5 Engineering more stable proteins

Summary. Protein function requires the folded protein form, but this form is unstable mainly because it readily unfolds into a flexible, unstructured form. Protein folding is favored by burying of hydrophobic side chains and hydrogen bonding between the amino acids. Protein unfolding is favored by the increase in flexibility of the main chain of amino acids upon unfolding. Protein stability is usually measured by the reversible unfolding of the protein with either heat or chemical additives like urea. Protein stabilization involves making substitutions that shift the folding-unfolding balance toward the folded form. Stabilizing substitutions can either stabilize the folded conformation or destabilize the unfolded ensemble. This tutorial emphasizes web-based tools to identify substitutions that stabilize proteins. Besides unfolding, other sources of protein instability are chemical modifications like oxidations or cleavage by proteases and aggregation of partly unfolded proteins into insoluble particles.

Key learning goals

- Proteins are dynamic structures that fold and unfold continuously. The primary source protein instability is protein unfolding. Stabilization shifts the folding-unfolding balance toward folding, but does not prevent unfolding.
- To measure protein stability, one measures the ratio of folded to unfolded protein. Under normal conditions, the amount of unfolded protein is too small to detect. Adding denaturants such as urea or heating the sample increases the amount of unfolded protein to measurable amounts. Extrapolation back to normal conditions reveals the stability of the protein.
- One way to stabilize proteins is to restore amino acid residues that are conserved in homologs. The rationale for this approach is evolution conserves residues that provide some benefit; one benefit is a contribution to stability.
- Another way to stabilize proteins is to destabilize the unfolded form. The main driving force for unfolding is the increased flexibility of the unfolded form. Decreasing the flexibility of unfolded form by adding disulfide crosslinks or introducing proline residues often stabilizes proteins.
- The most direct way to stabilize proteins is to create or strengthen attractive interactions between amino acids in the folded conformation. Although proteins will remain dynamic and still unfold, the stronger interactions either slow down unfolding or speed up refolding.

5.1 Introduction

Protein stability usually refers to resistance to unfolding. Stresses like high temperatures, organic cosolvents, high substrate or product concentrations, extremes of pH or high ionic strength can all cause unfolding. In many cases, stabilizing a protein to one stress also stabilizes it to other stresses. For example, proteins that tolerate high temperatures often also tolerate organic

cosolvents.[1]

One advantage of more stable proteins is the ability to use them as biocatalysts in artificial environments for which they have not evolved. For example, using biocatalysis at higher than physiological temperatures yields faster reaction rates, higher substrate solubility, and lower solution viscosities. Reactions at high temperature can also avoid microbial contamination since most contaminating microbes cannot grow at high temperature. Biocatalysts that tolerates organic cosolvents and high concentrations of substrates and products simplify the scale-up of industrial processes by requiring less solvent, smaller equipment, and less complicated product isolation.

The second advantage of more stable proteins is extended lifetime or storage at ordinary temperatures. For example, manufacture of β-lactam antibiotics with penicillin G acylase recycles the immobilized enzyme many times to lower the overall cost. More stable therapeutic proteins last longer during storage and are more resistant to proteases in serum.

A third advantage of more stable proteins is higher yields of soluble protein during manufacture, especially for small, single-domain proteins. Over-expression of recombinant proteins in microbes creates high concentrations of protein. The unfolded forms can aggregate into insoluble particles called inclusion bodies. More stable single-domain proteins aggregate less and yield higher amounts of soluble protein.[2]

A fourth advantage is that more stable proteins are also more 'evolvable' or able to acquire beneficial traits. Substitutions that change protein function may destabilize proteins. If the destabilized protein fails to fold, then the potential improvement is lost. More stable proteins can tolerate greater numbers of destabilizing mutations and are thus more evolvable than their marginally stable variants.[3]

One source of stable proteins is thermophiles and other extremophiles.[4] For example, the *Pfu* DNA polymerase used for the polymerase chain reaction (PCR) comes from the hyperthermophile *Pyrococcus furiosus*, Figure 5.1. This polymerase catalyzes DNA synthesis at 72 °C and tolerates the high temperatures (95 °C) needed to dissociate complementary DNA strands.

Figure 5.1 The archaeon *Pyrococcus furiosus* is a hyperthermophile that grows best at 100 °C. The heat-stable *Pfu* DNA polymerase used in PCR comes from this microbe. This painting simulates a scanning electron micrograph. Image by Fulvio314 from en.wikipedia.org/wiki/Pyrococcus_furiosus (CC BY-SA 3.0).

One disadvantage of enzymes from thermophiles is the typically low catalytic activity of these

enzymes at room temperatures.[5] Since thermophiles do not live at room temperature, there is no selective pressure for high activity at room temperature. If the application requires activity at room temperature, enzymes from thermophiles may not be suitable. *Pfu* DNA polymerase has negligible activity at room temperature. Other limitation of enzymes from thermophiles is the lack of other needed protein characteristics. For example, applications may require unusual substrate specificity unavailable in thermophiles or human proteins to minimize immune response.

Many human disease-causing mutations are associated with amino acid substitutions that decrease in protein stability.[5a] For example, mutations that decrease the stability of fructose bisphosphate aldolase cause hereditary fructose intolerance and mutations of the tumor suppressor protein p53 cause some cancers.[5b] The ability to predict substitutions that alter protein stability may help identify mutations that cause genetic diseases.

5.2 Native and denatured conformations equilibrate

The native protein state, N, is the folded, functional form. It is compact with the hydrophobic side chains mainly buried and the polar side chains mainly exposed to solvent. It has a specific structure (or similar set of structures), typically with α-helices, β-sheets, and turns folded in a particular arrangement, Figure 5.2. While the structure is dynamic and moves, it has an overall stable structure. Most of the amino acids interact with each other; only amino acids on the protein surface interact with solvent water. In contrast, the denatured protein state, D, is not functional and is not a single state, but a collection or ensemble of more or less folded states. The main chain makes large, random motions and the amino acids interact mainly with solvent water, not with each other.

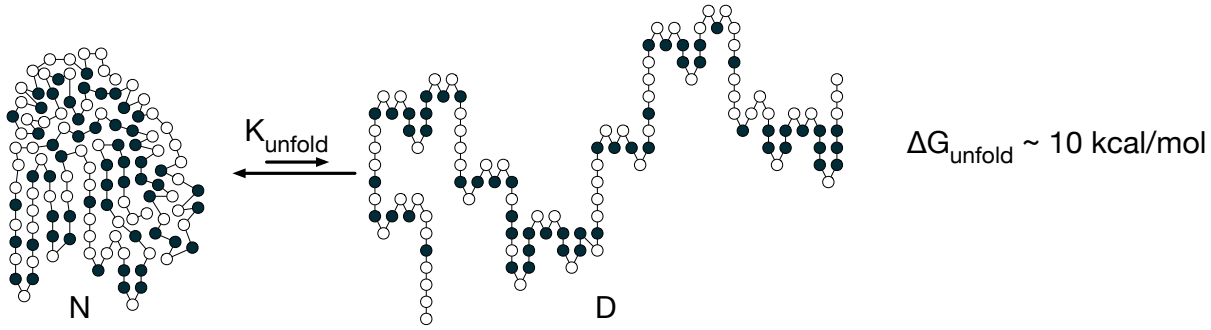

Figure 5.2 Schematic of protein unfolding showing the equilibrium between a compact folded native structure (N) and an ensemble of flexible unfolded states (D). Folding creates a specific structure that mostly buries hydrophobic amino acids (filled circles) in the interior of the folded structure while mostly exposing hydrophilic amino acids (open circles). Unfolding exposes more the amino acids to solvent. For clarity, the figure shows only one example unfolded structure, but in reality it is an ensemble of rapidly interchanging unfolded structures.

The native form of a typical protein is more stable than the denatured state by ~10 kcal/mol, Figure 5.3. A Gibbs energy of unfolding, ΔG^{unfold}, of +10 kcal/mol corresponds to an equilibrium constant $e^{-(10{,}000/RT)}$ or 4.6×10^{-8} at room temperature indicating that unfolding is rare, eq. 5.1. Folding and unfolding are fast for single domain proteins. Native forms continuously unfold to denatured states and these continuously refold to native forms.

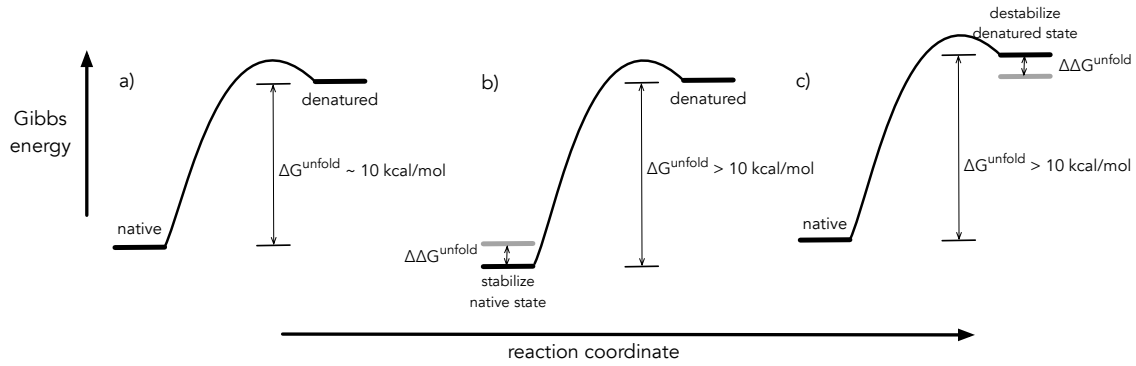

Figure 5.3 The Gibbs energy difference between the folded, native state of a protein and the denatured ensemble of states determines protein stability. The barrier to unfolding for single domain proteins is low. a) The denatured state for the original protein lies ~10 kcal/mol above the native state. b) Stabilization of the native form stabilizes the protein because it increases the energy difference between the native and denatured forms. c) Destabilization of the denatured form also stabilizes the protein because it increases the energy difference between the native and denatured forms.

$$\Delta G^{unfold} = -RT\ln\left(K^{unfold}\right) \text{ or } K^{unfold} = e^{-\Delta G^{unfold}/RT} \tag{5.1}$$

Protein solutions contain unfolded protein molecules even for stably folded proteins. In the case where the native form is 10 kcal/mol more stable than the unfolded form, one in twenty-two million protein molecules unfolds at any given time in solution at room temperature, eq. 5.2. This tiny fraction is too small to detect by spectroscopic methods, but it is nevertheless a large number of molecules. A solution containing 1 mg of a 30 kDa protein contains 2×10^{16} molecules; of these, nearly a billion (10^9) are in the denatured form at any time.

$$K^{\text{unfold}} = \frac{[D]}{[N]} = 4.6\times10^{-8} \text{ or } [N] = 22\times10^{6}\cdot[D] \tag{5.2}$$

The dominant force driving protein folding: the hydrophobic effect. The main driving force for protein folding is the hydrophobic effect, which is the tendency of non-polar solutes to cluster in water. Burying a $-CH_2$ group contributes 1.1 ± 0.5 kcal/mol to protein stability. The hydrophobic effect provides ~60% of the driving force to collapse the amino acid chain into a compact structure.[6] The folded form of a protein is its lowest energy conformation in dilute solutions. (In concentrated protein solutions, oligomeric or aggregated states may be

lower in energy.) The chains orient to maximize the hydrophobic effect and also to make attractive interactions between amino acids: hydrogen bonds and other electrostatic interactions.

The hydrophobic effect is not a bond between non-polar solutes, but rather the result of the favorable release of water molecules from the smaller hydrophobic surface upon clustering. In bulk water, individual water molecules form a network of hydrogen bonds with each other, Figure 5.4. These hydrogen bonds exchange rapidly, and the water molecules move rapidly. In contrast, water molecules at the non-polar surface make fewer hydrogen bonds. These remaining hydrogen bonds are stronger than those in bulk water, making these water molecules at the non-polar surface less mobile. They form an ice-like cage structure around the non-polar solute. The restricted mobility of the water molecules in this cage decreases their entropy making this arrangement less stable than bulk water. Clustering of non-polar solutes reduces the contact area between water and non-polar surface, releases water molecules from the cage-like structure and lowers their energy. Weak van der Waals interactions between the non-polar atoms also contribute a small amount to the hydrophobic effect.

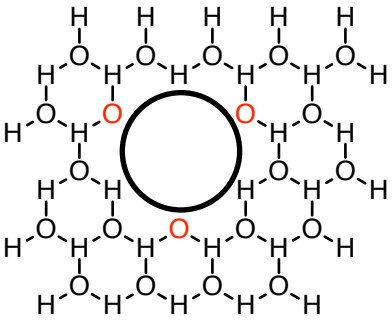

Figure 5.4 Water surrounds hydrophobic solutes with an ice-like structure. In this 2-D diagram of a hydrophobic circle in water, three oxygens near the circle (red color) have only two hydrogen partners, while the rest of the oxygen atoms have three hydrogen partners. Fewer partners strengthen the remaining bonds leading to a rigid, ice-like structure at the hydrophobic interface. This 2-D diagram simplifies the true 3-D structure, where each oxygen atom has four hydrogen partners (two covalent bonds, two hydrogen bonds) in the bulk water and two or three hydrogen partners (two covalent, zero or one hydrogen bonds) in the water at the hydrophobic interface.

Hydrogen bonds contribute the remaining ~40% of the driving forces for protein folding. The hydrogen bonds between protein atoms each contribute on average 1.1 ± 0.8 kcal/mol each to protein stability. The unfolded protein makes hydrogen bonds between protein and water, so this energy contribution is the net gain in hydrogen bond strength. Hydrogen bonds also define how the protein will fold into helices and sheets.

The dominant force driving unfolding of the main protein chain: flexibility. The denatured ensemble is a collection of highly flexible forms. This flexibility creates many conformations or micro-states and the existence of these states is the main driving force for unfolding. The micro-states create a favorable entropy contribution, $-T\Delta S$, to the Gibbs energy. To find the Gibbs energy contribution of entropy one multiplies by temperature and a negative sign since an increase in entropy lowers Gibbs energy, eq. 5.3, where W_1 and W_2 are the different numbers of micro-states in the states being compared.

$$\Delta G_{2\text{-}1} = -T^*\Delta S_{2\text{-}1} = -RT\ln(W_2/W_1) \quad (5.3)$$

To find the effect of a molecule's conformational flexibility on Gibbs energy, one compares the number of conformations in the flexible state to the non-flexible state. For example, one can estimate the difference in Gibbs energy between the folded and unfolded states for one amino acid in a protein at 300 °K only due to differences in backbone flexibility while ignoring any differences in side-chain flexibility. Assume that the backbone has a single conformation in the folded state, N, but can adopt three staggered conformations along each of the two rotatable bonds (ψ, ϕ) in the unfolded state, D, Figure 5.5.

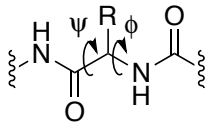

Figure 5.5 Each amino acid residue contains two rotatable bonds (ψ, ϕ) along the main chain. The C–N bond in the amide link does not readily rotate because it has partial double bond character.

The approach is to estimate the difference in the number of conformations available to the folded and unfolded states and then convert this difference to Gibbs energy using eq 5.3 above. The amino acid in the folded state has one available conformation while in the unfolded state, an amino acid can adopt 3*3 = 9 possible conformations. Converting this difference in possible conformations into Gibbs energy yields:

$$\Delta G_{N-D} = -T^*\Delta S_{N-D} = -RT\ln(W_f/W_u) = -300\ °K \cdot 1.987\ cal/°K{\cdot}mol \cdot \ln(1/9) = +1.3\ kcal/mol \quad (5.4)$$

Thus, entropy due to differences in the backbone flexibility favors the unfolded state by 1.3 kcal/mol for each amino acid residue. More accurate estimates that account for the unequal likelihood of the nine conformations and differences in side chain flexibility yield a similar number: 1.4 kcal/mol per amino acid residue.[7]

For a typical protein of 300 amino acids, this entropy effect of backbone flexibility contributes ~400 kcal/mol, which is large as compared the balance of ~10 kcal/mol in favor of the folded state. Hydrophobic interactions and hydrogen bonds in the native state counterbalance this flexibility advantage of the denatured state with a slightly larger Gibbs energy contribution. Protein stability is a balance between large forces favoring either the folded or unfolded states.

5.3 Measuring the folding-unfolding equilibrium

Measuring the equilibrium constant requires measuring the ratio of folded and unfolded protein at equilibrium. Under normal conditions, the equilibrium amount of unfolded protein is too small to measure. Instead, researchers use destabilizing conditions where easily detectable amounts of both the folded and unfolded protein exist, Figure 5.6. Typical destabilizing conditions are additives like urea or guanidinium chloride, changes in the solution pH, or increases in temperature. After measuring the equilibrium constant at these destabilizing

conditions, researchers extrapolate back to normal conditions to get the desired equilibrium constant under normal conditions.

