes. However, many other pharmaceuticals are not natural products, so the ability of enzymes to act on them is accidental. Engineering of enzymes to improve their ability to act on unnatural substrates and tolerate the harsh conditions of a chemical reactor is a common goal. Engineered enzymes can also create new biochemical pathways within cells, and the whole cells can be used for synthesis. For example, Ran and Frost (2007) expanded the substrate range of an aldolase to create a new metabolic pathway to make shikimic acid for an influenza drug synthesis. Keasling and coworkers combined enzymes from different organisms and biochemical pathways to create a new biochemical pathway for synthesis of artemisinin, an anti-malarial (Paddon et al., 2013). One reason to use enzymes for manufacture of pharmaceuticals is to decrease the environmental impact of the process. The risk associated with a chemical depends both on how dangerous it is (hazard) and on one's contact with it (exposure). risk = hazard • exposure (1.3) In the past, governments and industry focused on reducing risk by minimizing exposure. Lab coats, safety glasses, and other chemical handling rules limit the exposure of workers to hazardous chemicals. Green chemistry is hazard prevention approach (Anastas & Warner, 2000). Preventing problems is inevitably easier and less expensive than contending with difficulties after they occur. Green chemistry is the design of chemical products and processes that reduce or eliminate hazardous substance. Replacing chemical manufacturing steps with enzyme-catalyzed steps is an essential tool in green chemistry. Enzyme-catalyzed steps often replace hazardous solvents with water and hazardous catalysts with biodegradable enzymes. One example of greener pharmaceutical manufacture is the improved synthesis of sitagliptin, the active ingredient in an oral type 2 diabetes drug. The difficult step is the addition of the amino group with the correct orientation, Figure 1.4. Figure 1.4 Two improved routes to sitagliptin showing the difficult step of inserting the amino group with the correct configuration. Improvements in the chemocatalytic process (rhodium catalyst) won a Green Chemistry award in 2006. Further improvements by switching to a biocatalytic process (transaminase) won a second Green Chemistry award in 2010. Researchers at Merck had already dramatically improved the original synthesis of sitagliptin by using a catalytic asymmetric hydrogenation with a rhodium catalyst. This improved synthesis won a Green Chemistry award in 2006. Biocatalysis further improved this synthesis. Biocatalysis used a transaminase-catalyzed formation of an amine, which eliminated four steps including one that required a high pressure reactor (Savile et al., 2010). It resulted in a 10-13% higher overall yield and 19% less total

waste. This improved synthesis won a second Green Chemistry award in 2010. Enabling this synthesis required extensive protein engineering of the transaminase. The starting transaminase did not react at all with the required substrate and was unstable under the reaction conditions. The engineering replaced 27 amino acids creating a highly active, highly selective and stable enzyme.

1.5 Site-directed mutagenesis Proteins are usually manufactured in bacteria such as Escherichia coli. The gene encoding the protein of interest is placed on a plasmid, which is a circular piece of DNA that replicates independently in bacteria. As the bacteria grow they manufacture of the protein of interest. The plasmid is constructed so that the bacteria manufactured a lot of this protein, so this process is called over expression. Site-directed mutagenesis changes the DNA sequence on this plasmid so the bacteria manufacture a slightly different protein. The most common method of protein engineering is site-directed mutagenesis to replace existing amino acids in a protein with different amino acids, which are predicted 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. Most of the first part of this text focuses on how to predict which substitutions will improve the protein. This section describes how to make the replacement. 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 (Estell et al., 1985). This replacement required replacing the ATG codon for methionine with GCG to encode alanine. 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 … 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 1.5 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 (Hutchinson et al., 1978), 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. Genencor used another method since their experiments were completed before the invention of the polymerase chain reaction. Inverse PCR is also used for other applications such as finding the locations of insertions (https://en.wikipedia.org/wiki/Inverse_polymerase_chain_reaction), which are not discussed in this text. The three features that made this application of inverse PCR a site-directed mutagenesis experiment were the use of 1) a mutagenic primer, 2) back to

