Amino Acid Substitutions: Effects on Protein Stability

Secondary article

Zhiping Weng, *Boston University, Boston, Massachusetts, USA*
Charles DeLisi, *Boston University, Boston, Massachusetts, USA*

The ability of a protein to function in its biochemical role(s) is determined by its geometry, which in turn is determined by the amino acid sequence and the environment. Amino acid mutations affect the thermodynamics of folding and their effects can be investigated experimentally by such techniques as site-specific mutation, random point mutations and shuffling.

Article Contents

Introduction

A living cell can be viewed as a biochemical factory, with as many as 100 000 distinct proteins providing the hardware to carry out cellular processes. The roles of these molecular machines include such diverse activities as pumping (e.g. voltage-gated membrane channels), motor functions (e.g. flagella), amplification (e.g. cyclic AMP), and catalysis. The ability of a protein to function in one or another role is determined by its geometry, which in turn is determined by the amino acid sequence and the environment.

A protein is said to be in its native state under normal laboratory conditions (room temperature; pH near 7; ionic strength near 0.15 mol L^{-1}). Its three-dimensional structure under these conditions invariably consists of a congeries of compactly folded stretches of regular secondary structure. Environmental stress can cause a protein to lose its native structure and hence to denature to a state that is much less compact, somewhat more flexible, and very highly hydrated. The stability of a protein reflects the extent to which its conformation resists change when subject to stress.

For a large number ($\sim 10^{20}$ molecules per litre) of proteins with the same sequence under normal experimental conditions, all possible conformations will in principle be present. These will range from the native (or folded) state, to the denatured (or unfolded) state, and include all partially ordered states. The overwhelming majority, however, are in their native state. Under a constant environmental stress, such as elevated temperature, the equilibrium distribution of proteins among the states will shift towards the denatured state. Continued increases in temperature of 20–30°C will shift the equilibrium distribution so that the denatured state becomes overwhelmingly favoured.

Although the transition from the folded to the unfolded state consists of many steps, it can, if carried out reversibly, be described in terms of a single effective equilibrium constant (eqn [1]).

$$u(\text{unfolded}) \overset{K_f}{\rightleftharpoons} f(\text{folded}) \qquad [1]$$

where K_f, the folding equilibrium constant, is defined as the ratio of the number (or concentration) of folded molecules to the number (or concentration) of unfolded molecules. The relative Gibbs free energy of folding, ΔG_f, the difference between the free energies of the folded (G_f) and the unfolded (G_u) states at constant pressure, is related to K_f via eqn [2] in a standard concentration state (taken here as 1 mol L^{-1}).

$$\Delta G_f = G_f - G_u = -RT\ln(K_f) \qquad [2]$$

where R is the gas constant (8.31 J K^{-1} mol^{-1}) and T is the temperature in kelvin. ΔG_f is a widely accepted measure of protein stability, the lower (more negative) the value the more stable the protein. ΔG_f can also be obtained from enthalpies (H_f, H_u) and entropies (S_f, S_u) (eqn [3]).

$$\Delta G_f = \Delta H - T\Delta S = (H_f - H_u) - T(S_f - S_u) \qquad [3]$$

The free energy change for folding at room temperature typically ranges between -21 and -63 kJ mol^{-1}, depending on the protein. At the higher value (-21), approximately 2 protein molecules in 10 000 will be in the denatured state at room temperature. At the lower value the corresponding number is 5 in 10^6. The unfolding free energies are high compared to the -125 to -418 kJ mol^{-1} associated with a covalent bond. Protein structures have evolved to be stable enough to function effectively, but not so stable that they cannot be readily processed and metabolized, as needed by normally active cells.

The effect of amino acid sequence on stability can be understood by comparing the measured unfolding free energy of the native, wild-type sequence, to well characterized mutant sequences. Since the *in vivo* biological function of a protein is sensitive to its conformation, activity serves as a phenotype for the mutated form. The mutant can thus be isolated and its sequence determined and correlated

with changes in stability – i.e. changes in ΔG_f ($\Delta\Delta G_f$), as well as in H_f, H_u, S_f and S_u.