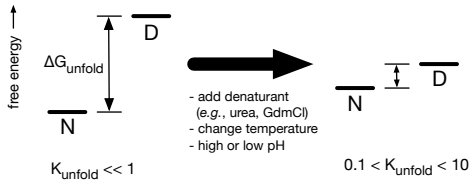

Figure 5.6 Changing the environment shifts the equilibrium between native and denatured states. The equilibrium constant can be measured when the relative amounts of the two forms are in the range of 0.1: 1 to 10:1, that is, when $|\Delta G_{unfold}| < \sim 1.4$ kcal/mol. GdmCl = guanidinium chloride.

Most of this tutorial treats protein folding and unfolding as a cooperative, two-state process (Fig. 5.2 above). The entire protein is either folded or denatured, and the process is entirely reversible. There are no partially folded proteins or folding intermediates. The cooperative folding might start with one amino acid adopting an α-helix conformation, quickly followed by neighboring amino acids adopting the α-helix structure to create intrachain hydrogen bonds. This assumption of a simple, reversible two-state process is reasonable for single domain monomeric proteins. In these cases, adding stabilizing substitutions anywhere in the protein can contribute to stabilization since the entire protein unfolds simultaneously.

Changing the environment does not gradually change protein structure from folded to unfolded. Proteins remain fully folded or fully unfolded, but their ratio changes with the environment. Unfolding is cooperative because the interactions that stabilize the folded protein are stronger in combination than their individual contributions. The protein remains folded even when some stabilizing interactions break, but breaking too many destabilizes the others leading to complete unfolding of the protein (Horovitz & Fersht, 1992).[7a]

Section 5.5 below will briefly consider multiple domain and oligomeric proteins. Multiple domain proteins usually fold and unfold stepwise via intermediates. The same stabilization strategies apply, but the substitutions must be in unfolded regions, not anywhere in the protein.

5.3.1 Denaturation with urea, ΔG_{H_2O}

Urea unfolds proteins because it 1) solvates exposed hydrophobic groups to reduce the hydrophobic effect and 2) forms hydrogen bonds to the backbone to disrupt secondary structures. Typical proteins unfold at 3-5 M urea; the solubility limit is ~18 M urea at room temperature.

The urea-induced denaturation experiment involves preparing solutions of protein in various concentrations of urea, waiting until the folding-unfolding reaches equilibrium, and measuring the amounts of native and denatured protein. The most common method to detect the folded and unfolded proteins is measuring the intensity of protein absorbance or fluorescence, Figure 5.7, but any method that distinguishes between the native and denatured states is suitable. Some examples are measuring shifts in the wavelength of the fluorescence maximum, changes in the circular dichroism spectra, changes in the NMR spectra, decreases in catalytic activity, or

changes in the dye fluorescence upon binding to hydrophobic regions of the unfolded protein.

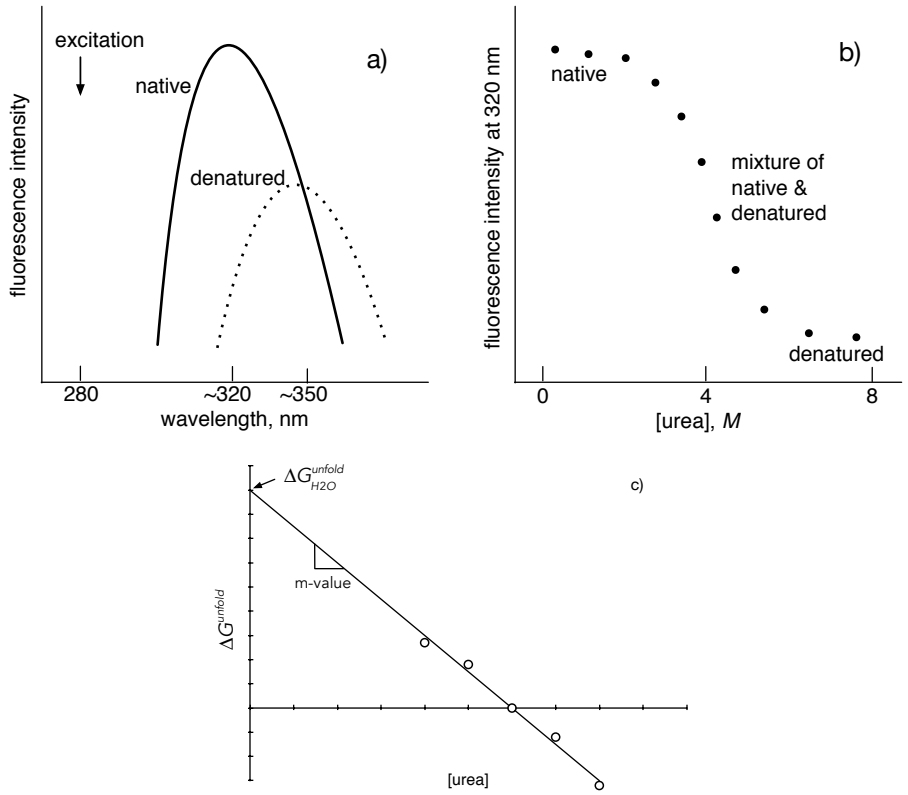

Figure 5.7 Unfolding of a protein in urea as monitored by tryptophan fluorescence. a) The indole ring of tryptophan absorbs light (absorbance maximum ~280 nm; not shown), then emits light at lower energy (higher wavelengths). The emission maximum of a folded protein is typically ~320 nm, which is similar to the emission maximum of indole in hexane. Upon unfolding the emission maximum shifts to ~350 nm, which is similar to the emission maximum of tryptophan in water. b) Hypothetical urea unfolding data showing a decrease in fluorescence intensity at 320 nm as the protein unfolds in increasing concentrations of urea. c) Linear extrapolation of the Gibbs energies calculated from the data in panel b reveals the Gibbs energy of unfolding in water.

The reasoning above predicts that the fluorescence emission wavelength increases for the denatured protein, but it is not easy to predict the relative intensity of this emission. The intensity change also depends wavelength chosen to monitor the fluorescence. For the example in Figure 5.7b, the fluorescence intensity at 320 nm (near the emission maximum of the native protein) decreases as the protein unfolds, but the fluorescence intensity at 360 nm (near the emission maximum of the denatured protein) increases (not shown).

To extract an equilibrium constant from the fluorescence changes, one needs to know the fractions of native, F_N, and denatured protein, F_D. The ratio of these fractions is the equilibrium constant, eq. 5.5. One assigns the fluorescence in low or no denaturant, Y_N, to the native protein and the fluorescence at high denaturant concentrations, Y_D, to the denatured protein. Intermediate values of fluorescence, Y_{obs}, correspond to a mixture of native and denatured

states. Only data in the transition region of the denaturation experiment contribute to the analysis.

$$N \overset{K^{unfold}}{\rightleftharpoons} D$$

$$K^{unfold} = \frac{F_D}{F_N} = \frac{Y_{obs} - Y_N}{Y_D - Y_{obs}} \tag{5.5}$$

The negative of the natural logarithm of the equilibrium constant multiplied by the gas constant and temperature yields the Gibbs energy of unfolding under the measurement conditions, eq. 5.1 above. To measure protein stability, researchers measure the equilibrium constant between native and denatured protein in increasing concentrations of a denaturant, for example, see Figure 5.7b above. These equilibrium constants at different urea concentrations are converted to Gibbs energies according to eq 5.1. A plot of Gibbs energy on the y-axis versus denaturant concentration on the x-axis is a straight line, eq 5.6. Extrapolation of this line to [urea] = 0 yields ΔG^{unfold} in pure water or ΔG^{unfold}_{H2O} as the y-intercept. This linear extrapolation approach[8] is a convenient way to measure protein stability.

$$\Delta G^{unfold} = \Delta G^{unfold}_{H2O} - m \cdot [urea] \tag{5.6}$$

Positive Gibbs energies of unfolding in pure water indicate that unfolding of the protein is unfavorable. A higher Gibbs energy of unfolding indicates a more stable folded protein.

The molecular basis for the linear relationship between unfolding Gibbs energy and urea concentration may be a binding interaction between urea and protein. The binding of urea destabilizes the protein and the fraction of urea bound increases linearly with concentration. An example of this linear extrapolation approach is in the supporting information.

The slope of the line for the linear extrapolation plot, eq 5.6, is negative since adding urea destabilizes the protein. The m-value is the absolute value of this slope of the line and indicates the sensitivity of the protein to denaturant. The m-value depends on the solvent-accessible surface area exposed upon unfolding.[8] A protein that unfolds completely has a higher m-value than a similar protein that unfolds only partially. Similar proteins with different m-values indicate different structures of the unfolded, denatured state. Typical m-values for urea-induced unfolding are ~1000 cal/(mol · M).

5.3.2 Heat denaturation, $\Delta G^{unfold}(T)$

Heat also unfolds proteins. Increases in temperature increase the entropy contribution to Gibbs energy of unfolding, eq. 5.7 where ΔH^{unfold} and ΔS^{unfold} are respectively the changes in enthalpy and entropy upon unfolding. The entropy contribution due to the increase in flexibility of the unfolded protein increases at higher temperature and eventually leads to unfolding.

$$\Delta G^{unfold} = \Delta H^{unfold} - T\Delta S^{unfold} \tag{5.7}$$

Here we consider reversible heat-induced unfolding. Section 5.6.1 below considers irreversible unfolding caused by aggregation of the unfolded protein. As with urea-induced denaturation, this unfolding can be monitored spectroscopically, Figure 5.9. For reversible unfolding, the equilibrium constant yields the Gibbs energy of unfolding at these high temperatures. The melting temperature, T_m, is the temperature where the amounts of folded and unfolded protein are equal; $\Delta G^{unfold} = 0$.

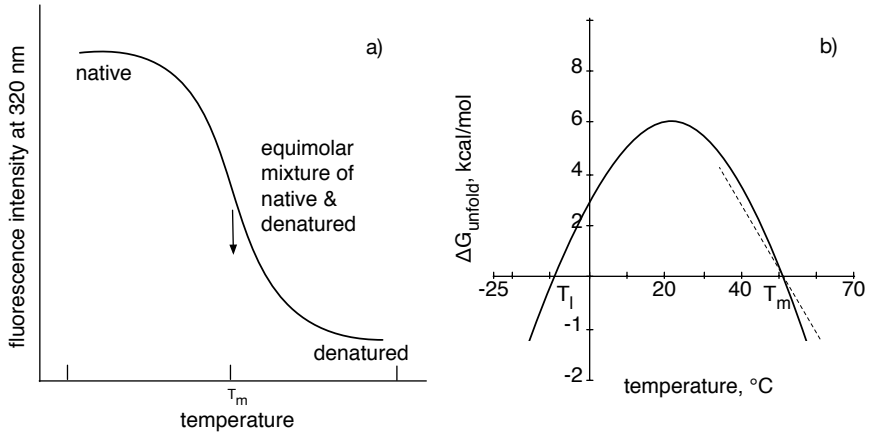

Figure 5.9 Heat induced unfolding of proteins. a) Hypothetical melting curve for a protein measured by the decrease in fluorescence intensity at 320 nm as the protein unfolds. The melting temperature, T_m, is the temperature where the amounts of folded and unfolded protein are equal. b) The Gibbs energy of unfolding of a protein varies non-linearly with temperature according to eq. 5.15. The stability of a protein also decreases at lower temperatures; T_l marks the the low temperature unfolding temperature. This plot uses values typical for a 300 aa protein: $T^0 = 300$ °K, $\Delta H^0 = 30$ kcal/mol, $\Delta S^0 = 80$ cal/mol · deg, $\Delta C_p = 4.2$ kcal/mol · deg. The dotted line is an approximation sometimes used near the melting temperature, see eq. 5.11.

These measured Gibbs energies of unfolding at high temperatures can be extrapolated to room temperature, but this extrapolation is more complex than extrapolating the Gibbs energy of unfolding to pure water from a urea denaturation curve.[9] One might expect a plot of ΔG^{unfold} versus temperature to yield a straight line according to eq. 5.7 above, where ΔS^{unfold} is the slope and ΔH^{unfold} is the y-intercept. However, the Gibbs energy of protein unfolding varies with temperature according to a more complex equation, eq. 5.10.[10] Extrapolating the melting curve data to lower temperatures also requires measuring ΔC_p which is the change in heat capacity upon unfolding, Figure 5.9.

$$\Delta G^{unfold} = \Delta H^{unfold0} - T\Delta S^{unfold0} + \Delta C_p[(T - T^0) - T\ln(T/T^0)] \qquad (5.10)$$

Here T^0 is the reference temperature and $\Delta H^{unfold0}$ and $\Delta S^{unfold0}$ are the enthalpy and entropy of unfolding at that temperature. For reactions involving small molecules, the heat capacity of starting materials and products are similar, and differences may be ignored. (Eq. 5.10 simplifies to eq. 5.7 when $\Delta C_p = 0$.) However, the heat capacities of the folded and unfolded proteins differ

significantly. Another way to state the reason for the more complex equation is that for protein unfolding ΔH^{unfold} and ΔS^{unfold} vary with temperature while eq. 5.11 assumes they are constant. The extra terms in eq. 5.15 account for the variation of ΔH^{unfold} and ΔS^{unfold} with temperature.

The unfolded protein has a higher heat capacity than the folded protein because unfolding exposes hydrophobic groups to water. Water surrounds these hydrophobic groups with an ice-like structure, which has slightly higher energy than bulk water, see Figure 5.4 above. The fluctuation of water between bulk and ice-like structure increases the heat capacity.

The change in heat capacity for protein unfolding, ΔC_p, is always positive with a typical value of ~14 cal/mol · deg per residue or 4.2 kcal/mol · deg for a 300 aa protein.[11] Differential scanning calorimetry can measure this change in heat capacity. Differential scanning calorimetry measures the heat required to increase the temperature of a sample. The baseline value before unfolding reveals the heat capacity of the folded protein, while the baseline value after unfolding reveals the heat capacity of the unfolded protein. The difference is the change in heat capacity.

The parabolic curve in Figure 5.9b predicts that proteins will also denature at low temperatures. For most proteins the predicted cold denaturation temperature, T_l, lies well below the freezing point of water as shown in the sketch, but for some proteins, the cold denaturation temperature lies in the liquid range of water. For example, the yeast protein frataxin denatures at both 7 °C and at 30 °C.[12] At high temperatures the increase in conformational entropy (flexibility) of the protein atoms drives unfolding, while at low temperatures a decrease of the hydrophobic effect drives unfolding. At lower temperatures, the cost of ordering water molecules around a hydrophobic group decreases. As the tendency of hydrophobic groups to aggregate in water decreases, the folded form loses a stabilizing force.

Changes in melting temperature, ΔT_m. Comparing the unfolding or melting temperatures of protein variants is a common, approximate measure of their stability. One assumes that the unfolding process is similar for protein variants so the parabolic curve of Gibbs energy of unfolding versus temperature (Fig. 5.9b above) remains similar. One heats a protein sample slowly while monitoring the protein tryptophan fluorescence (Fig 5.9a). The temperature corresponding to the midpoint of the change in fluorescence indicates the melting temperature. A comparison of pairs of homologous proteins from thermophiles and mesophiles showed that the melting temperatures, T_m, were 31.5 °C higher and the Gibbs energies of unfolding 8.7 kcal/mol higher for the thermophiles.[13] From this comparison, we can derive a rule of thumb that stabilizing a protein by 1 kcal/mol will increase the melting temperature by ~3.6 °C, eq. 5.11, see dotted line in Figure 5.9b.

$$\Delta\Delta G_{unfold}\,(\text{kcal/mol}) \sim \Delta T_m\,(°C)/3.6°C/kcal/mol \qquad (5.11)$$

In many cases, the Gibbs energy of heat-induced unfolding, $\Delta G^{unfold}(T)$, is comparable to that obtained by urea unfolding. When different experiments yield different Gibbs energies of

unfolding for the same protein, it is most likely because the unfolded states differ in the two experiments.

5.4 Stabilization to cooperative unfolding

Stabilizing a dynamic object like a protein differs from stabilizing a rigid object like a chair. To stabilize a chair one can add a supporting brace, which keeps the chair in the desired structure. In contrast, a protein folds and unfolds continuously; it is dynamic. Adding a stabilizing interaction to a protein does not prevent unfolding. Instead, the stabilizing interaction shifts the folding-unfolding equilibrium toward folding so the protein spends more time in the folded form. Since the equilibrium constant is the ratio of the forward and reverse rate constants, eq. 5.12, shifting the equilibrium constant also changes the rates of folding and unfolding. A stabilizing interaction may slow down unfolding, speed up folding, or both; but, unlike stabilizing a chair, stabilizing a protein does not *prevent* unfolding.

$$\mathrm{N} \underset{k_{fold}}{\overset{K_{unfold}}{\rightleftharpoons}} \mathrm{D} \qquad K_{unfold} = \frac{k_{unfold}}{k_{fold}} \tag{5.12}$$

Stabilizing a protein also differs from stabilizing a rigid object because one can stabilize a protein by destabilizing the unfolded form.[14] Everyday objects like chairs are not dynamic and cannot be stabilized this way. This book uses the expression 'stabilize a protein' or 'increase protein stability' to mean either stabilizing the folded form or destabilizing the unfolded form.