back primers, and 3) a circular template. The site-directed mutagenesis method uses an inverse polymerase chain reaction (inverse PCR) to copy the whole plasmid using back-to-back mutagenic primers. New England BioLabs sells kits for this procedure (Q5 Site-directed mutagenesis kit). The mutagenesis involves three steps: design of the mutagenic primers, copying of the whole plasmid using these primers and circularization of the linear product. Copying the whole plasmid avoids the cut and paste steps that complicated older site-directed mutagenesis methods. The polymerase chain reaction copies a section of DNA defined by two DNA primers, Figure 1.6. The DNA primers are single-stranded 20-30 nucleotide long oligomers. One primer is complementary to one DNA strand at the start of the copying; the other primer is complementary to the other DNA strand at the end of the copying. In addition, the primers should have similar melting temperatures and not form dimers with themselves or each other. These primers are synthesized chemically by commercial service. The DNA polymerase extends these primers in the 3' to 5' direction. The result after multiple cycles of copying is a linear double-stranded section of DNA. Figure 1.6 Both normal PCR (top) an inverse PCR on a circular plasmid (bottom) yield linear copies of DNA. In normal PCR, the two primers bind at the start and end of the region to be copied. DNA synthesis proceeds inward toward the other primer. For inverse PCR on a circular plasmid, the two primers anneal to the template DNA back-to-back and DNA synthesis proceeds outward. Since the template is circular, the result is a linear copy of the whole plasmid. To copy an entire plasmid, one can use inverse PCR, Figure 1.6. 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. Adjusting the design of the DNA primers converts the inverse PCR copying to a site-directed mutagenesis. First, 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. Second, 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. Figure 1.7 shows an example of primers designed to replace methionine 222 in subtilisin with alanine. The result of this copying linear, double-stranded copies with a mutation at the desired location.

Figure 1.7 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). Finally, an enzymatic clean-up step circularizes this linear DNA fragment for insertion and replication in bacteria, Figure 1.8. 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. Figure 1.8 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 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. The two main strategies of protein engineering are rational design and directed evolution. Rational design predicts the changes needed to improve the protein based on the protein structure, molecular basis of its action and from computer modeling. In many cases, this approach yields dramatic improvements, but sometimes the needed changes are difficult to predict. In these cases, researchers use directed evolution. Directed evolution generates variants, screens these variants for improved properties and repeats the cycle. The first part of the text describes rational design approaches to protein engineering, while the second part describes directed evolution.

References P. T. Anastas, J. C. Warner (2000) Green Chemistry: Theory and Practice, Oxford U. Press. E. L. Bell, W. Finnigan, S. P. France, A. P. Green, M. A. Hayes, L. J. Hepworth, S. L. Lovelock, H. Niikura, S. Osuna, E. Romero, K. S. Ryan, N. J. Turner & S. L. Flitsch (2021) Biocatalysis. Nat. Rev. Methods Primers, 1, 46. https://doi.org/10.1038/s43586-021-00044-z J. D. Bloom, A. Raval, C. O. Wilke (2007) Thermodynamics of neutral protein evolution. Genetics 175, 255–66. https://doi.org/10.1534/genetics.106.061754 G. B. Bolli, R. D. Di Marchi, G. D. Park, S. Pramming, V. A. Koivisto (1999) Insulin analogues and their potential in the management of diabetes mellitus. Diabetologia 42, 1151-67. https://doi.org/10.1007/s001250051286 J. A. Brannigan, A. J. Wilkinson (2002) Protein engineering 20 years on. Nat. Rev. Mol. Cell Biol. 3, 964–70. https://doi.org/10.1038/nrm975 A. Bravo, M. Soberon (2008) How to cope with insect resistance to Bt toxins? Trends Biotechnol. 26, 573–9. https://doi.org/10.1016/j.tibtech.2008.06.005 A. Bruggink, Ed. (2001) Synthesis of β-lactam antibiotics: chemistry, biocatalysis and process integration, Springer Netherlands; https://doi.org/10.1007/978-94-010-0850-1 J. R. Cantor, T. H. Yoo, A. Dixit, B. L. Iverson, T. G. Forsthuber, G. Georgiou (2011) Therapeutic enzyme deimmunization by combinatorial T-cell epitope removal using neutral drift. Proc. Natl. Acad. Sci. U. S. A. 108, 1272–7. https://doi.org/10.1073/pnas.1014739108 P. Carter,