Although mutation studies provide insight into the effect of sequence on stability, they cannot easily be used predictively without making the connection between the mutation and the changes in the various contributions to the overall free energy difference between the folded and unfolded states.

The free energy change ΔG_f accompanying protein folding can conveniently be divided into two components: desolvation (ΔG_s) and conformational (ΔG_c). Each has enthalpic and entropic components as defined by eqn [3], resulting in eqn [4]. ΔG_s has both enthalpic (ΔH_s) and entropic (ΔS_s) components, and the enthalpy change has both van der Waals (vdW) and electrostatic components. The difference between vdW energies in the folded and unfolded state tends to be small compared to the electrostatic energy difference; consequently ΔH_s is almost entirely electrostatic (Vajda *et al.*, 1994; Weng *et al.*, 1996).

The change in solvation entropy (ΔS_s) results from a decrease in the surface area of the protein, and hence in the number of water molecules in contact with it, when it is folded (**Figure 1**). Even though the entropy of the folded protein is somewhat less than that of the denatured form, the favourable entropy change from freeing water is sufficient to drive folding.

The enthalpic component ΔH_c of ΔG_c reflects the changes in covalent energy terms, such as the bond streching, bending and rotation energies upon folding. The folded protein conformation may be strained and the above terms may not have the lowest value. The conformational entropy change ΔS_c stems from the fact that both the backbone and the side-chain of an unfolded protein can take on multiple conformations that have more or less the same enthalpy, while the backbone and the interior side-chains of a folded protein have only a single conformation.

A mutation can affect any of the four terms in eqn [4], the expression for the free energy change upon folding.

$$\Delta G_f = \Delta H_s - T\Delta S_s + \Delta H_c - T\Delta S_c \qquad [4]$$

In so doing, it must affect the folded and unfolded states differently. Unfortunately, the unfolded state is difficult to characterize; consequently, much of what is understood applies to the folded state.

Experimental Methods

Techniques for altering protein primary structure (sequence) fall into three categories: site-specific mutagenesis, random point mutations and shuffling. Numerous variants of each category exist, but the principles described are general.

Site-specific mutagenesis

If a protein is produced in the laboratory by expression of its gene, point mutations can readily be introduced by site-directed mutagenesis, using the polymerase chain reaction (PCR). Typically, the gene has already been cloned into a plasmid. The experimental procedure is then as follows (**Figure 2**).

1. Oligonucleotides (PCR primers) with the desired mutation at the centre, but otherwise identical to the corresponding gene sequence, are synthesized. For example, if an alanine (codon GCU) is to be replaced with a valine (codon GUU), the primer would have 'GUU' at the centre but with the same flanking bases

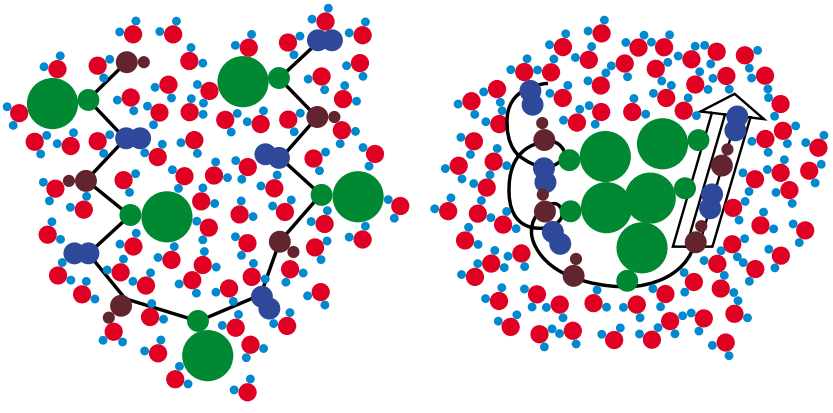

Figure 1 (a) Unfolded and fully solvated polypeptide chain; (b) folded chain, right. Large green circles represent hydrophobic side-chains; red circles represent oxygen; blue circles represent hydrogen. Arrow and coil represent ordered regions of the folded protein: α helix and β strand. Note that, in the unfolded state, both the backbone and side-chain atoms of protein interact with water molecules and in the folded state they are secluded from water by forming hydrophobic packing and internal hydrogen bonds.