One reason that it is hard to predict which substitutions will stabilize a protein is that the structure of the denatured state (an ensemble of rapidly equilibrating structures) is unknown. Stabilizing substitution must affect the folded and unfolded forms differently. A substitution that equally stabilizes (or destabilizes) the folded and unfolded forms does not change stability of the protein. X-ray crystal structures reveal changes to the folded form, but changes to the denatured state are invisible because its structure is unknown. Stabilization can occur without strengthening the folded structure. For example, seven amino acid substitutions dramatically stabilized xylanase (T_m increased by 25 °C), but the structures of the wild-type and stabilized proteins showed similar interactions between amino acids in both structures.[15] In this case, the stabilizing substitutions likely destabilized the unfolded state and changed its structure, but nature of these changes is unknown because the structure of the unfolded state is unknown.

A second reason that it is hard to predict stabilizing substitutions is that the substitutions must not hinder catalytic or binding ability of the target proteins. Substitutions that increase stability often also decrease activity and vice versa.[16] One reason for this trade-off is that catalysis and binding rely on interactions of the substrate or transition state with unsatisfied hydrogen bonds, exposed hydrophobic groups and unpaired charges in the protein. Satisfying these interactions with amino acid substitutions will stabilize the protein, but also disrupt binding and catalysis. Researchers typically avoid substitutions near the active site when searching for stabilizing substitutions to prevent disrupting binding and catalysis.

Some protein stabilization strategies seek to stabilize the native protein, others destabilize the denatured conformation, and, for a few, the molecular mechanism is unknown, Table 5.1. The design strategies that seek to stabilize the native protein require a protein structure or at least a homology model to start the design. A homology model is an extrapolated 3D-structure of a protein based on the known structure of a similar protein (https://swissmodel.expasy.org). Strategies that seek to destabilize the unfolded ensemble also require a protein structure to ensure that the changes do not distort the folded form. Strategies with unknown mechanism do not require a protein structure. For example, restoring residues that are conserved in homologs (consensus sequence approach, see below) requires only a protein sequence. Random mutagenesis followed by screening to find stabilized variants also does not require a protein structure, but this tutorial does not discuss random mutagenesis methods.

Table 5.1. Design strategies to stabilize proteins and web tools to implement these strategies.

strategy	rationale	example of implementation
restore residues conserved in homologs[a]	not specified	Consensus finder identifies conserved amino acids in homologs that are missing from target protein
add disulfide links	destabilize unfolded protein	SS bond identifies suitable locations, additional analysis needed to narrow choices
add Pro residues	destabilize unfolded protein	Analysis of structure to identify locations suitable for proline (PROMOTIF)
substitutions in or near flexible regions	stabilize folded protein	Random mutagenesis in or near flexible regions (B-FITTER); or molecular dynamic modeling to identify missing interactions
random mutagenesis & screening	not specified	Random mutagenesis anywhere followed by screening for stabilized variants
optimize electrostatic interactions	stabilize folded protein	TKSA-MC identifies destabilizing electrostatic interactions
optimize many interactions	stabilize folded protein	Modeling with Rosetta or FoldX to identify stabilizing substitutions

[a]Bioinformatics strategy; stabilization rationale not specified, see section 5.4.

To engineer a change, one needs to choose both the location of the substitution and the replacement amino acid. The 'restore residues conserved in homologs' approach predicts both, so it is easy to implement. The 'substitutions in or near flexible regions' approach defines the location, but not the replacements. Additional modeling or experiments are needed to identify the replacements. The 'add disulfide links' approach specifies that cysteine is the replacement amino acid, but additional modeling is needed to find suitable locations. Design strategies like 'improve hydrophobic packing' are the least specific and require computer modeling to predict both the location and the replacement amino acid. Each of the approaches in Table 1 will be described in more detail below. The effect of each stabilizing mutation is typically small (~0.5 kcal/mol), so substantial stabilization of the protein (2-4 kcal/mol) requires multiple substitutions and often more than one design strategy.

5.4.1 Restore conserved amino acids

Evolution conserves amino acids that contribute to protein function. This contribution may be to structure, catalysis, stability or other protein property. The consensus approach hypothesizes that conserved amino acids residues are more likely to contribute to protein stability than a non-conserved amino acid. To use the consensus approach to stabilize a protein, one identifies amino acid residues conserved within homologous proteins, but which are missing in the target protein. Restoring these conserved amino acids is more likely to stabilize the protein than random substitutions.

Consensus design works because natural proteins rarely follow the consensus sequence at all structurally important positions. Natural proteins only have to be stable enough to fulfill their biological function; proteins with stabilities above a certain threshold will have no further selection advantage. There are many stabilizing residues, but individual proteins do not need all of them to reach the threshold of being stable enough. Thus, replacing a residue with the corresponding consensus amino acid may improve the stability or folding efficiency of a protein of interest. The same reasoning leads to the conclusion that the stability of a protein designed by consensus can be higher than that of the proteins in the alignment.

The consensus sequences approach does not hypothesize any molecular basis for the stabilization. It relies only on amino acid sequence comparisons from bioinformatics and does not require any structural information.[17,17a] It is easy to implement, and several web-based tool to generate a consensus sequence for a protein are available: Consensus Finder[17b] or FireProt.[17c]

Typically researchers restore one or a few conserved amino acids. Usually at least half of these predicted substitutions prove to be stabilizing. Pantoliano and coworkers first suggested that substitutions to the consensus amino acid may stabilize proteins and showed that the consensus-type mutation Met50Phe increased the unfolding temperature of subtilisin BPN′ by 1.8 °C[18] and many others have used this approach.

One can also restore all the conserved amino acids simultaneously. Lehmann and coworkers predicted the consensus sequence of fungal phytases based on the sequence alignment of thirteen sequences, Table 5.2.[17] The consensus phytase differed by at least 80 amino acids from the parent sequences. A synthetic gene encoded the consensus phytase containing all these changes. The consensus protein was 15–26°C more thermostable than any of its parents.

Table 5.2. Part of the sequence alignment (amino acid positions 54-73) of fungal phytases and the derived consensus sequence.[17] Amino acid sequences from the same species are weighted so that each species contributes equally to the consensus sequence. The '-' in the consensus sequence indicates that the consensus amino acid is ambiguous at this position.

source	weight	sequence
Aspergillus terreus 9A-1	0.50	54- QVLARHGARS PTHSKEKAYA
Aspergillus terreus cbs	0.50	QVLARHGARS PTDSKTKAYA
Aspergillus niger var. *awamori*	0.33	QVLSRHGARY PTESKGKKYS
Aspergillus niger T213	0.33	QVLSRHGARY PTESKGKKYS
Aspergillus niger NRRL3135	0.33	QVLSRHGARY PTDSKGKKYS
Aspergillus fumigatus 13073	0.20	QVLSRHGARY PTSSKSKKYK
Aspergillus fumigatus 32722	0.20	QVLSRHGARY PTSSKSKKYK
Aspergillus fumigatus 58128	0.20	QVLSRHGARY PTSSKSKKYK
Aspergillus fumigatus 26906	0.20	QVLSRHGARY PTSSKSKKYK
Aspergillus fumigatus 32239	0.20	QVLSRHGARY PTASKSKKYK
Emericella nidulans	1.0	QVLSRHGARY PTESKSKAYS
Talaromyces thermophilus	1.0	QLLSRHGARY PTSSKTELYS
Myceliophthora thermophilia	1.0	QVLSRHGARA PTLKRAASYV
Consensus		**QVLSRHGARY PTSSK-KAYS**

At some positions, counting the numbers of occurrences identifies the consensus amino acid. For example, all thirteen sequences start with QVL, and most have S in the next position (boxed), so the consensus sequence begins with QVLS, Table 2. At other positions, identifying the consensus amino acid requires care to ensure unbiased sampling of sequences. Identifying the most common residues requires a uniform sampling of amino acid sequences throughout the evolutionary tree, but the available amino acid sequences may not be evenly distributed. For the phytase example above, researchers reduced the bias of multiple sequences from different strains of the same species by weighting the amino acid sequences so that each *Aspergillus* species had equal weight. For example, the list included five sequences from *Aspergillus fumigatus* strains, Table 2. Each of these sequences was weighted by 0.2 in deriving the consensus to avoid bias toward this species. For example, consider the last amino acid in Table 2 (position 73, boxed). The first two sequences (*A. terreus*) have A at this position, the next three (*A. niger*) and two others (*Emericella* and *Talaromyces*) have S, the five *A. fumigatus* sequences have K, and the last sequence (*Myceliophthora*) has V. Although S and K both occur five times, K occurs only in the five *A. fumigatus* sequences and receives a total weight of 1.0, while S occurs in three different species and gets a total weight of 3.0. The consensus amino acid at this position is S. The consensus sequence is not unique, but depends on the set of comparison sequences. In later work, as more phytase sequences became available, researchers tested a revised consensus sequence, which further increased stability.

In other cases, a database search may yield thousands of similar sequences. To minimize phylogenetic bias in these cases, one clusters similar sequences (typically >90% identity) together and uses only one sequence from each cluster for the consensus calculation. The web-based tools

to find consensus sequences mentioned above use this clustering to minimize phylogenetic bias.

Ancestral proteins are extinct proteins that correspond to branch points in a phylogenetic tree. Gene synthesis and expression of these proteins in microbial hosts resurrects these proteins. In many cases, ancestral proteins are more stable than modern proteins. Some ancestral proteins may be more stable because the earth was warmer in prehistoric times. However, ancestral proteins from times when the earth's temperature was only slightly warmer are also more stable, so other effects may contribute.[19] One contribution is that reconstructed ancestral proteins are often similar to the consensus sequence for that group since many descendants retain the ancestral residues. However, ancestral proteins differ from consensus proteins is two ways. First, ancestral proteins identify amino acids conserved within a related group, while consensus proteins include data from all homologous proteins, including those outside a group. Second, the consensus approach weights all sequences equally, but ancestral sequence reconstruction takes branch lengths into account, so more recent changes count less in an ancestral sequence prediction.

5.4.2 Destabilize unfolded ensemble

The main property favoring the unfolded ensemble is its flexibility. Reducing this flexibility destabilizes the unfolded ensemble, thus shifting the folding-unfolding equilibrium toward folding. Two ways to reduce this flexibility are to add disulfide cross-links and to replace amino acid residues with proline. In both cases the replacement amino acid is defined; the challenge is to find suitable locations for the replacement. The ideal site in the folded structure would fit the replacement perfectly, so the only effect of the substitution is to reduce the flexibility of the unfolded ensemble.

Add disulfide cross links. Replacing a nearby pair of amino acid residues with cysteines followed by spontaneous oxidation creates a disulfide cross link, Figure 5.10. Such disulfide links often stabilize proteins to unfolding. For example, the replacement of Ala43 and Ser80 in *Bacillus* RNAse with cysteines followed by spontaneous oxidation yielded a more stable protein that unfolded in 5.77 M urea as compared to 4.58 M urea for the wild type.[19a] This engineering involved both amino acid replacements and the formation of a cross link, both of which could contribute to changes in stability. To measure the effect of the cross link alone, the researchers compared the stabilities of the oxidized disulfide form with the reduced dithiol form of the protein. These two forms differ only in the presence or absence of a cross link. For the example above, the Gibbs energy of unfolding of the oxidized form was 3.2 kcal/mol higher than for the reduced dithiol form demonstrating that the cross link stabilized the protein.

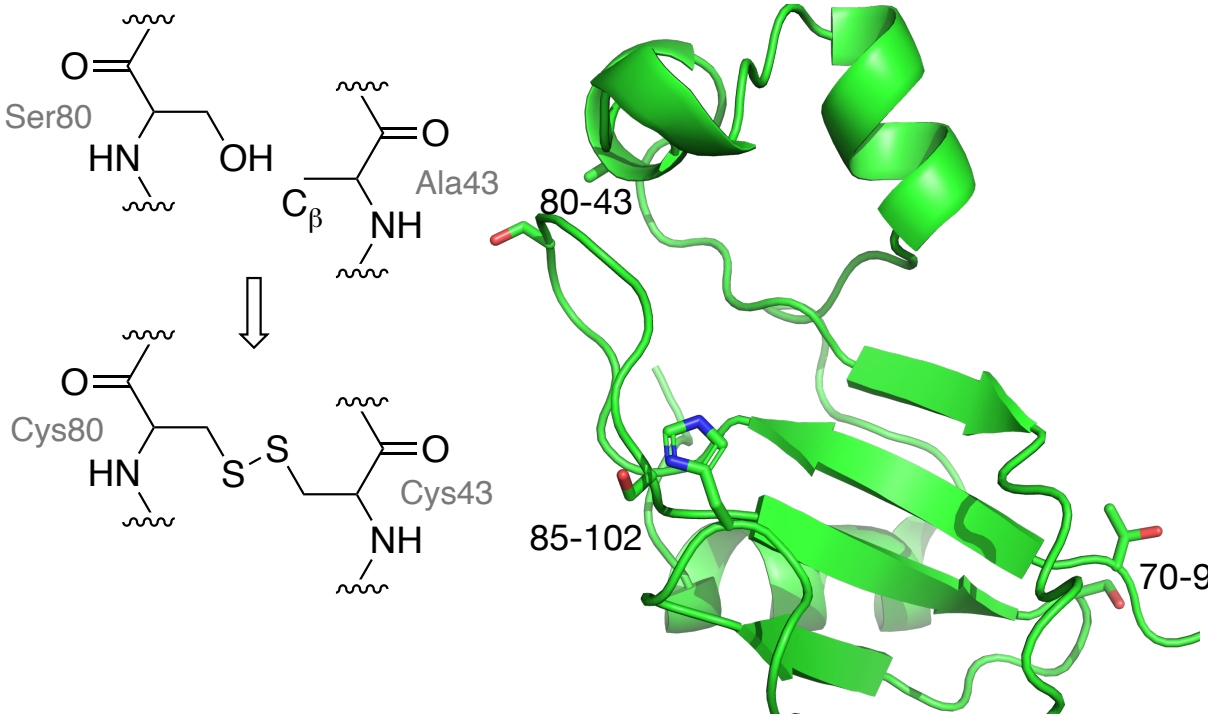

Figure 5.10 Replacement of Ser80 and Ala43 in *Bacillus* RNAse with cysteines followed by spontaneous oxidation created a disulfide cross link. This cross link stabilized the protein by 3.2 kcal/mol. Similar cross links introduced at two other locations had variable effects, see text below.

The reason that disulfide cross links stabilize proteins is that cross links reduce the flexibility of the unfolded protein. Upon unfolding, the cross link remains and limits the movement of the unfolded protein. Flexibility creates disorder (entropy), which is stabilizing. Reducing flexibility of the unfolded protein destabilizes it. The cross link constrains the distance between the linked cysteines. In the folded form, the cysteines are already close, so this constraint has little effect. In contrast, this constraint dramatically limits the movement and flexibility of the unfolded protein. Since increased flexibility and the associated increased entropy is the main reason that proteins unfold, reducing the flexibility of the unfolded form shifts the folding-unfolding equilibrium toward folding. This shift stabilizes the protein to unfolding.

A common misconception is that disulfide cross links prevent unfolding; they do not. Proteins remain flexible, dynamic structures even after adding disulfide cross links. They still unfold in urea solutions and upon heating. The RNAse protein in the example above still unfolded in urea solutions, although unfolding required higher urea concentrations that for wild-type. A similar misconception is that protein cross links strengthen and stabilize the folded protein state. They do not. The locations chosen for a disulfide cross link are already nearby in the folded protein. The introduction of the cysteines and the cross link may slightly change the folded protein stability, but this change is most often a destabilization, see below.