L. Presta, C. M. Gorman, J. B. B. Ridgway, D. Henner, W. L. T. Wong, et al. (1992) Humanization of an anti-p185HER2 antibody for human cancer therapy. Proc. Natl. Acad. Sci. U. S. A. 89, 4285–9. http://www.pnas.org/content/89/10/4285.long L. A. Castle, D. L. Siehl, R. Gorton, P. A. Patten, Y. H. Chen, S. Bertain et al. (2004) Discovery and directed evolution of a glyphosate tolerance gene. Science 304, 1151–4; https://doi.org/10.1126/science.1096770 D. A. Estell, T. P. Graycar, J. A. Wells (1985) Engineering an enzyme by site-directed mutagenesis to be resistant to chemical oxidation. J. Biol. Chem., 260, 6518–21; http://www.jbc.org/content/260/11/6518.long A. R. Fersht, L. Serrano (1993) Principles of protein stability derived from protein engineering experiments. Curr. Opin. Struct. Biol. 3, 75–83; https://doi.org/10.1016/0959-440X(93)90205-Y S. B. Harvey, M. D. Stone, M. B. Martinez, G. L. Nelsestuen (2003) Mutagenesis of the γ-carboxyglutamic acid domain of human factor VII to generate maximum enhancement of the membrane contact site. J. Biol. Chem. 278, 8363–9.; http://www.jbc.org/content/278/10/8363.long C. L. Hsieh, J. A. Goldsmith, J. M. Schaub, A. M. DiVenere, H.-C. Kuo, K. Javanmardi, K. C. Le, D. Wrapp, A. G. Lee, Y. Liu, C.-W. Chou, P. O. Byrne, C. K. Hjorth, N. V. Johnson, J. Ludes-Meyers, A. W. Nguyen, J. Park, N. Wang, D. Amengor, … J. S. McLellan (2020) Structure-based design of prefusion-stabilized SARS-CoV-2 spikes. Science, 369, 1501–5. https://doi.org/10.1126/science.abd0826 C. A. Hutchison, 3rd, S. Phillips, M. H. Edgell, S. Gillam, P. Jahnke, M. Smith (1978) Mutagenesis at a specific position in a DNA sequence. J. Biol. Chem. 253, 6551–60; http://www.jbc.org/content/253/18/6551.long I. Kurek, T. K. Chang, S. M. Bertain, A. Madrigal, L. Liu, M. W. Lassner, G. Zhu (2007) Enhanced thermostability of Arabidopsis Rubisco activase improves photosynthesis and growth rates under moderate heat stress. Plant Cell 19, 3230–41; https://doi.org/10.1105/tpc.107.054171 M. R. Ladish, K. L. Kohlman (1992) Recombinant human insulin. Biotechnol. Prog. 8, 469-78; https://doi.org/10.1021/bp00018a001 H. Lin, Y. K. Lee (2017) Genetic engineering of medium-chain-length fatty acid synthesis in Dunaliella tertiolecta for improved biodiesel production. J. Appl. Phycol. 29, 2811–9. https://doi.org/10.1007/s10811-017-1210-7 S. M. Loosmore, G. R. Zealey, H. A. Boux, S. A. Cockle, K. Radika, R. E. Fahim, G. J. Zobrist, R. K. Yacoob, P. C. Chong & F. L. Yao (1990) Engineering of genetically detoxified pertussis toxin analogs for development of a recombinant whooping cough vaccine. Infect. Immun. 58, 3653–62. https://doi.org/10.1128/iai.58.11.3653-3662.1990 C. Morrison, R. Lähteenmäki (2018) Public biotech in 2017-the numbers. Nat. Biotechnol. 36, 576–84. https://doi.org/10.1038/nbt.4175 C. N. Pace (1973) The stability of globular proteins. Crit. Rev. Biochem. 3, 1–43; https://www.ncbi.nlm.nih.gov/pubmed/238787 C. J. Paddon, P. J. Westfall, D. J. Pitera, K. Benjamin, K. Fisher, D. McPhee et al. (2013) High-level semi-synthetic production of the potent antimalarial artemisinin. Nature 496, 528–32; https://doi.org/10.1038/nature12051 S. R. Padgette, K. H. Kolacz, X. Delannay, D. B. Re, B. J. LaVallee, C. N. Tinius et al. (1995) Development, identification, and characterization of a glyphosate-tolerant soybean line. Crop Sci. 35, 1451-61; https://doi.org/10.2135/cropsci1995.0011183x003500050032x E. Persson, O. H. Olsen, S. E. Bjørn, M. Ezban (2012) Vatreptacog alfa from conception to clinical proof of concept. Semin. Thromb. Hemost. 38, 274–81; https://doi.org/10.1055/s-0032-1302442 W. J. Quax, N. T. Mrabet, R. G. M. Luiten, P. W. Schunrhuizen, P. Stanssens, I. Lasters (1991) Enhancing the thermostability of glucose isomerase by protein engineering.