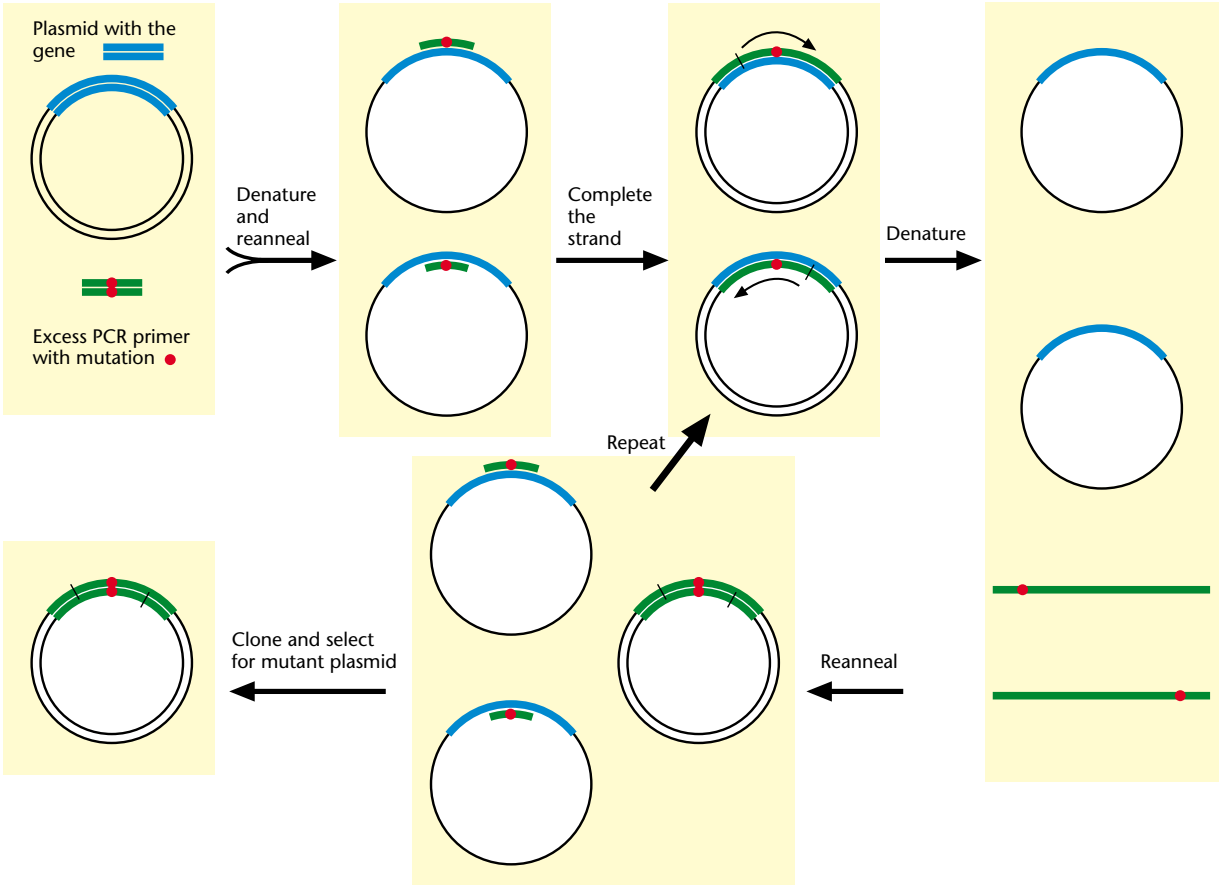

Figure 2 A wild-type plasmid (upper left) is used as a template for producing a plasmid with a specified mutation at a particular site. After temperature-induced denaturation, each primer strand binds the complementary single strand of the plasmid DNA. In the presence of polymerase, DNA replication proceeds from the 3′ end of the primer, producing two mutant plasmids, each having an unligated site (|) where replication terminates. That completes the first cycle. Another round of heating produces unmutated single-stranded circular DNA and mutated linear DNA. Reannealing produces primer bound to unmutated DNA, which serves as template for another mutated strand, and inert double-stranded mutant DNA. If there are *N* plasmids initially, each round produces 2*N* mutant strands, with 2*rN* mutant strands after *r* rounds.

as in the wild-type gene. The primers need to be sufficiently long to anneal only with the correct segment of the gene, despite the destabilizing effect of the mutation.

2. Wild-type plasmid is mixed with excess primer in a solution that includes heat-resistant polymerase and nucleoside triphosphates (NTPs). The solution is placed in a thermocycler, which is designed to change temperature rapidly and precisely.
3. Raising the temperature above 95°C separates the double-stranded plasmid and primers into single strands. The solution is then cooled to approximately 55°C so that primers can anneal with single-stranded plasmid. Since the primers are in excess, most of single-stranded plasmid will anneal with a primer instead of with its complementary strand.
4. The temperature is increased to approximately 72°C, at which temperature polymerase extends the primer to a complete strand in a matter of minutes. Except for the desired mutation, the new strand has exactly the same sequence as one of the strands in the wild-type plasmid.
5. Steps 3 and 4 are repeated approximately 20 times to make more mutant strands. Since polymerase extends only in the 5′ to 3′ direction and primers anneal only to the 3′ end of the mutant strands, the mutant strands cannot serve as primer and consequently cannot be replicated. The amplifying power of this system is therefore linear, unlike ordinary PCR, which amplifies exponentially.