The chain entropy model predicts the amount of expected stabilization from a disulfide cross link. This model calculates the change in entropy due to reduced flexibility of the unfolded protein state caused by the disulfide cross link. This model assumes that the introduction of

cysteines and formation of the disulfide cross link has no effect on the stability of the folded protein state. The amount of protein stabilization expected from disulfide cross link depends on the size of the ring created by the cross-link. Larger rings limit the flexibility of more amino acids and are therefore more stabilizing. Equation 5.13 below estimates the chain entropy effect of disulfide cross link on the unfolded state.[20]

$$\Delta\Delta S^{unfold} = -2.1\,\text{cal/mol}\cdot{}^{\circ}\text{K} - \frac{3}{2}R\ln n \tag{5.13}$$

The $\Delta\Delta S_{unfold}$ refers to change due to the crosslink for the entropy change associated with unfolding; the *n* is the number of residues in the ring created by the crosslink and R is the gas constant. A link between amino acid residues 43 and 80 creates a ring of 38 residues; one more than the difference between 80 and 43. This equation comes from from polymer theory estimates of the probability that the two ends of the chain will coincide in the same volume element. If linked, the two ends must remain in the same volume element, while if unlinked they may move apart and allow greater numbers of conformations. To estimate the Gibbs energy change of the disulfide cross link using the chain entropy model, one multiplies by −T where T is the temperature (in °K) yields the anticipated Gibbs energy contribution, eq. 5.14.

$$\Delta\Delta G^{unfold}\ (\text{cal/mol}) = -T\Delta\Delta S^{unfold} = T\left[2.1 + \frac{3}{2}\cdot 1.987\cdot\ln n\right] \tag{5.14}$$

For example, the chain entropy model predicts that at 25 °C a cross link between amino acids 43 and 80 will stabilize a protein by 3.9 kcal/mol, which is slightly higher than the measured value of 3.2 kcal/mol in *Bacillus* RNase, Figure 5.11. The stabilization predicted by eq. 5.14 increases non-linearly with the number of amino acid residues in the ring, Figure 5.11. The estimated stabilization for the smallest possible ring (two amino acids) is 1.2 kcal/mol. The stabilization increases to 3.0 kcal/mol at n = 15, which is the average separation of disulfide links in natural proteins.[21] With larger numbers of amino acid residues in the ring, the increase slows because the cross link restricts a smaller proportion of the motions in larger rings as compared to smaller rings.

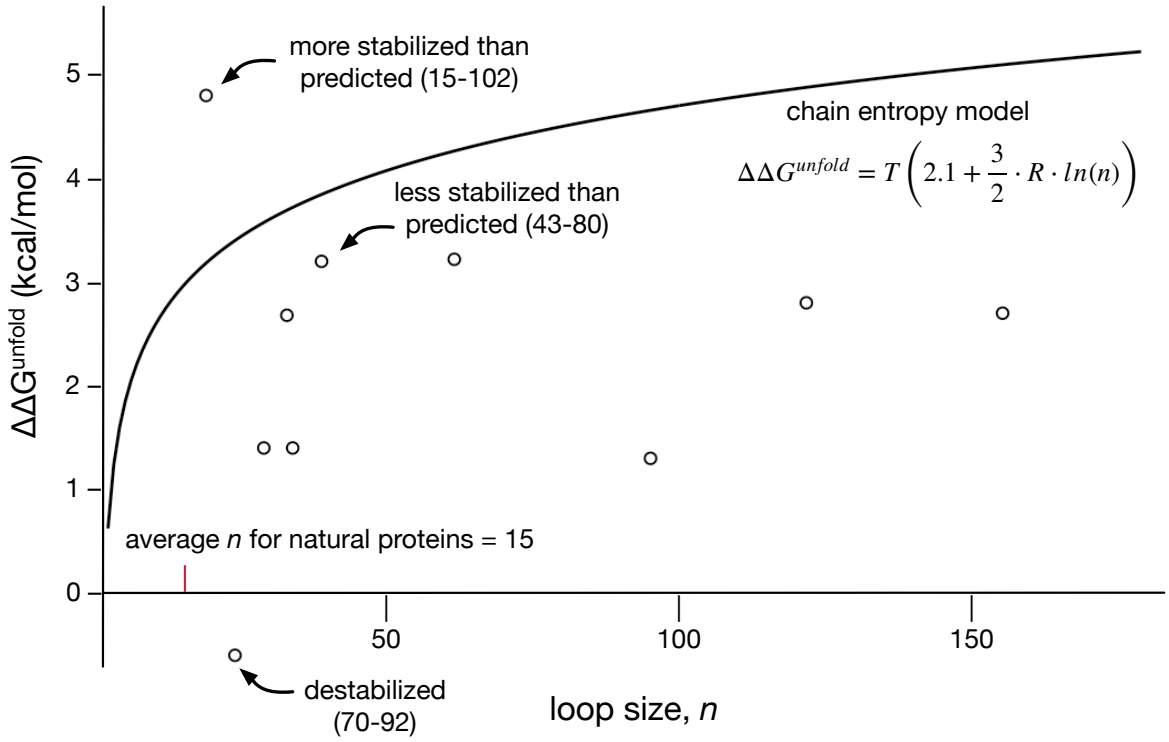

Figure 5.11 The predicted and measured effects of a cross link on protein stability. The open circles are the measured differences in stability between the the oxidized protein (the two cysteines form a cross link) and the reduced protein (contains the two cysteines, but no cross link). The line is the predicted difference at 24 °C according to the chain entropy model. Loop size, *n*, is the number of amino acids connected by the disulfide. In all but one case the measured effect is lower than predicted, which may be due to disruption of the folded protein by the cross link. Five data points are for T4 lysozyme,[21a, 21b] three (marked by arrows) are for *Bacillus* RNase,[19a] and one each for chymotrypsin inhibitor 2 (Roesler & Rao, 2000) and ribonuclease H (Kanaya et al., 1991). Estimated change in the Gibbs energy of unfolding due to creation of a ring with a disulfide link. Inset shows equation of the plotted line. T = 25 °C (298 °K).

The measured values for stabilization often do not match that predicted by the chain entropy model, likely because the chain entropy model ignores any effect on the folded state. The x-ray structure of *Bacillus* RNase 43-80 disulfide link mentioned above shows a small reorganization of backbone structure that may account for the slightly lower than predicted stabilization (3.2 vs. 3.9 kcal/mol).[19a] Most of the examples in Figure 5.11 seem to fit in this category. Two other disulfide links introduced in *Bacillus* RNase also are unusual cases. The disulfide at positions 85-102 stabilized the protein more than predicted by the chain entropy model, while the one at 70-92 was so much lower that it destabilizes the protein. The x-ray structures of the more-stable-than-expected protein showed no structural changes as compared to wild type, while the destabilized cross-linked protein showed substantial reorganization that apparently destabilized the folded protein.

To stabilize proteins by introducing a disulfide link, one must identify locations where the disulfide cross link fits in the folded protein without disrupting it. This step requires a structure of the target protein, although a homology model may also be used. Most prediction programs use

geometry-only calculations where the distances and angles required for a disulfide crosslink are fit to the current fixed locations of possible amino acid pairs. Typically the programs first identify amino acid pairs with suitable Cβ-Cβ distances (2.9-4.6 Å) and second, estimate the energy cost of distorting a disulfide bond to fit. Links with high predicted strain energies should be avoided. For example, conformation 1 in Figure 5.11 between Ala4 and Asp89 is predicted to be destabilizing. The expected strain energy (total energy, TOTNRG column) is 6.2 kcal/mol, which is larger than the predicted stabilization energy for this link using eq. 5.14 (4.6 kcal/mol). Since disulfide cross links that create larger rings are predicted to be more stabilizing by the chain entropy model, one normal favors larger rings over smaller rings.

Three geometry-only programs available as web tools are SSBOND (currently offline), [22] Disulfide by Design (http://cptweb.cpt.wayne.edu/DbD2/index.php),[22a] and YOSSHI (https://biokinet.belozersky.msu.ru/yosshi). [22b] The web interface requires only a pdb file of the protein structure for the calculation. YOSSHI also identifies disulfide cross links that occur in homologs. Since these cross links have been tested by natural selection, they may be prioritized for testing. Figure 5.12 shows part of the results of an SSBOND prediction.

C_β–C_β distance test

NR	RES1	--	RES2	NAME1	NAME2	CB DIST	CA DIST
1	4	--	89	ALA	ASP	4.561	6.400
2	15	--	18	LEU	TYR	3.902	5.848
3	16	--	201	ASP	ALA	3.945	6.239

C_β–S_γ, S_γ–S_γ distance & angle test

	SGDIST	X1	X2	X3	X2'	X1'	CHINRG	TAUNRG	DISNRG	TOTNRG
1	CONFORMATION BETWEEN :			4 -	89	ALA ASP				
1	2.030	173.2	-134.8	-132.4	-62.1	-156.9	5.83	.24	.10	6.18
2	CONFORMATION BETWEEN :			15 -	18	LEU TYR				
1	2.030	69.2	156.5	-78.3	-88.6	120.1	4.40	.53	.42	5.35
2	2.030	-156.7	-98.7	-66.0	153.3	-26.6	5.48	.65	.21	6.35
3	CONFORMATION BETWEEN :			16 -	201	ASP ALA				
1	2.030	34.5	171.5	109.6	84.9	97.0	4.45	.15	.04	4.64
2	2.030	-90.4	117.7	85.9	-149.2	-159.6	4.32	.39	.02	4.73
3	2.030	1.2	-122.6	-103.9	-75.2	-165.5	5.25	.11	.02	5.37
4	2.030	-99.2	174.3	-114.1	137.2	82.4	5.41	.37	.87	6.65

Figure 5.12 The program SSBOND predicts potential locations for a disulfide link in a protein. A possible pair of amino acids in a protein: Gly and Asp. The program first identifies amino acid pairs with Cβ-Cβ distances in the range of 2.9-4.6 Å. In the case of glycine, which does not have a Cβ atom, SSBOND predicts its possible location from the location of C, N, and Cα. An example SSBOND calculation for a haloalkane dehalogenase protein (pdb ID = 1ede) with default settings found 81 pairs that match the Cβ-Cβ distances. The first three are shown here. In the second stage, SSBOND checks for suitable Cβ-Sγ distances, Sγ-Sγ distances, and proper angles; then calculates the energy cost for deviation from ideal geometry (TOTNRG = total energy). SSBOND ignores any potential interactions of the disulfide link with rest of the protein. This second stage eliminated two pairs leaving 79 possible pairs. The third pair (Asp16-Ala201) has predicted strain energy of at least 4.6 kcal/mol, while equation 3 above estimates 5.3 kcal/mol of stabilization at 25 °C. The prediction is stabilization by 0.7 kcal/mol. An experimental test revealed an increase in the melting transition of 5.2 °C[23] which corresponds to ~1.5 kcal/mol of stabilization according to ΔG_{unfold} (kcal/mol) ~ ΔT_m(°C)/3.6.[13]

Since geometry-only calculations ignore interactions between amino acids, one must consider these before choosing sites for mutagenesis. First, most researchers also avoid cross links near the active site to prevent possible disruption of catalysis. Second, geometry-only calculations do not consider possible loss of favorable interactions due to removal of existing amino acids, nor do they consider potential bumping interactions between the cysteines and adjacent amino acids. Examination of the protein structure can identify these potential problems. Some researchers also include a molecular dynamics simulations to generate alternative protein conformations.[23] These simulations could identify additional locations where a disulfide cross link can fit or identify locations that should be avoided since the cross link causes large shifts in the backbone that may destabilize the folded protein state.

Disulfide bonds usually form spontaneously upon air oxidation of the cysteines. If the application requires the protein to remain in the cytoplasm, which is a reducing environment, the disulfide bonds may not form spontaneously making this stabilization approach unsuitable. Mutant strains of *E. coli* where the enzyme thioredoxin reductase is inactivated can form disulfide cross links within the cytoplasm.[23a]

While disulfides are the most common way to cross link the amino acid chain for protein stabilization, there are other possibilities.[24] For example, connecting the N- and C-termini into a ring also restricts the main chain motion of the unfolded protein and stabilizes proteins to unfolding. Formation of this ring requires special methods as well as a protein fold where the N- and C-termini are near one another.

Introduce proline residues. One can also reduce the backbone flexibility of the unfolded ensemble without creating crosslinks by selective amino acid substitutions. Proline is the least flexible amino acid because its ring limits its backbone conformations. Similarly, glycine is the most flexible amino acid since no side chain limits the possible backbone conformations. Replacing any amino acid with proline is expected to reduce the flexibility of the denatured ensemble and stabilize the folded protein. Replacing glycine with any other amino acid should have a similar effect. To achieve a net stabilization, these substitutions should not otherwise stabilize the denatured ensemble and must not destabilize the folded form with unfavorable interactions. As in other cases, one should avoid changes in the active site of an enzyme to prevent disrupting catalysis.

flexibility: Gly > Ala & 17 other aa >Pro

Introducing proline is a reliable approach to stabilizing proteins. Replacing a typical amino acid with proline reduces the number of conformation in the denatured ensemble by an estimated factor of 7.5 or a ΔS of -4 cal/mol $\cdot$ deg,[26] eq. 5.15, which corresponds to a Gibbs energy contribution of ~1.2 kcal/mol at 298 °K.

$$\Delta S = R\ln\left(\frac{\#\,of\ conformations\ for\ typical\ aa}{\#\,of\ conformations\ for\ proline}\right) = \Delta S = R\ln(7.5) = 4\,cal/mol\cdot deg \qquad (5.15)$$

This approach requires identifying a suitable location for the proline. The main chain angles for proline fit well in three places: near the start of an α-helix, the i+1 position in a type I or II β-turn[26] or the i position of a type II β-turn.[27] These choices of proline location consider only main chain angles; the success of any substitution also assumes minimal changes to side chain interactions. The WHAT-IF web tool at https://swift.cmbi.umcn.nl/servers/html/index.html can identify these locations in a protein structure (Click on 'mutation prediction', then 'Suggest proline mutations.'). The first example of protein stabilization by proline substitution was an Ala82Pro substitution in T4 lysozyme, which occurs near the start of an α-helix and stabilized the protein by 0.8 kcal/mol, Table 5.4. The restricted backbone angles of proline fit well near the start of an α-helix and the rest of the structure also fit a proline residue. The first four residues of a helix lack a partner to which the main chain N-H can donate a hydrogen bond, so the lack of an N-H in proline is not a disadvantage at the start of a helix. The typical success rate for stabilization upon proline substitution is ~50%.

Table 5.4 Stabilization of lysozyme from phage T4 by substitutions that reduce the flexibility of the denatured state.[25]

protein	location of substitution	T_m, °C	ΔΔG (kcal/mol)
wt	-	64.7	0
Ala82Pro	near start of α-helix	66.8	0.8
Gly77Ala	near end of α-helix[a]	65.6	0.4

[a]Stabilization likely due to stabilization of the helix and not due to changes in flexibility of the denatured state.

Although replacing glycine with alanine also reduces the flexibility of the denatured state, other considerations limit the effectiveness of this substitution. Replacing glycine residues with alanine reduces the number of conformations available to the unfolded ensemble by approximately a factor of three, which corresponds to an Gibbs energy contribution of ~0.4 kcal/mol at 298 °K, which is smaller than that of proline substitution, 1.2 kcal/mol. As with the addition of proline, one must ensure that the glycine to alanine substitution does not strain the folded form. Since glycine can adopt conformations that are not accessible to alanine (e.g., a left-handed helix conformation which occurs in some β-turns), replacements at these locations would destabilize proteins. Second, alanine can create stabilizing interactions within the denatured state that counterbalance the reduced flexibility so the net effect may be zero.[28] The Gly77Ala example in Table 4 above stabilized the protein, but the x-ray structure suggests that it stabilized the folded form due to the burial of the hydrophobic methyl group.

5.4.3 Stabilize folded form

Strengthening interactions between amino acids in the folded protein is the most obvious approach to stabilizing proteins. Starting from the three-dimensional structure of the protein, one designs various improved interactions. One difficulty with this approach is that proteins already create stabilizing interactions, so this design searches for additional interactions. One must predict both the location for the changes and the replacement amino acids.

Flexible regions identify weak spots in proteins. The effect of flexible regions in proteins can be confusing. Flexibility is a weakly stabilizing feature because it increases entropy.[28a] This flexibility may contribute a factor of two, or 0.4 kcal/mol, to stability. While this flexibility is stabilizing, it also indicates that interactions with other amino acids are weak. Creating strong interactions between amino acids such as hydrogen bonds and hydrophobic interactions (several kcal/mol) in place of the weak stabilization of flexibility can be a net gain for stability, Figure 5.13. Thus, it is not removing flexibility that stabilizes the protein, but the addition of new interactions between the amino acids. Loops on the surface and the N- and C-termini of the protein are often the most flexible.

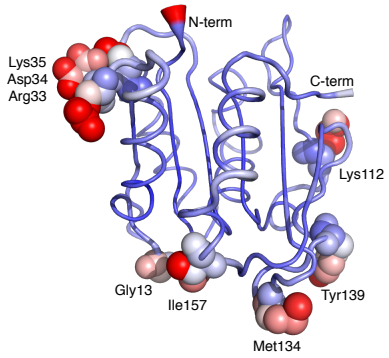

Figure 5.13. X-ray structure of a lipase (pdb code 1isp) colored to show the B-factor from low (blue) to high (red). Reetz and coworkers[29] replaced the residues in eight regions with the highest flexibility (B-factor), excluding the N- and C-termini, using random mutagenesis. Stabilizing substitutions (2-4 °C increase in T_m) occurred in the six regions where the residues shown as spheres. Side chain atoms are more flexible than the main chain atoms.