Bio/Technology 9, 738–42; https://www.ncbi.nlm.nih.gov/pubmed/1367634 N. Ran, J. W. Frost (2007) Directed evolution of 2-keto-3-deoxy-6-phosphogalactonate aldolase to replace 3-deoxy-D-arabino-heptulosonic acid 7-phosphate synthase. J. Am. Chem. Soc. 129, 6130–9; https://doi.org/10.1021/ja067330p T. H. Richardson, X. Tan, G. Frey, W. Callen, M. Cabell, D. Lam et al. (2002) A novel, high performance enzyme for starch liquefaction. Discovery and optimization of a low pH, thermostable α-amylase. J. Biol. Chem. 277, 26501–7; https://doi.org/10.1074/jbc.M203183200 J. K. Ryu, H. S. Kim, D. H. Nam (2012) Current status and perspectives of biopharmaceutical drugs. Biotechnol. Bioproc. E., 17, 900–11; https://doi.org/10.1007/s12257-012-0095-1 C. A. Sarkar, K. Lowenhaupt, T. Horan, T. C. Boone, B. Tidor, D. A. Lauffenburger (2002) Rational cytokine design for increased lifetime and enhanced potency using pH-activated "histidine switching". Nat. Biotechnol. 20, 908–13; https://doi.org/10.1038/nbt725 C. T. Savile, J. M. Janey, E. C. Mundorff, J. C. Moore, S. Tam et al. (2010) Biocatalytic asymmetric synthesis of chiral amines from ketones applied to sitagliptin manufacture. Science, 329, 305–9; https://doi.org/10.1126/science.1188934 D. M. Taverna, R. A. Goldstein (2002) Why are proteins marginally stable? Proteins Struct. Function Genet. 46, 105–9; https://dx.doi.org/10.1002/prot.10016 E. Whittle, J. Shanklin (2001) Engineering Δ9-16:0-acyl carrier protein (ACP) desaturase specificity based on combinatorial saturation mutagenesis and logical redesign of the castor Δ9-18:0-ACP desaturase. J. Biol. Chem. 276, 21500–5. http://www.jbc.org/content/276/24/21500.full M. Z. Yao, Y. H. Zhang, W. L. Lu, M. Q. Hu, W. Wang, A. H. Liang (2012) Phytases: crystal structures, protein engineering and potential biotechnological applications. J. Appl. Microbiol. 112, 1–14; https://doi.org/10.1111/j.1365-2672.2011.05181.x

Questions 1. Number of possible dipeptides. a) How many possible dipeptides are there? b) If one starts with the dipeptide Asp-Phe, how many single substitution variants are possible? c) How many double substitution variants are possible? d) Show that the sum of wild-type, single substitution variants, and double substitution variants equals the total number of possibilities. Goals Gibbs energies Site-directed mutagenesis Answers 1. If you include the original amino acids, then number of possibilities is 20^2 = 400. This value includes the 361 (19^2) two-substitution variants, 38 single-substitution variants (2 * 19) and one wild-type enzyme.