6. Under ideal conditions, after 20 rounds, the number of mutant strands should be 20-fold greater than the number of wild-type strands. A mutant can form a

circular plasmid with a complementary mutant or with a wild type, except that it has a nick where ligation halted. A restriction enzyme *Dpn*I, which cuts methylated DNA, is then added to the mixture. Since the wild-type plasmid (which is usually made in *E. coli* by cloning) is methylated and the mutant plasmid made by PCR is not, all wild-type plasmids are destroyed.

7. An aliquot of the final solution is used to transform bacteria. Bacterial enzymes repair the nicks in the mutant plasmid. Since the majority of the final plasmids have mutations, a randomly picked colony has a very high chance of carrying the mutant gene.

Random mutations at specified positions

It is often desirable to investigate the effect of more than one amino acid on protein stability and function. For example, if we had reason to believe that a particular position was critical to folding, we would want to determine which if any substitutions at that position increased stability. The most direct approach is to construct 19 site-directed mutations (because there are 20 amino acids), each with the codon of a different amino acid at the centre of the primer, and measure the folding free energies of the wild type and all mutants.

An alternative is to generate all possible mutants and screen for the most stable. To do this, only the first and the last steps in the site-directed mutagenesis procedure need to be modified. The first step is modified to produce a mixture of primers with different central codons (**Figure 3**). These are obtained during synthesis simply by using a mixture of NTPs, rather than a single type of NTP, for one or more of the nucleotides of the central codon. NTPs will be selected randomly in accordance with their frequencies in the mixture, resulting in primers with different sequences.