There are two steps to this stabilization approach: first to identify the flexible regions of the protein and second, to identify stabilizing substitutions. One way to identify flexible areas is molecular dynamics simulation, which directly models protein motion.[23] Another way to identify flexible regions is the high B-factors in the x-ray crystal structure.[29] B-factors describe the spreading of electron density assigned to that atom. This spread may be due to movement of the atom during the x-ray analysis (temperature-dependent atomic vibrations), or may be due to the atom occupying several fixed positions in the structure (static disorder), which suggests motion in solution. Errors in model-building can also increase B-factors. The B-FITTER program (https://www.kofo.mpg.de/en/research/biocatalysis) identifies the amino acids in a protein that have the highest B-factors. B-FITTER calculates the average of the B-factors for all atoms (excluding

hydrogen) for each amino acid and displays a list of the twenty amino acid residues with the highest B-factors.

The second step is to identify stabilizing substitutions. One can make changes within the flexible region or next to flexible region. Interactions between amino acids require partners so either approach can yield improvements, but one comparison found larger stabilization with the replacements in the neighboring areas.[30] Some researchers targeted the flexible regions with random mutations,[29] while others used modeling to predict stabilizing substitutions.[30]

Copy features from thermotolerant homolog. Proteins from microorganisms that live in extreme environments (such as high temperatures >80 °C; extremes of pH, high salt, high pressure) tolerate their extreme environment better than homologous proteins from mesophiles (organisms that grow best at moderate temperatures). One protein stabilization strategy is to transfer the amino acids responsible for this extra stability in proteins from extremophiles to the corresponding proteins from mesophiles. This difficulty of this strategy is identifying which amino acids to transfer since some sequence differences impart increased stability, but most of the differences reflect random genetic drift.

One approach relies on amino acid sequence comparison within the extremophiles. The expectation is that stabilizing amino acids are conserved in the proteins from extremophiles, but missing from the target protein from mesophiles. This approach differs from the consensus sequence approach, identifies residues conserved among all homologs, including both those from extremophiles and mesophiles. One pitfall of this approach is that residues may be conserved within extremophiles because they are closely related, not because they contribute to stability. For example, comparison of a pectate lyase with four homologs from thermophiles identified nine substitutions common to the enzymes from thermophiles, but absent in the target enzyme.[31] However, only one of these nine substitutions (Arg236Phe) significantly increased the thermostability of the target lyase (12-fold), likely by improving hydrophobic packing.

A second approach relies on the structures of proteins from extremophiles and the identification of stabilizing interactions within them. The stabilizing interactions are not apparent from the structure, so identifying them requires modeling. For example, researchers identified which ion pairs in adenylate kinase from a thermophile contribute to stability using molecular dynamics simulations.[32] During the simulated movement of the protein some ion pairs broke apart, while three of them remained tightly paired. Transfer of these tightly-paired ion pairs into a less stable homolog increased its melting temperature by 0.8 to 3.3 °C.

Stabilizing interactions are not apparent from a comparison of structures because no single factor dominates.[33] Each protein uses a unique mixture of interactions similar to those in Table 1 above to increase protein stability. For example, many proteins from thermophiles contain more extensive hydrogen bond networks to strengthen electrostatic interactions, but some proteins from thermophiles do not include more extensive hydrogen bond networks but are nevertheless stable

at high temperature. Similarly, some, but not all, proteins from thermophiles contain increased atomic packing to strengthen hydrophobic interactions, increased numbers of ion pairs, shortened loops to minimize interactions with the solvent, increased occurrence of Ala in helices and increased oligomerization to enhance interactions between amino acids in the folded state. Other stabilizing features are not apparent in the folded structure because their effect is to destabilize the denatured state. This wide range of features strengthens the argument that there are multiple paths to protein stabilization.

Remove destabilizing electrostatic interactions on the protein surface. Charged residues create favorable electrostatic interactions between residues with opposite charge and unfavorable interactions between residues with the same charge. Optimizing these interactions in the folded protein, especially by removing destabilizing interactions on the surface of the protein, stabilizes the protein. For example, five substitutions on the surface of an acylphosphatase increased the Gibbs energy of unfolding by 2.2 kcal/mol without affecting catalytic activity.[34] Three of these substitutions reversed the charge of the residues and two substitutions introduced new charges.

The two reasons to choose substitutions on the protein surface are that they are likely to be far from the active site and that they are likely to maintain good solvation. Avoiding mutations near the active site increases the probability that the modified protein will maintain catalytic activity (or binding in the case of a binding protein). Choosing residues with good solvation for substitution makes the prediction of stabilizing or destabilizing more accurate. The effect of a charged amino acid residue on protein stability depends on the differences in both electrostatics and solvation in the folded versus unfolded states. Since all residues are assumed to be well-solvated in the unfolded state, choosing residues that are well solvated in the folded state allows the prediction to ignore solvation in the comparison.

One reason to avoid substitutions on the protein surface is that they may have a smaller effect on stability than substitutions in the interior of the protein. The environment of residues on the protein surface changes less drastically upon unfolding than the environment of residues in the interior of the protein. Nevertheless, residues on the surface move apart when the protein unfolds, thereby changing the electrostatic interactions.

TKSA-MC (http://tksamc.df.ibilce.unesp.br) is a web-tool to identify residues for replacement due to unfavorable electrostatic interactions (Contessoto et al., 2017). It calculates the electrostatic contribution of all the charged residues in the protein and suggests replacing them if they make an unfavorable electrostatic contribution and lie on the protein surface. It does not suggest what the replacement amino acid should be, but uncharged polar residues or oppositely charged residues are the obvious choices. The electrostatic calculation requires a protein structure and considers the pH of the solution, the distance between the charges, different dielectric constants for the protein and solvent, and the fraction of the residue exposed to solvent. For example, TKSA-MC predicts that replacement of residues Asp36, Asp40, and Glu42 in

Staphyloccocal binding protein would stabilize it, Figure 5.x

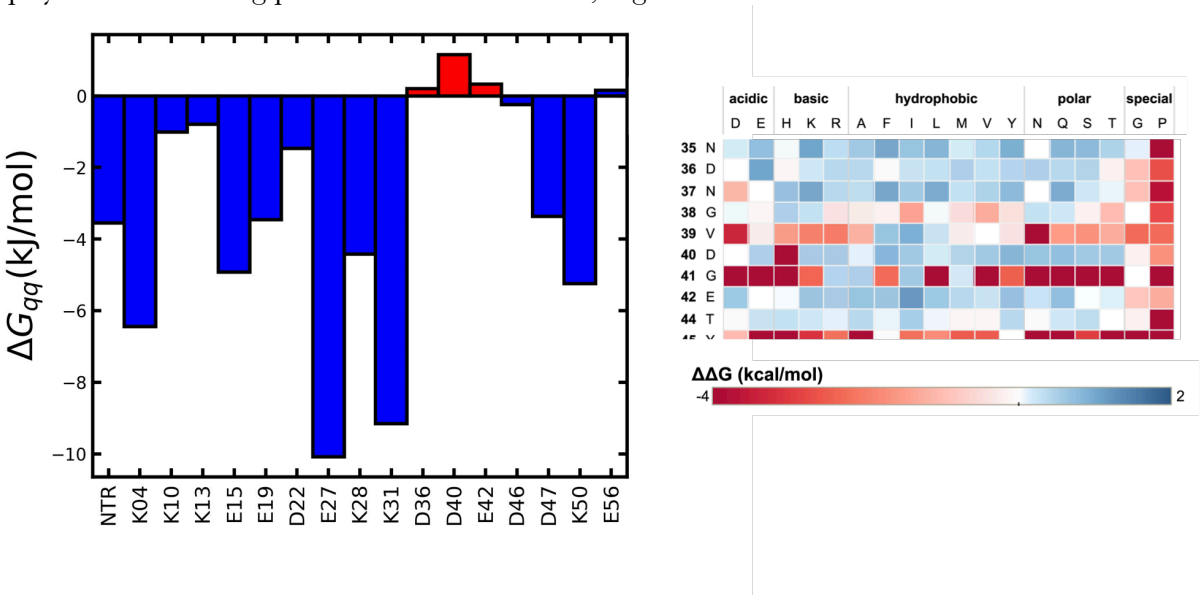

Figure 5.x. Stabilization of Staphyloccocal binding protein by replacement of destabilizing residues on the protein surface. a) Prediction of electrostatic interactions at pH 7 by web tool TKSA-MC based on the x-ray crystal structure (PDB id: 1pga). Positive values indicate destabilizing interactions. Replacement of the destabilizing residues filled in red (Asp36, Asp40, and Glu42), which indicates those on the surface, is predicted to stabilize this protein. b) Blue squares indicate replacements of amino acids in Staphyloccocal binding protein that stabilize the protein. In agreement with predictions, most replacements of Asp36, Asp40, and Glu42 lead to a more stable protein.

Most electrostatic interactions are stabilizing and only a few are destabilizing, so this approach identifies only a few locations for mutagenesis. The ability to predict new stabilizing interactions would yield additional locations for mutagenesis, but this approach has not been tested with this web tool.

Modeling to identify stabilizing substitutions. Rosetta[37] and FoldX[38] are two modeling programs widely used to predict protein stability. Both programs model the physical interactions of the atoms in the folded form, but Rosetta adds statistical analysis of different properties extracted from protein databases. A web service for Rosetta design is available at http://rosettadesign.med.unc.edu and for FoldX at https://loschmidt.chemi.muni.cz/fireprot/.

FoldX is force field specifically for predicting protein stability (Schymkowitz et al., 2005). FoldX requires an experimental structure of a protein as the starting point. It does not do geometry optimization or conformational searching; it only calculates unfolding Gibbs energy for the protein and variants of the protein. For protein engineering, it predicts substitutions that stabilize or destabilize a protein. The force field equation contains terms for steric clash and electrostatic interaction like the physics-based force fields, but it also contains terms for solvation and entropy, eq. 5.16.

$$\begin{aligned}\Delta G = {} & a\cdot\Delta G_{vdw} + b\cdot\Delta G_{solvH} + c\cdot\Delta G_{solvP} + d\cdot\Delta G_{wb} && \Leftarrow \text{ solvation terms}\\ & + e\cdot\Delta G_{hbond} + f\cdot\Delta G_{el} + g\cdot\Delta G_{kon} && \Leftarrow \text{ electrostatics terms}\\ & + h\cdot T\Delta S_{mc} + k\cdot T\Delta S_{sc} && \Leftarrow \text{ entropy terms}\\ & + l\cdot\Delta G_{clash} && \Leftarrow \text{ steric clash term}\end{aligned} \tag{5.16}$$

These added terms estimate the Gibbs energies of various interactions of the amino acids with each other in the folded form as compared to interaction with solvent, which mimics the unfolded form. For example, the four solvation terms estimate the attractive van der Waals interaction between water and the protein (from the energy of removal of a protein atom from water to vapor phase), the desolvation of hydrophobic and polar groups upon folding (from the energy of removal of a protein atom from water to an organic solvent), and specific interactions with water when the water molecules that make more than two hydrogen bonds with the protein. In these cases the water molecules are included as atoms in the calculation. The terms in the FoldX force field were scaled with respect to each other to match the predicted Gibbs energy to experimental measurements of >1000 stabilizing and destabilizing amino acid substitutions. A FoldX calculation is part of the HotSpot Wizard web tool (Sumbalova et al., 2018) at: http://loschmidt.chemi.muni.cz/hotspotwizard.

Rosetta's equation for energy is a hybrid of simple terms that model physical interactions combined with knowledge-based terms to improve accuracy. These knowledge-based terms come from statistical analysis of known protein structures. For example, while FoldX estimates electrostatic interactions using Coulomb's law and hydrogen bonding terms, Rosetta also weights the estimate by the probability that the two charged atoms occur nearby in PDB structures. The supporting information contains a detailed example of using Rosetta to identify a stabilizing substitution.

The energies from a molecular mechanics calculation like Rosetta are energy relative to a hypothetical unstrained molecule. A geometry optimization gives you a stable, reasonable structure, but the energy value needs a comparison value to make sense. For example, if you calculate the energy of the wt and a mutant, you can conclude which one of those folded structures is more stable. (The folded state is a collection of many conformations, so you are assuming that this single geometry-optimized structure is a good representative all folded structures that contribute to the folded state.) If you further assume that the energies of the unfolded states are identical, then the difference between the energies of the wt and mutant is the difference in unfolding energies. In your case, Rosetta predicts that the mutant is less stable than the wt (less negative). Recall that the success rate of these predictions was about 10% in the paper that we discussed in class. The units in Rosetta are 'Rosetta energy units' which are similar to kcal/mol.

Engineering increased stability by introducing new interactions is difficult because substitutions can perturb the protein structure in unexpected ways. These perturbations can destabilize the structure and offset the intended stabilizing effect of the substitution. The

reliability of the predicted stabilizing substitutions by these and similar modeling programs remains low. Some things that are missing from the models are alternative conformations of the protein backbone, explicit water molecules, knowledge of the unfolded state and good models of entropic effects. Simulating alternative conformations using molecular dynamics with explicit water molecules adds these interactions and improves the success rate to 13-19%.[23,39] Another approach to increasing success is to average the predictions of numerous methods. The assumption is that errors arising from different simplifications will cancel out. A web-tool to predict the protein stabilization effect of a substitution using eleven different modeling methods is available at http://meieringlab.uwaterloo.ca/stabilitypredict/.

Molecular modeling methods typically identify substitutions to increase hydrophobic interactions and various electrostatic interactions. For example, Rosetta predicted improved hydrophobic interactions in cytosine deaminase with three substitutions (A23L, I140L, V108I), which increased the apparent melting temperature by 10 °C.[40]

The web server [PROSS](https://pross.weizmann.ac.il/step/pross-terms/) (Protein Repair One-Stop Shop,[40a]) combines the homology search used by the consensus sequence approach with computational design. First, PROSS searches for homologs to identify which substitutions are allowed at each position. The rationale for this evolution-based constraint is to favor variants that maintain their original molecular function. While the consensus sequence approach predicts stabilizing substitutions from the frequency of occurrence, PROSS adds an energy calculation to predict which substitutions are best. This calculation requires a three-dimensional structure of target protein. The rationale for adding energy calculations is that individual proteins differ so that an amino acid residue that is suitable in most homologs may not be suitable in the target protein. For example, stabilization of acetylcholine esterase involved a replacement for glycine at position 416. There was no clear choice for a replacement based on the frequency of occurrence because nine amino acids, including glycine, appeared at this position with similar frequency. Computational modeling of the replacements using Rosetta eliminated the most commonly-occurring histidine because it fit poorly and suggested third-most-frequently occurring glutamine because it formed a hydrogen bond with a nearby tyrosine.

5.5 Stabilization to stepwise unfolding

So far the discussion assumed that the protein is a single domain protein that unfolds cooperatively, reversibly, and in a single step. Multiple domain proteins likely unfold stepwise with one domain unfolding first followed by other domains. For these proteins, stabilizing substitution must stabilize the domain that unfolds first. Proteins may also associate into oligomers. Oligomerization of folded proteins creates stabilizing interactions between amino acids at the protein-protein interface. (If the interactions were destabilizing, the proteins would not associate.) Creating additional interactions at the protein-protein interface is a new stabilization strategy available for oligomeric proteins.

5.5.1 Multiple domain proteins

Multiple domain proteins may unfold in a single step like single domain proteins, but more often they unfold stepwise with each domain unfolding separately. Stepwise unfolding creates an metastable intermediate with part of the protein folded and part unfolded. The stabilization must focus on the domain that unfolds first.

Bacterial cocaine esterase consists of three domains: an α/β-hydrolase domain, a cap domain, and a jelly roll domain. Molecular dynamics simulations indicated that the cap domain unfolded first. Two substitutions in the cap domain (T172R/G173Q) increased the half-life from ~0.2 h to ~6 h at 37 °C,[41] which corresponds to stabilization of 2.1 kcal/mol. Modeling suggested that the T172R substitution strengthened interactions within the cap domain, while the G173Q substitution strengthened interactions between the cap and α/β-hydrolase domains by creating a new hydrogen bond.

Eijsink and coworkers dramatically stabilized a neutral protease from *Bacillus*, NprT, to self-degradation by stabilizing it to unfolding, Table 5.6.[42] Proteolytic degradation proceeds from the unfolded form and does not depend strongly on amino acid sequence for this broad specificity protease, so stabilization to unfolding also slows the proteolytic degradation.The stabilizing substitutions were either copied from a heat tolerant homolog or rationally designed to stabilize the protein. The half-life at 70 °C increased dramatically from <0.5 to 170 min, which corresponds to 4.0 kcal/mol. The researchers named this stabilized enzyme 'boilysin' because of its ability to remain active at 100 °C.

Table 5.6 Substitutions that stabilize neutral protease NprT to unfolding at >100 °C.[42]

ΔT_m[a]	substitution	how identified	molecular basis for stabilization
+12.3 °C	Ala69Pro	copied from heat tolerant homolog	reduce flexibility of unfolded domain
	Thr63Phe	copied from heat tolerant homolog	create hydrophobic interaction near surface
+7.6 °C	Gly58Ala	copied from heat tolerant homolog	possible reduced flexibility of unfolded domain
	Thr56Ala	copied from heat tolerant homolog	not determined
	Ala4Thr	copied from heat tolerant homolog	not determined
+6.9 °C	Ser65Pro	rational design	reduce flexibility of unfolded domain
	8-60 Cys crosslink	rational design	reduce flexibility of unfolded domain

[a]The increases in apparent melting temperature associated with the substitutions shown. Upon melting, the protease self-degrades rapidly.

The stabilizing mutations all cluster in one region of the protein, Figure 5.14. This multiple domain protein unfolds stepwise, and this region of the protein is the part that unfolds first and is

then cleaved by other protease molecules. Stabilizing this region to unfolding was the key to preventing proteolytic degradation.

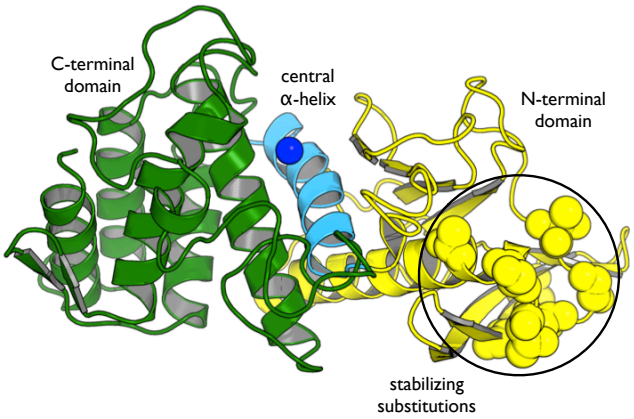

Fig. 5.14 Cartoon representation of a model of thermolysin-like protease. The active site zinc ion (blue sphere) lies near the central a-helix (light blue) between the N-terminal domain (yellow) and C-terminal domain (green). The protease's stability depends on partial unfolding, so stabilizing substitutions (spheres within circle) cluster in that region: the N-terminal domain of the protein, in particular in the 55–69 surface loop. This model of protease NprT (P06874) created by SwissModel using the structure of thermolysin as the template, which has 87% identical amino acids.

5.5.2 Oligomeric proteins

Dimerization or oligomerization of monomeric native state proteins creates new protein-protein interactions while decreasing protein-water interactions and therefore stabilizes the folded native form, Figure 5.15. In contrast, if unfolded or partially unfolded forms dimerize and oligomerize, then these interactions stabilize the unfolded structures, which eventually leads to aggregation.

$$N \underset{}{\overset{K_{unfold}}{\rightleftharpoons}} D \qquad \text{no oligomerization}$$

$$O_N \overset{K_{dis}}{\rightleftharpoons} N \overset{K_{unfold}}{\rightleftharpoons} D \qquad \text{N oligomerization stabilizes N}$$

$$N \overset{K_{unfold}}{\rightleftharpoons} D \overset{K_{olig}}{\rightleftharpoons} O_D \qquad \text{D oligomerization destabilizes N}$$

Figure 5.15 Dimerization or oligomerization of the native state (N) stabilizes the native state by shifting the overall distribution of conformations away from the denatured state (D). Strengthening interactions between native monomers stabilizes proteins to denaturation. In contrast, dimerization or oligomerization of the denatured state shifts the distribution of conformations away from the native state and eventually leads to aggregation. K_{dis} = equilibrium constant for dissociation oligomers of native form, O_N, to monomers; K_{olig} = equilibrium constant for association of denatured form into oligomers, O_D.

Substitutions at the oligomer interface that strengthen the interaction between monomers will stabilize proteins. In dramatic example, a replacement to remove a negative charge at the dimer interface of malate dehydrogenase (Glu165Gln or Glu165Lys) increased the melting temperature by 24 °C.[43] Another approach is to covalently link the monomers into dimers by introducing a

disulfide link. Comparison of the structures of a peptidase and a thermostable homolog identified a cross-link between monomers in the thermostable homolog. Adding the this link to the peptidase increased its denaturation temperature by 30 °C.[44]

5.6 Weaknesses other than unfolding

While unfolding is the weak point of most proteins, in some cases, other weaknesses dominate protein instability. Aggregation of egg white proteins into insoluble particles was mentioned above. Other examples are chemical modification of proteins such oxidation or degradation by proteases. These other weaknesses may be related to unfolding. Proteins unfold at least partially before they aggregate or are degraded by proteases. Chemical modification can promote unfolding. A different type of protein instability is short serum half-life, which depends on biochemical processes that clear proteins from the bloodstream and is outside the scope of this tutorial.

5.6.1 Irreversible unfolding (aggregation)

Protein aggregation is the irreversible association of partially or fully unfolded proteins into insoluble particles. Aggregation-prone regions assemble by intermolecular β-structured interactions to form the core of the aggregate. Aggregation can occur when unfolded proteins encounter one another. For example, cooking an egg first unfolds the egg white proteins to expose hydrophobic regions to the solvent. These regions associate with similar exposed hydrophobic regions of other egg white protein molecules leading to oligomerization and eventually insoluble aggregates, eq. 5.20, Figure 5.8. Here k_{agg} is the aggregation rate constant.

$$N \underset{}{\overset{K_{unfold}}{\rightleftharpoons}} D \xrightarrow{k_{agg}} \text{aggregate} \tag{5.20}$$

Aggregation can also occur under mild conditions. For example, overexpressing protein in bacteria creates high concentrations of unfolded protein. If protein folding is slow, the unfolded proteins can aggregate into insoluble particles called inclusion bodies. Biotherapeutic proteins can also aggregate during storage, which creates a danger of an immune response to the aggregates.

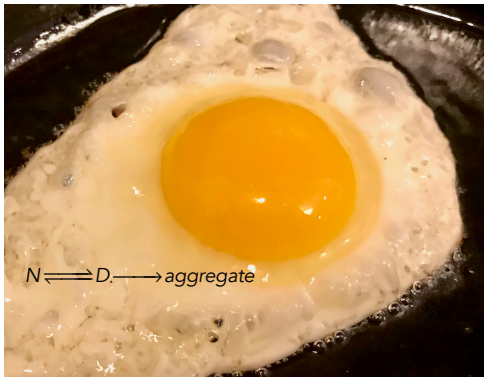

Figure 5.8 Cooking an egg involves heat-induced unfolding of the egg white proteins, followed by their aggregation into an insoluble gel. Egg white consists of ~10 wt% protein in water; this high concentration of protein promotes aggregation after unfolding.

Measuring a loss of function after heating identifies aggregation as a contribution to protein instability. For example, one measures the enzyme activity at room temperature, incubates the sample at an elevated temperature, then cools it to room temperature and measures enzyme activity again. The decrease in activity reveals the fraction of enzyme irreversibly unfolded due to heating. The values reported are typically the half-life at a specific temperature. For non-catalytic proteins, the change in intrinsic fluorescence can measure the amount of natively folded protein remaining. The activity loss depends on both how much of the protein unfolds upon heating (its inherent stability) and on its ability to refold upon cooling (which may be prevented by aggregation), eq. 5.x above.

To convert the measured loss of activity to half-life, researchers assume the inactivation follows first-order kinetics, eq. 5.8, where A is the activity at time t, A_0 is the initial enzyme activity and k is the inactivation rate constant. The natural logarithm of enzyme activity ($\ln A$) decreases linearly with time. A plot of several measurements of the natural logarithm of the remaining activity at different times yields a straight line, with a slope of $-k$ and a y-intercept of $\ln A_0$.

$$\ln A = -kt + \ln A_0 \tag{5.8}$$

By measuring the inactivation rate constants for both the wild-type enzyme and the engineered variant, eq. 5.9, yields the change in Gibbs energy of activation, $\Delta\Delta G^{\ddagger}$, for the rate of enzyme inactivation. If the mutant is more stable, then k_{mut} / k_{wt} is less than one, so the value of $\Delta\Delta G^{\ddagger}$ is positive, indicating an increase in the activation energy to unfold the protein.

$$\Delta\Delta G^{\ddagger} = -RT \ln\left(\frac{k_{mut}}{k_{wt}}\right) \tag{5.9}$$

In most cases, the rate determining step in enzyme inactivation is the unimolecular unfolding step, so the assumption of first-order kinetics is justified. However, refolding of the enzyme may also depend on aggregation of the protein into an insoluble precipitate, so the rate determining step may involve several enzyme molecules. In these cases, the inactivation will not fit equation 8, which assumes first order kinetics.

Another way to measure irreversible denaturation by heat is to measure an apparent melting temperature, $T_{m,\,app}$. It is not a true melting temperature, only an apparent one, because it is an irreversible process.

Two strategies to reduce aggregation are 1) to reduce unfolding (K_{eq} in eq. 5.20) or 2) slow down aggregation (k_{agg} in eq. 5.20). Minimizing unfolding of the protein will use strategies above in section 5.4. For example, a combination of three substitutions designed to stabilize the

native cytosine deaminase slowed its aggregation 30-fold.[40]

Since aggregation occurs mainly through hydrophobic interactions, one method to slow aggregation is to minimize the hydrophobicity of solvent-exposed regions. To reduce the aggregation of antibodies, Chennamsetty and coworkers[45] first identified hydrophobic areas on the surface including hydrophobic patches that become exposed as the protein moves. Replacing hydrophobic residues with hydrophilic ones within these patches reduced aggregation. Several web tools predict substitutions needed to reduce protein aggregation: Aggrescan3D (http://biocomp.chem.uw.edu.pl/A3D2)[46a] and Solubis (https://solubis.switchlab.org/).[46] After predicting the aggregation-prone regions (hydrophobic, β-sheet forming stretches) the web tools predict substitutions to 1) reduce their unfolding and exposure to solvent by strengthening interactions between the aggregation-prone area and the rest of the protein or 2) eliminate the aggregation propensity of the region. These substitutions are charged residues (R, K, D, or E) which reduce hydrophobic interactions needed for aggregation or proline, which hinders the formation of β-sheet structures in aggregates. The tools use FoldX to choose the substitution that is expected to best stabilize the folded form.

Hindering aggregation can even overcome a decrease in stability. Substitution A281E in *Candida antarctica* lipase B decreased its melting temperature from 58 to 51 °C indicating a decline in stability. However, this substitution increased the half-life of this enzyme at 70 °C 22-fold demonstrating a reduced propensity to aggregate.[47] This substitution makes a hydrophobic region of the enzyme less hydrophobic and less likely to aggregate.

A more drastic approach to reducing the association between protein molecules is supercharging. Supercharging is extensive substitutions on the protein surface to increase the net charge as high as +48, so the protein molecules repel one another and cannot aggregate.[48] Introducing large numbers of charge residues on the protein surface also reduce its hydrophobicity as charged amino acids replace uncharged amino acids. For example, green fluorescent protein (GFP, net charge = −7) loses its fluorescence at 100 °C due to unfolding, eq. 5.21. Cooling to room temperature does not restore fluorescence because the unfolded protein has aggregated and precipitated. In contrast, green fluorescent protein variants engineered to have a +36 or −30 net charge also lost their fluorescence upon heating to 100 °C due to unfolding, but upon cooling, they did not aggregate and regained 62 and 28%, respectively, of their original fluorescence. This supercharging did not make the GFP variants more stable; in contrast, the GFP variants unfolded more readily than wild-type in urea, but supercharging did reduce their propensity to aggregate in the unfolded state.

$$GFP^{7-}_{folded} \underset{}{\overset{heat}{\rightleftharpoons}} GFP^{7-}_{unfolded} \rightarrow aggregate$$

$$GFP^{30-}_{folded} \underset{}{\overset{heat}{\rightleftharpoons}} GFP^{30-}_{unfolded} \qquad (5.21)$$

Two disadvantages of supercharging are the potential to destabilize the protein and to disrupt its function. Residues with the same charge create unfavorable electrostatic interactions, which

destabilize proteins. A change in electrostatic environment can also shift the pK_a of residues in the active site, which may affect binding or catalysis. For example, catalysis (k_{cat}) by a supercharged glutathione-S-transferase (net charge −40 for the dimer) was three-fold slower than wild-type (net charge +2 for the dimer).

5.6.2 Chemical modification

The side chains of three amino acids readily undergo spontaneous chemical modification: 1) asparagine deamidate to aspartate, 2) methionine oxidizes to a sulfoxide and 3) cysteine oxidizes to disulfide protein oligomers as well as sulfur oxides such as sulfenic acids (RS-OH). Replacement of problematic residues with a non-reactive residue eliminates the possibility of modification, but replacement of every asparagine, methionine, and cysteine in a protein is rarely needed. Many of the residues may not undergo modification, and even when they do, some modified residues will have little effect on protein properties.

Spontaneous deamidation of asparagine to aspartate is a common chemical modification of proteins. The most reactive are Asn-Gly sequences in sterically unhindered regions which have half-lives as short as 6 d in physiological conditions.[49] The least reactive asparagines are 10^5-fold more stable. Deamidation converts the neutral Asn residue to a negatively charged Asp residue. This change can impair function, destabilize the protein or have no effect. The web tool at www.deamidation.org predicts asparagine deamidation rates based on the 3-D structures of proteins. Glutamine can also deaminate to glutamate, but the rate is about a thousand-fold slower, except for the case of an N-terminal Gln, which deaminates at a rate similar to Asn.

Oxidation of sulfur atoms in methionine and cysteine alters a proteins properties and may destabilize or inactivate it. The first protein engineering of an industrial enzyme was the removal of an oxidation sensitive methionine from the detergent protease subtilisin.[50] The existing subtilisin could tolerate most of the harsh conditions of laundry (heat, high pH, and detergent), but it could not tolerate bleach. Bleach oxidized a methionine near the active site to the methionine sulfoxide (R-S(O)-CH_3), which hindered binding of the substrate proteins to inactivate the protease. Replacement of this problematic methionine with alanine created a bleach-tolerant protease.

A cysteine-to-serine replacement stabilizes several cytokine drugs. This change does not affect biological activity, but avoids oligomerization by the formation of non-native intermolecular disulfide links between proteins. The specific activity of interferon-β, when expressed in *E. coli*, was about 10-fold lower than the native protein. The researchers hypothesized that oligomerization through intermolecular disulfide bonds caused the lower activity.[51] Replacement of Cys17 with Ser eliminated oligomerization and restored the specific activity to that of the native protein. The commercial drug, Betaseron®, contains this substitution. A similar Cys125Ser substitution stabilizes human interleukin 2, marketed as Proleukin®.

Some applications require proteins to work under conditions where other chemical modifications are possible. For example, glucose, an aldehyde, can react with lysine to form an imine. The enzymatic conversion of corn syrup (glucose) to high fructose corn syrup to increase its sweetness requires the enzyme xylose isomerase to tolerate high concentrations of glucose. Stabilization involved replacing a surface lysine that reacted with the glucose.[52] At very high or low pH, irreversible chemical reactions can degrade the protein. For example, base-catalyzed β-elimination at pH > 8 alters cystine (disulfide) residues' side chains. Peptide backbone links next to aspartic acid residues can hydrolyze at pH < 4.

5.6.3 Serum half-life

Extending the serum half-life of a therapeutic protein enhances its efficacy because it remains active for a longer time. It also lowers cost because less therapeutic protein is required and improves delivery because injections are less frequent.[52a, 52b] Serum half-lives for small proteins (< 50 kDa) are short (5–50 min) due to: (i) rapid filtration by the kidneys; (ii) receptor-mediated endocytosis and (iii) degradation by proteolytic enzymes and peripheral tissues. Methods to slow the first two processes are 1) to increase hydrodynamic radius above renal filtration threshold and 2) to enhance FcRn-mediated recycling after endocytosis.

Larger proteins have a higher hydrodynamic radius. Chemical methods to increase the size of the protein are covalent modification with poly(ethylene glycol). Protein engineering methods include adding glycosylation sites or fusion to another protein. For example, darbepoetin-α is a re-engineered form of erythropoietin containing two additional glycosylation sites.[52c] Native erythropoietin has a molecular weight of 30 kDa and a serum half life of approximately 5 h, while the extra glycosylation on darbepoetin-α increase the molecular weight to 37 kDa and lengthen the serum half-life 3.5-fold. Fusion of a therapeutic protein to another protein increases the hydrodynamic radius, but the added protein may cause an immune response. Fusion to an unstructured designed protein, which is not immunogenic, avoids this risk.[52d] The fusion protein approach also has the advantage that it works in bacterial expression hosts, while the glycosylation approach requires expressing the proteins in eukaryotes, which are more complex. In the erythropoietin case, the native protein already contains several glycosylation sites and required expression in eukaryotes, so the added glycosylation sites did not increase complexity.