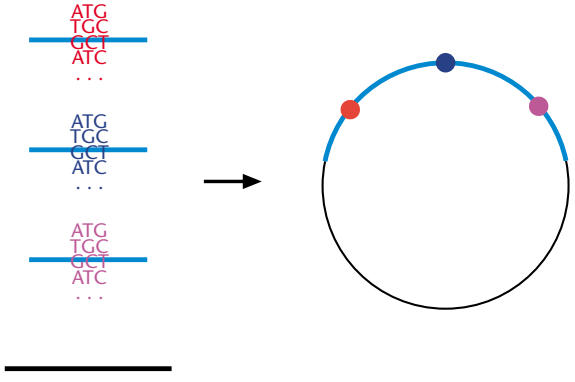

Figure 3 Production of mutant plasmids by cassette synthesis and combinatorial ligation. A gene (blue line) is excised and fragmented, in the example shown into three pieces, and different mutant codons are introduced (coloured circles) into the fragments. If codons are introduced at equal frequency, then the probability is 1/64 that a randomly picked fragment will have a particular codon. Fragments are selected in random triplets for ligation to reform an intact and triply mutated gene. The example shown can produce more than 2.6×10^5 different mutants.

Selecting the most stable mutant is more difficult, and depends on the availability of a readily identifiable phenotype. For example, the enzyme β-galactosidase can be detected by its ability to cleave a bond between galactose and an indicator dye that changes colour when the bond is cleaved. Therefore the more stable the enzyme, the longer it remains active and the more intense the colour change. Selection and amplification of the bacteria carrying the phenotype and DNA sequencing follow detection. Thus, in this example, sequences corresponding to the most intense colour change can be identified and the stabilities of the encoded proteins determined.

As this example illustrates, the selection step for random mutagenesis is independent of the number of mutations, and plasmid synthesis only needs to be modified slightly to accommodate mutations at multiple positions (Sauer, 1996). This is in contrast to site-directed mutagenesis, where effort increases linearly with the number of mutations. Random mutagenesis is therefore particularly advantageous when multiple positions must be mutated simultaneously. Since it is difficult to synthesize oligonucleotides with more than 60 bases, multiple oligonucleotides are synthesized, each with a mutation position at the centre (**Figure 3**) followed by ligation to reconstruct the entire gene. These oligonucleotides are called cassettes and the procedure is termed 'cassette mutagenesis'.

Random mutations throughout the whole gene

The easiest way to construct random mutations throughout the whole gene is to do PCR with a low-fidelity polymerase, which makes random mistakes during DNA replication. Such 'error-prone PCR' can be combined with DNA shuffling (**Figure 4**) so that diverse sequences can be rapidly generated and selected. The method is intended to mimic recombination used by nature to generate biological diversity. A pool of identical or closely related sequences is fragmented randomly, and these fragments are reassembled into full-length genes via self-priming PCR and extension. This process, called 'assembly PCR', yields crossovers between related sequences due to template switching. Such shuffling allows rapid combination of positive-acting mutations and simultaneously flushes out negative-acting mutations from the sequence pool. When coupled with effective selection, and applied iteratively, such that the output of one cycle is the input of the next cycle, DNA shuffling is an efficient process for directed molecular evolution.

DNA shuffling is a recent invention, with the ability to sample much larger sequence spaces than other mutagenesis techniques. Most of its applications have been focused on discovering mutations leading to higher activities