To enhance FcRn-mediated recycling, researchers fuse the protein of interest with the Fc fragment of human IgG1. Besides increasing the molecular weight of the protein to slow renal filtration, this fusion enables an IgG-specific recycling mechanism. IgG proteins have an unusually long serum half life, approximately three weeks, due to a special recycling mechanism. Endothelial cells lining the bloodstream internalize proteins including IgG proteins. These proteins transfer to the endosome, then to the lysosome for degradation. However, the endosome is acidic and contains FcRn receptor proteins, which bind the Fc region of IgG proteins. Instead of degradation, this complex moves back to the cell surface where the higher pH of the blood weakens the binding between IgG and FcRn, causing release of the IgG back to the bloodstream.

This recycling mechanism accounts for the longer half-life of IgG in the blood. Etanercept is an example of a protein therapeutic (anti-inflammatory for rheumatoid arthritis) where the active protein is fused to the Fc fragment of human IgG to increase serum half life, Fig 14.[52e] Two copies of the active protein (extracellular domain of p75 tumor necrosis factor receptor) are fused to the Fc fragment of human IgG, which increases serum half life.

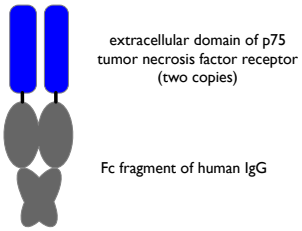

Fig 14. Etanercept is a dimeric TNF-receptor-IgG Fc fragment fusion chimera. Two copies of the extracellular domain of tumor necrosis factor receptor (blue) increase its affinity 100-4000-fold over monomeric counterparts, while the single copy of the Fc fragment (gray) increases serum half life by enabling recycling of filtered protein back to the bloodstream.

Cleavage of proteins by proteases involves two steps similar to aggregation discussed above, eq 5.20. First, at least part of the protein must unfold so that it can fit into the active site of the protease. This unfolding is usually reversible. Second, the protease must cleave the protein. This cleavage is irreversible, as is the analogous aggregation step above. Some proteases are non-specific so amino acid substitutions in the protein will not slow down proteolysis, but other proteases cleave only at specific locations. For example, trypsin cleaves mainly after a lysine or arginine residue. Engineering to remove these readily cleaved sites on the protein surface decreases proteolysis by these proteases. PeptideCutter is a web tool to predict common protease cleavage sites at: https://web.expasy.org/peptide_cutter/.[52f]

5.7 Concluding remarks

Most single substitutions increase the stability of a protein by ≤1 kcal/mol (3-4 °C increase in melting temperature), so large stabilizations require multiple stabilizing mutations. In many cases, especially when the substitutions are far from one another, their stabilization effects will be approximately additive. For example, a heat-stabilized α-amylase contains at least ten amino acid substitutions, Table 5.7.[53] This enzyme catalyzes the hydrolysis of starch to glucose oligomers in the manufacture of corn syrup. High temperatures (90 °C) increase the solubility of the starch and speed up hydrolysis, but require a heat-tolerant α-amylase. Stabilizing substitutions include replacement of amino acid residues that can undergo chemical modification, substitutions to stabilize the native form, substitutions to destabilize the unfolded form and substitutions discovered by random mutagenesis where the stabilization mechanism is unknown.

Table 5.7 Heat-stabilizing substitutions in α-amylase from *Bacillus licheniformis*.

approach	specifics
prevent chemical modification	remove oxidation (M197) and deamidation sites (N188, Q264)
stabilize N	bury Ca^{2+} ion (A181T), minimize electrostatic destabilization (H156Y)
destabilize U	introduce proline (R124P)
random mutagenesis	M15T, H133I, N188S, A209V

In another example, twelve substitutions in a halohydrin dehalogenase increased its apparent melting temperature by 28 °C and increased its ability to tolerate organic solvents.[45] The researchers first identified stabilizing single substitutions and then combined the best twelve into a single variant. The increased stability was mainly due to redistributed surface charges and improved interactions between subunits in this homotetrameric enzyme.

References

1. E. Vazquez-Figueroa, V. Yeh, J. M. Broering, J. F. Chaparro-Riggers, A. S. Bommarius (2008) Thermostable variants constructed via the structure-guided consensus method also show increased stability in salts solutions and homogeneous aqueous-organic media, *Prot. Eng. Design Select.* **21**, 673–80; https://doi.org/10.1093/protein/gzn048

2. S. Mayer, S. Rüdiger, H. C. Ang, A. C. Joerger, A. R. Fersht (2007) Correlation of levels of folded recombinant p53 in *Escherichia coli* with thermodynamic stability *in vitro*. *J Mol Biol.* **372**, 268–76; https://doi.org/10.1016/j.jmb.2007.06.044

3. J. D. Bloom, S. T. Labthavikul, C. R. Otey, F. H. Arnold (2006) Protein stability promotes evolvability. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 5869–74; https://doi.org/10.1073/pnas.0510098103

4. M. W. Adams, R. M. Kelly (1998) Finding and using hyperthermophilic enzymes. *Trends Biotechnol.* **16**, 329–32; https://doi.org/10.1016/S0167-7799(98)01193-7

5. A. Merz, M.-C. Yee, H. Szadkowski, G. Pappenberger, A. Crameri, W. P. C. Stemmer, C. Yanofsky, K. Kirschner (2000) Improving the catalytic activity of a thermophilic enzyme at low temperatures. *Biochemistry* **39**, 880–9; https://doi.org/10.1021/bi992333i

5a. Z. Wang, J. Moult (2001) SNPs, protein structure, and disease. *Hum. Mutat.* **17**, 263–70. https://doi.org/10.1002/humu.22

5b. A. C. Joerger, H. C. Ang, A. R. Fersht (2006) Structural basis for understanding oncogenic p53 mutations and designing rescue drugs. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 15056–61; https://doi.org/10.1073/pnas.0607286103

6. C. N. Pace, K. L. Shaw (2000) Linear extrapolation method of analyzing solvent denaturation curves, *Prot. Struct. Func. Bioinform.*, **41 S4**, 1-7; https://doi.org/10.1002/1097-0134(2000)41:4+<1::AID-PROT10>3.0.CO;2-2

7. M. C. Baxa, E. J. Haddadian, J. M. Jumper, K. F. Freed, T. R. Sosnick (2014) *Proc. Natl. Acad. Sci., U. S. A.* **111**, 15396–401; https://doi.org/10.1073/pnas.1407768111

7a. A. Horovitz, A. R. Fersht (1992) Co-operative interactions during protein folding. J. Mol. Biol. 224, 733–40; https://doi.org/10.1016/0022-2836(92)90557-Z

8. J. M. Scholtz, G. R. Grimsley, C. N. Pace (2009) Solvent denaturation of proteins and interpretations of the m

value, *Meth. Enzymol.*, **466**, 549–65; https://doi.org/10.1016/S0076-6879(09)66023-7

9. A. R. Fersht (1998) Protein stability in *Structure and Mechanism in Protein Science*, Freeman, Chapter 17.

10. W. Becktel, J. Schellman (1987) Protein stability curves. *Biopolymers* **26**, 1859–77; https://doi.org/10.1002/bip.360261104

11. A. D. Robertson, K. P. Murphy (1997) Protein structure and the energetics of protein stability, *Chem. Rev.* **97**, 1251-67; https://doi.org/10.1021/cr960383c

12. A. Pastore, S. R. Martin, A. Politou, K. C. Kondapalli, T. Stemmler, P. A. Temussi (2007) Unbiased cold denaturation: Low- and high-temperature unfolding of yeast frataxin under physiological conditions. *J. Amer. Chem. Soc.* **129** 5374–5. http://doi.org/10.1021/ja0714538

13. A. Razvi, J. M. Scholtz, (2006) Lessons in stability from thermophilic proteins. *Protein Sci.* **15**, 1569–78; https://doi.org/10.1110/ps.062130306

14. V. G. H. Eijsink, A. Bjørk, S. Gåseidnes, R. Sirevåg, B. Synstad, B. van den Burg. G. Vriend (2004) Rational engineering of enzyme stability. *J. Biotechnol.* **113**, 105–20; https://doi.org/10.1016/j.jbiotec.2004.03.026

15. C. Dumon, A. Varvak, M. A. Wall, J. E. Flint, R. J. Lewis, J. H. Lakey, C. Morland, P. Luginbühl, S. Healey, T. Todaro, G. DeSantis, M. Sun, L. Parra-Gessert, X. Tan, D. P. Weiner, H. J. Gilbert (2008) Engineering hyperthermostability into a GH11 xylanase is mediated by subtle changes to protein structure. *J. Biol. Chem.* **283**, 22557–64; https://doi.org/10.1074/jbc.M800936200

16. B. K. Shoichet, W. A. Baase, R. Kuroki, B. W. Matthews (1995) A relationship between protein stability and protein function. *Proc. Natl. Acad. Sci. U. S. A.* **92**, 452–6; https://doi.org/10.1073/pnas.92.2.452

17. M. Lehmann, L. Pasamontes, S. F. Lassen, M. Wyss (2000) The consensus concept for thermostability engineering of proteins. *BBA-Protein Struct. M.* **1543**, 408–15; https://doi.org/10.1016/s0167-4838(00)00238-7

17a. T. J. Magliery (2015) Protein stability: computation, sequence statistics, and new experimental methods. *Curr. Opin. Struct. Biol.* **33**, 161–8. https://doi.org/10.1016/j.sbi.2015.09.002

17b. B. J. Jones, C. N. E. Kan, C. Luo, R. J. Kazlauskas (2020) Consensus Finder web tool to predict stabilizing substitutions in proteins, *Meth. Enzymol.* **643**, 129–48. https://doi.org/10.1016/bs.mie.2020.07.010

17c. M. Musil, J. Stourac, J. Bendl, J. Brezovsky, Z. Prokop, J. Zendulka, T. Martinek, D. Bednar, J. Damborsky (2017) FireProt: web server for automated design of thermostable proteins. *Nucleic Acids Res.* **45**, W393–9. https://doi.org/10.1093/nar/gkx285

18. M. W. Pantoliano, M. Whitlow, J. F. Wood, S. W. Dodd, K. D. Hardman, M. L. Rollence, P. N. Bryan (1989) Large increases in general stability for subtilisin BPN' through incremental changes in the free-energy of unfolding. *Biochemistry* **28**, 7205–13; https://doi.org/10.1021/bi00444a012

19. D. L. Trudeau, M. Kaltenbach, D. S. Tawfik (2016) On the potential origins of the high stability of reconstructed ancestral proteins. *Mol. Biol. Evol.* **33**, 2633–41; https://doi.org/10.1093/molbev/msw138

19a. J. Clarke, K. Henrick, A. R. Fersht (1995) Disulfide mutants of barnase. I. Changes in stability and structure assessed by biophysical methods and X-ray crystallography. *J. Mol. Biol.* **253**, 493–504.

20. C. N. Pace, G. R. Grimsley, J. A. Thomson, B. J. Barnett (1988) Conformational stability and activity of ribonuclease T1 with zero, one, and two intact disulfide bonds. *J. Biol. Chem.* **263**, 11820–5. http://www.jbc.org/content/263/24/11820.full.pdf

21. J. M. Thornton (1981) Disulphide bridges in globular proteins. *J. Mol. Biol.* **151**, 261–87. https://doi.org/10.1016/0022-2836(81)90515-5

21a. L. J. Perry, R. Wetzel (1984) Disulfide bond engineered into T4 lysozyme: stabilization of the protein toward thermal inactivation. *Science* **226**, 555-7. https://www.jstor.org/stable/1693872

21b. M. Matsumura, W. J. Becktel, M. Levitt, B. W. Matthews (1989) Stabilization of phage T4 lysozyme by engineered disulfide bonds. *Proc. Natl. Acad. Sci. U. S. A.* **86**, 6562-6. https://doi.org/10.1073/pnas.86.17.6562

22. B. Hazes, B. W. Dijkstra (1988) Model building of disulfide bonds in proteins with known three-dimensional structure. *Protein Eng.* Des. Sel. **2**, 119–25. https://doi.org/10.1093/protein/2.2.119

23. M. G. Pikkemaat, A. B. M. Linssen, H. J. C. Berendsen, D. B. Janssen (2002) Molecular dynamics simulations as a tool for improving protein stability. *Protein Eng. Des. Sel.* **15**, 185–92. https://doi.org/10.1093/protein/15.3.185

23a. E. J. Stewart, F. Åslund, J. Beckwith (1998) Disulfide bond formation in the *Escherichia coli* cytoplasm: an invivo role reversal for the thioredoxins. *EMBO J.* **17**, 5543–50. https://doi.org/10.1093/emboj/17.19.5543

24. H.-X. Zhou (2004) Loops, linkages, rings, catenanes, cages, and crowders: entropy-based strategies for stabilizing proteins. *Acc. Chem. Res.* **37**, 123–30. https://doi.org/10.1021/ar0302282

25. B. W. Matthews, H. Nicholson, J. W. Becktel (1987) Enhanced protein thermostability from site-directed mutations that decrease the entropy of unfolding. *Proc. Natl. Acad. Sci. U. S. A.* **84**, 6663–7. https://doi.org/10.1073/pnas.84.19.6663

26. K. Watanabe, Y. Suzuki (1998) Protein thermostabilization by proline substitutions. *J. Mol. Catal. B: Enzym.* **4**, 167–80. https://doi.org/10.1016/s1381-1177(97)00031-3

27. H. Fu, G. R. Grimsley, A. Razvi, J. M. Scholtz, C. N. Pace (2009) Increasing protein stability by improving beta-turns. *Proteins* **77**, 491–8. https://doi.org/10.1002/prot.22509

28. K. A. Scott, D. O. V. Alonso, S. Sato, A. R. Fersht, V. Daggett (2007) Conformational entropy of alanine versus glycine in protein denatured states. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 2661–6. https://doi.org/10.1073/pnas.0611182104

28a. Z. Liu, S. Lemmonds, J. Huang, M. Tyagi, L. Hong, N. Jain (2018) Entropic contribution to enhanced thermal stability in the thermostable P450 CYP119. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E10049–58; https://doi.org/10.1073/pnas.1807473115

29. M. T. Reetz, J. D. Carballeira,, A. Vogel (2006) Iterative saturation mutagenesis on the basis of B factors as a strategy for increasing protein thermostability. *Angew. Chem. Int. Ed. Engl.* **45**, 7745–51. https://doi.org/10.1002/anie.200602795

30. R. J. Floor, H. J. Wijma, D. I. Colpa, A. Ramos-Silva, P. A. Jekel, W. Szymanski, B. L. Feringa, S. J. Marrink, D. B. Janssen (2014) Computational library design for increasing haloalkane dehalogenase stability. *ChemBioChem* **15**, 1660–72. https://doi.org/10.1002/cbic.201402128

31. Z. Xiao, H. Bergeron, S. Grosse, M. Beauchemin, M.-L. Garron, D. Shaya, T. Sulea, M. Cygler, P. C. K. Lau (2008) Improvement of the thermostability and activity of a pectate lyase by single amino acid substitutions, using a strategy based on melting-temperature-guided sequence alignment. *Appl. Environ. Microbiol.* **74**, 1183–9. https://doi.org/10.1128/AEM.02220-07

32. E. Bae, G. Phillips, Jr. (2005) Identifying and engineering ion pairs in adenylate kinases. *J. Biol. Chem.* **279**, 30943–8. https://doi.org/10.1074/jbc.M504216200

33. G. A. Petsko (2001) Structural basis of thermostability in hyperthermophilic proteins, or "There's more than one way to skin a cat". *Meth. Enzymol.* **334**, 469–78. https://doi.org/10.1016/S0076-6879(01)34486-5

34. A. V. Gribenko, M. Patel, J. Liu, S. McCallum, C. Wang, G. I. Makhatadze (2009) Rational stabilization of enzymes by computational redesign of surface charge–charge interactions, *Proc. Natl. Acad. Sci. U. S. A.* **106**, 2601–6. https://doi.org/10.1073/pnas.0808220106

35. C.-H. Chan, C. C. Wilbanks, G. I. Makhatadze, K.-B. Wong (2012) Electrostatic contribution of surface charge residues to the stability of a thermophilic protein: benchmarking experimental and predicted pK_a values. *PLoS*

ONE, 2012, **7**, e30296. https://doi.org/10.1371/journal.pone.0030296

36. C. R. Søndergaard, M. H. M. Olsson, M. Rostkowski, J. H. Jensen (2011) Improved treatment of ligands and coupling effects in empirical calculation and rationalization of pK_a values. *J. Chem. Theory Comput.*, 2011, **7**, 2284–95; https://doi.org/10.1021/ct200133y. A web server version of PROPKA 3.0 is available at http://nbcr-222.ucsd.edu/pdb2pqr_2.0.0/ (accessed April 2018).

37. K. W. Kaufmann, G. H. Lemmon, S. L. DeLuca, J. H. Sheehan, J. Meiler (2010) Practically useful: What the Rosetta protein modeling suite can do for you. *Biochemistry* **49**, 2987–98. https://doi.org/10.1021/bi902153g

38. J. Schymkowitz, J. Borg, F. Stricher, R. Nys, F. Rousseau, L. Serrano (2005) The FoldX web server: an online force field. *Nucl. Acid. Res.* **33**, W382–8. https://doi.org/10.1093/nar/gki387

39. H. Arabnejad, M. Dal Lago, P. A. Jekel, R. J. Floor, A.-M. W. H. Thunnissen, A. C. Terwisscha van Scheltinga, H. J. Wijma, D. B. Janssen (2017) A robust cosolvent-compatible halohydrin dehalogenase by computational library design. *Protein Eng. Des. Sel.* **30**, 173–87; https://doi.org/10.1093/protein/gzw068

40. A. Korkegian, M. E. Black, D. Baker, B. L. Stoddard (2005) Computational thermostabilization of an enzyme. *Science* **308**, 857–60; https://doi.org/10.1126/science.1107387

40a. A. Goldenzweig, M. Goldsmith, S. E. Hill, O. Gertman, P. Laurino, Y. Ashani, O. Dym, T. Unger, S. Albeck, J. Prilusky, R. L. Lieberman, A. Aharoni, I. Silman, J. L. Sussman, D. S. Tawfik, S. J. Fleishman (2016) Automated structure- and sequence-based design of proteins for high bacterial expression and stability. *Mol. Cell* **63**, 337–46; https://doi.org/10.1016/j.molcel.2016.06.012

41. D. Gao, D. L. Narasimhan, J. Macdonald, M.-C. Ko, D. W. Landry, J. H. Woods, R. K. Sunahara, C.-G. Zhan (2009) Thermostable variants of cocaine esterase for long-time protection against cocaine toxicity. *Mol. Pharmacol.* **75**, 318–23; https://doi.org/10.1124/mol.108.049486

42. B. van den Burg, G. Vriend, O. R. Veltman, V. G. H. Eijsink (1998) Engineering an enzyme to resist boiling *Proc. Natl. Acad. Sci. U. S. A.* **95**, 2056–60; https://doi.org/10.1073/pnas.95.5.2056

43. A. Bjørk, B. Dalhus, D. Mantzilas, R. Sirevåg, V. G. H. Eijsink (2004) Large improvement in the thermal stability of a tetrameric malate dehydrogenase by single point mutations at the dimer-dimer interface. *J. Mol. Biol.* **341**, 1215–26; https://doi.org/10.1016/j.jmb.2004.06.079

44. T. Kabashima, Y. Li, N. Kanada, K. Ito, T. Yoshimoto (2001) Enhancement of the thermal stability of pyroglutamyl peptidase I by introduction of an intersubunit disulfide bond. *Biochim. Biophys. Acta* **1547**, 214−20; https://doi.org/10.1016/s0167-4838(01)00185-6

45. N. Chennamsetty, V. Voynov, V. Kayser, B. Helk, B. L. Trout (2010) Prediction of aggregation prone regions of therapeutic proteins. *J. Phys. Chem. B*, **114**, 6614–24; http://doi.org/10.1021/jp911706q

46. J. van Durme, G. de Baets, R. van der Kant, M. Ramakers, A. Ganesan, H. Wilkinson, R. Gallardo, F. Rousseau, J. Schymkowitz (2016) Solubis: A webserver to reduce protein aggregation through mutation. *Protein Eng. Des. Sel.* **29**, 285–9; http://doi.org/10.1093/protein/gzw019

46b. A. Kuriata, V. Iglesias, J. Pujols, M. Kurcinski, S. Kmiecik, S. Ventura (2019) Aggrescan3D (A3D) 2.0: prediction and engineering of protein solubility. *Nucl. Acids Res.* **47**, W300–7. https://doi.org/10.1093/nar/gkz321

47. N. Zhang, W.-C. Suen, W. Windsor, L. Xiao, V. Madison, A. Zaks (2003) Improving tolerance of *Candida antarctica* lipase B towards irreversible thermal inactivation through directed evolution. *Protein Eng. Des. Sel.* **16**, 599–605. http://doi.org/10.1093/protein/gzg074

48. M. S. Lawrence, K. J. Phillips, D. R. Liu (2007) Supercharging proteins can impart unusual resilience. *J. Am. Chem. Soc.* **129**, 10110-2; https://doi.org/10.1021/ja071641y

49. N. E. Robinson (2002) Protein deamidation. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 5283–8. http://doi.org/10.1073/

pnas.082102799

50. 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.full.pdf

51. D. F. Mark, S. D. Lu, A. A. Creasey, R. Yamamoto, L. S. Lin (1984) Site-specific mutagenesis of the human fibroblast interferon gene. *Proc. Natl. Acad. Sci. U. S. A.* **81**, 5662–6; http://www.pnas.org/content/81/18/5662.long

52. 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; http://doi.org/10.1038/nbt0891-738

52a. R.E. Kontermann (2009) Strategies to extend plasma half-lives of recombinant antibodies. *BioDrugs* **23**, 93–109. https://doi.org/10.2165/00063030-200923020-00003

52b. W. R. Strohl (2015) Fusion proteins for half-life extension of biologics as a strategy to make biobetters. *BioDrugs* **29**, 215–39. https://doi.org/10.1007/s40259-015-0133-6

52c. J. C. Egrie, E. Dwyer, J. K. Browne, A. M. Hitz, A. Lykos (2003) Darbepoetin alfa has a longer circulating half-life and greater in vivo potency than recombinant human erythropoietin. *Ex. Hematol.* **31**, 290–299. https://doi.org/10.1016/S0301-472X(03)00006-7

52d. V. Schellenberger, C. W. Wang, N. C. Geething, B. J. Spink, A. Campbell, W. To, M. D. Scholle, Y. Yin, Y. Yao, O. Bogin, J. L. Cleland, J. Silverman, W. P. C. Stemmer. (2009) A recombinant polypeptide extends the in vivo half-life of peptides and proteins in a tunable manner. *Nat. Biotechnol.* **27**, 1186–90. https://doi.org/10.1038/nbt.1588

52e. M. Feldmann (2002) Development of anti-TNF therapy for rheumatoid arthritis. *Nat. Rev. Immunol.* **2**, 364–371. https://doi.org/10.1038/nri802

52f. E. Gasteiger, C. Hoogland, A. Gattiker, S. Duvaud, M. R. Wilkins, R. D. Appel, A. Bairoch (2005) Protein identification and analysis tools on the ExPASy server in J. M. Walker, Ed., *The Proteomics Protocols Handbook*, Humana Press, Totowa, NJ, pp 571–607. https://doi.org/10.1385/1-59259-890-0:571

53. J. E. Nielsen, T. V. Borchert (2000) Protein engineering of bacterial α-amylases. *Biochim. Biophys. Acta Protein Struct. Mol. Enzymol.* **1543**, 253–74; https://doi.org/10.1016/s0167-4838(00)00240-5

Supporting information

Contents

Derivation of equation relating spectral changes and equilibrium constant, eq. 5.5.

Example of linear extrapolation of protein unfolding data in urea.

Example of using Rosetta Design to predict a stabilizing substitution.

Derivation of equation relating spectral changes and equilibrium constant, eq. 5.5.

Deriving the relationship between spectral changes and equilibrium constant is straightforward. The protein is either in the native state or in the denatured state, so the sum of the two fractions is one: $F_N + F_D = 1$. The observed fluorescence at each urea concentration, Y_{obs}, is the sum of the contributions from the native and denatured states:

$$Y_{obs} = Y_N \cdot F_N + Y_D \cdot F_D \qquad \text{(S1)}$$

Substituting $F_N = 1 - F_D$ yields:

$$F_D = (Y_{obs} - Y_N)/(Y_D - Y_N) \quad \text{(S2)}$$

Similarly substituting $F_D = 1 - F_N$ yields:

$$F_N = (Y_D - Y_{obs})/(Y_D - Y_N) \quad \text{(S3)}$$

Dividing these two equations yields eq S4, which is the same as eq 8 in the main text.

$$K_{unfold} = F_D/F_N = (Y_{obs} - Y_N)/(Y_D - Y_{obs}) \quad \text{(S4)}$$

Example of linear extrapolation of protein unfolding data in urea

Protein A was dissolved in solutions of different concentrations of urea and allowed to reach equilibrium. The fluorescence spectra of these solutions showed an increase in fluorescence at 2-5 M urea, Fig. S1. Using this fluorescence data, calculate the Gibbs energy of unfolding in pure water for protein A.

[urea]	Y_{obs}
0	72
1	70
2	81
3	95
4	140
5	150
6	155
7	152
8	156
9	156

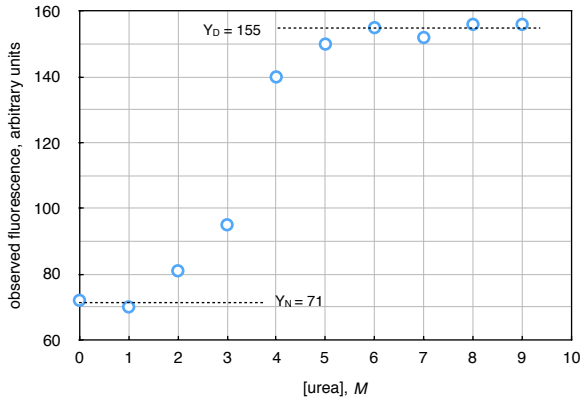

Fig. S1. Typical experimental data from a urea-unfolding experiment to measure the Gibbs energy of unfolding of a protein. The intensity of fluorescence can either increase or decrease upon protein unfolding; in this example, it increases. Analysis and replot of this data are in Fig 9 below.

Solution

1. Estimate the fluorescence of the native and denatured states.

The fluorescence of the native state is the flat part of the curve before unfolding begins. The average of the values at 0 and 1 M urea yields $Y_N = 71$. The fluorescence of the denatured state is the flat part of the curve after unfolding is complete. The average of the values at 6-9 M urea yields $Y_D = 155$.

2. Estimate the equilibrium constant between the native and denatured forms at each urea concentration in the unfolding region.

For this example: $K_{unfold} = (Y_{obs} - 71)/(155 - Y_{obs})$, so one can calculate K_{unfold} at different urea concentrations by substituting the experimental values of Y_{obs}. This procedure yields the values in the table below.

[urea]	K_{unfold}	ΔG_{unfold} (kcal/mol)
2	0.14	1.2
3	0.40	0.54
4	4.6	–0.90
5	16	–1.6

3. Convert the equilibrium constant to Gibbs energy according to eq 9 in main text using R = 1.987 cal/mol °K and T = 298 °K, which yields the data in the table above.

4. Plot the Gibbs energy of unfolding as a function of urea concentration (eq 10 in main text), fit the data to a straight line and extrapolate to pure water ([urea] = 0). The slope of the line in the linear extrapolation plot is called the m-value.

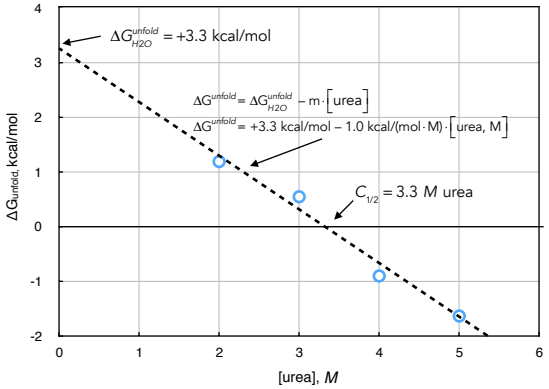

Fig S2. Linear extrapolation plot of the data from Fig S1 above yields $\Delta G_{unfold} = 3.3 - 1.0[urea]$ as the best fit.

The extrapolation yields a y-intercept of +3.3 kcal/mol, which is the Gibbs energy of unfolding of this protein in pure water. This protein is less stable than typical proteins, but even for this one, the equilibrium amount of denatured protein in pure water is only 0.38%, which would be difficult to measure directly.

Example of using Rosetta Design to predict a stabilizing substitution

Predict a single amino acid substitution to stabilize the esterase SABP2 (structure pdb code = 1y7i). Previous modeling suggested that a replacement at position 60 may stabilize the protein. Use the RosettaDesign server http://rosettadesign.med.unc.edu to predict the substitution.[1] The server requires registration, but it is free to use.

Computational approach: Test all 19 possible replacements at position 60 to identify any that have lower energy while repacking the side chains at nearby residues to adjust for the amino acid substitutions. To do these calculations, we need to provide Rosetta with 1) the pdb file of SABP2 and 2) instructions for the desired calculation. An example set of instructions is below:

NATRO
start

47 A NATAA
56 A NATAA
57 A NATAA
58 A NATAA
59 A NATAA
60 A ALLAA
61 A NATAA
62 A NATAA
63 A NATAA
64 A NATAA

To send these instructions to Rosetta Design, copy the text above and save them as a text file with the name SABP2_LEU60.res or something similar. The file extension must be .res. Submit the pdb file and the res file to the Rosetta design server.

Hint: The results do not indicate the instructions that you used for the calculations. Once you submit the job and get a job number, rename your res file by adding the job number to later match it to the results (e.g., 35161_SABP2_LEU60.res)

Explanation of the instructions:

The first line "NATRO" is the default for the whole protein. NATRO means natural rotamer and tells Rosetta Design to keep the existing amino acid and the existing side chain rotamer. That is, don't change anything.

The next line, "start" indicates the start of specific instructions for particular amino acids.

The following lines are all formatted as [amino acid number] [chain letter] [instructions]. In this case, several amino acids from chain A are being redesigned.

At positions 47, 56-59, 61-64, NATAA means native amino acid and tells Rosetta design to keep the existing amino acid, but allow the side chain to move to a new rotamer. This flexibility allows the surrounding region to adjust as the amino acid at position 60 changes.

At position 60, ALLAA means all amino acids and tells Rosetta Design to try all amino acids and all rotamers of these amino acids.

Results

The results consist of a log file and a pdb file. Open the pdb file with a text editor.

Scroll to residue 60 of chain A; it has been redesigned to Arg.

Scroll to nearly the end of the file to the row starting with the word 'pose'. The last number in this row is the total energy = 41.7 (the units are Rosetta Energy Units)

As a control, run another calculation that only readjusts the positions of the side chains. The res file is the same, except line 60 should read: 60 A NATAA. In this case the total energy is higher, 41.8, confirming that the Arg substitution is an improvement. Experiments confirmed that the Leu60Arg substitution improved heat stability at 60 °C by about a factor of two.[2]

More advanced calculations

1. Try the calculation several times

Change the "Number of independent trajectories" from 1 to 5, which will run the same calculation five times. Each calculation may give slightly different results. You will receive five pdb files.

2. Expand the range of possible side chain rotamers

The conformation of the amino acid side chains is defined by angles chi 1, chi 2, chi 3, and chi 4. For phenylalanine, two angles, chi 1 and chi 2, define the conformation; for lysine, all four angles are needed to define the conformation. The command EX 1 EX 2 tells Rosetta Design to include additional rotamers for chi 1 and chi 2 at ±1 standard deviation from the mean chi angle for each rotamer for buried residues. The command USE_INPUT_SC means include the native rotamer, since it may be an unusual one. For 60, there is no USE_INPUT_SC because the initial amino acid is being replaced. The EX 1 and EX 2 tells Rosetta Design to try extra rotamers. In this case, the prediction is the same - Arg - but the energy values will be lower.

NATRO
start

47 A NATAA USE_INPUT_SC EX 1 EX 2
56 A NATAA USE_INPUT_SC EX 1 EX 2
57 A NATAA USE_INPUT_SC EX 1 EX 2
58 A NATAA USE_INPUT_SC EX 1 EX 2
59 A NATAA USE_INPUT_SC EX 1 EX 2
60 A ALLAA EX 1 EX 2
61 A NATAA USE_INPUT_SC EX 1 EX 2
62 A NATAA USE_INPUT_SC EX 1 EX 2
63 A NATAA USE_INPUT_SC EX 1 EX 2
64 A NATAA USE_INPUT_SC EX 1 EX 2

Other suggestions

It may be easier to view the results pdb file if you open the file as a spreadsheet with Excel or similar program. (The data is in tabular form.) The energy scores section has energy values broken down by a number of categories across different columns and values for each amino acid going down in rows.

References

1. Y. Liu and B. Kuhlman, *Nucl. Acids Res.*, 2006, **34**, W235−W238.

2. B. J. Jones, H. Y. Lim, J. Huang and R. J. Kazlauskas, *Biochemistry*, 2017, **56**, 6521−6532.