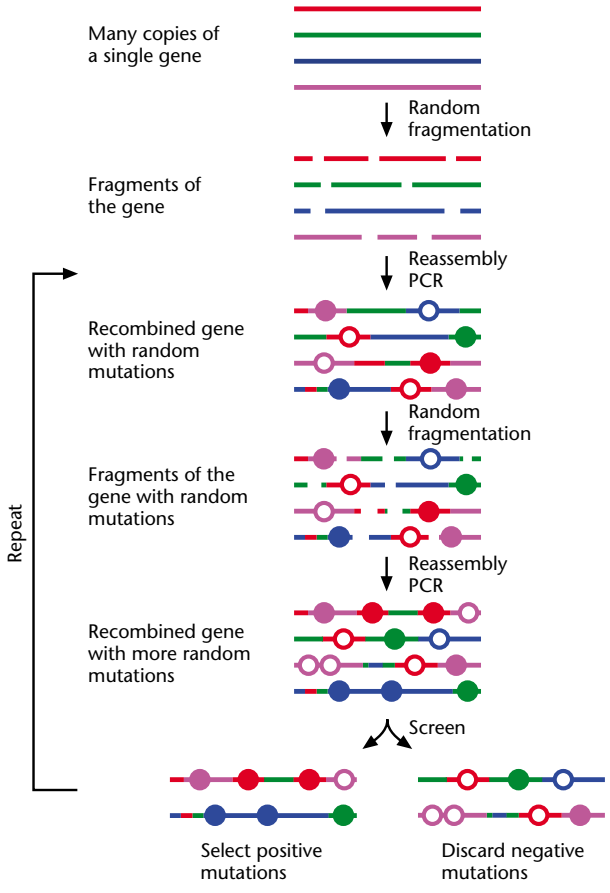

Figure 4 The generation of sequence diversity by shuffling and error-prone PCR. The genes are randomly fragmented into strands of different length. Overlapping fragments will bind one another, and the single-stranded 3′ end of the duplex will serve as template for continuation of the 3′ end of the bound partner, adding nucleotides without error correction. The resulting strands now differ from the wild type by a small fraction of point mutations, and by random recombination. Another round of random fragmentation and PCR is followed by selection of desired mutants (for example increased affinity for some substrate) and repetition of the cycle on the selected strands.

(e.g. resistance to antibiotics, higher enzymatic activities and stronger cell fluorescence signals). Dramatic improvements in activity have been achieved using DNA shuffling, and it will not be surprising if this technique uncovers mutated proteins that are much more stable than the wild type.

Stabilizing/Destabilizing Mutations

Sequences of the same protein in different species can be very different, more so if the species are evolutionarily distant. For example, in yeasts and humans, the sequences of calmodulin, an important calcium-binding protein, differ in 42% of positions. In fact, yeast and human sequences of most proteins differ at this level, yet many yeast proteins can be replaced by a human counterpart – replacing the yeast molecule with a human molecule does not produce any observable change in phenotype. This means that the sequence space is so large that a function can be achieved by many different sequences. It is therefore of some interest to understand which mutations a protein tolerates, and why. If we can achieve a predictive understanding, we can design mutant proteins with desired properties.

The easiest and most direct procedure for carrying out such a programme is to construct a series of mutants of a protein *in vitro* and measure their stabilities. A few proteins (e.g. Arc repressor, T4 lysozyme, and staphylococcal nuclease) have been studied extensively using site-directed mutagenesis and cassette mutagenesis. Most mutations (or sets of mutations) have been found to be destabilizing, though some can lead to slightly more stable mutants (Gassner *et al.*, 1996). Very few mutants are actually significantly more stable than the wild type, and those that are sometimes suffer from slow folding kinetics or impaired biological function.

Effects on the Native and/or Denatured State: Changes in Solvent-exposed versus Buried Residues

Most mutations affect both folded and unfolded states. If a mutation increases the free energy of the folded state while simultaneously decreasing the free energy of the unfolded state, it increases the overall folding free energy and is thus destabilizing. A mutation that decreases the free energy of the folded state and simultaneously increases the free energy of the unfolded state is stabilizing. However, if a mutation increases (or decreases) the free energies of both the folded and unfolded states, it can be stabilizing or destabilizing, depending upon which of the two quantities dominates. As seen from eqn [4], it is the combination of the effects on the solvation and conformational enthalpies and entropies that determines the actual free energy change.

Our state of understanding is most easily summarized by thinking of mutated positions as either on the surface of the folded protein or buried. Effects on the enthalpies and entropies of the folded and unfolded states are analysed qualitatively in **Table 1**. Subscripts c and s indicate that the mutated position is buried (core) or surface. $\Delta\Delta G_f$ is the difference between the folding free energy of the mutant (mut) and the wild type (wt), and can be further expanded

Table 1 Thermodynamic changes accompanying point mutations

Mutation	ΔH_s^f	ΔS_s^f	ΔH_c^f	ΔS_c^f	ΔH_s^u	ΔS_s^u	ΔH_c^u	ΔS_c^u	$\Delta\Delta G_f$
Interior: a large hydrophobe to a small hydrophobe	0	0	+[a]	0	0	++	0	0	+
Interior: a small hydrophobe to a large hydrophobe	0	0	+++[b]	0	0	– –	0	0	?
Interior: a charged residue to a hydrophobe of similar size and shape	0	0	++[c]	0	+	– –	0	0	+
Interior: a hydrophobe to a charged residue of similar size and shape	0	0	+[d]	0	– –	++	0	0	+++
Interior: a salt bridge to two equal size and same shape hydrophobes	0	0	++[e]	0	+++	+++	0	0	– –
Surface: a large hydrophobe to a small hydrophobe	0	++	0	0	0	++	0	0	0
Surface: a small hydrophobe to a large hydrophobe	0	– –	0	0	0	– –	0	0	0
Surface: a charged residue to a hydrophobe of similar size and shape	+	–	0	0	+	–	0	0	0
Surface: a hydrophobe to an equal size and same shape charge	–	+	0	0	–	+	0	0	0
At an α helix: a leucine to an isoleucine	0	0	+[f]	0	0	0	0	0	+
Surface: an alanine to a glycine	0	+	0	+[g]	0	+	0	++[h]	+
Surface: a salt bridge to two equal size and same shape hydrophobes	+++	– – –	+	0	+++	– – –	0	0	+

+, indicates an increase; –, indicates a decrease.
[a]Creates a cavity.
[b]May cause vdW clashes.
[c]The electrostatic environment of the charge partner becomes unfavourable.
[d]The charge may end up at a like-charge environment.
[e]Loses the coulombic interaction of the salt bridge.
[f]Ile is a helix breaker and can increase the internal energy of the protein chain.
[g]Glycine can increase the conformational entropy of some local loop.
[h]Glycine can increase the conformational entropy of the unfolded chain.

into combinations of enthalpies and entropies (eqn [5]).

$$\begin{aligned}\Delta\Delta G_f &= (\Delta G_f)_{mut} - (\Delta G_f)_{wt} \\ &= \Delta H_s^f - T\Delta S_s^f + \Delta H_c^f - T\Delta S_c^f - \Delta H_s^u \\ &\quad + T\Delta S_s^u - \Delta H_c^u + T\Delta S_c^u \end{aligned} \quad [5]$$

This summary assumes that the wild-type protein is optimal, i.e. that the core packing is perfect and there is no strain on the chain. It further considers the unfolded state to be largely extended and fully solvated and those immediately adjacent residues along the chain to have the same interaction energy on average in folded and denatured states. These are reasonable first approximations and help to provide a conceptual framework that ties together disparate results; however, there is some evidence that the unfolded state does have some local structure.

Use of Amino Acid Substitutions to Test Protein Stability Predictions

The change of folding free energy $\Delta\Delta G_f$ due to a mutation can be calculated using free energy perturbation, in which a protein in some definite conformation is placed in the centre of a periodic box of water molecules and the dynamics of the system (the protein and all water molecules) is followed by solving the equations of motion. The free energy change $(G^f)_{mut} - (G^f)_{wt}$ of mutating a residue reversibly (using a large number of very small steps to ensure equilibration at each step) in the folded state is calculated by integrating the enthalpy and entropy along the mutation path. The same calculation is carried out for the unfolded state, to obtain $(G^u)_{mut} - (G^u)_{wt}$. The difference between the folded and unfolded state gives us $\Delta\Delta G_f$.

The major difficulty involved in free energy perturbation is the accurate estimation of solvation entropies, which requires averaging over water conformations for a long period of time. Accuracy is compromised by energy fluctuations that are larger than the average value and the results. An alternative to estimating solvation entropy relies on an assumed linear relation with solvent-exposed hydrophobic surface areas. The exact values can be calibrated using transfer data of amino acids from hydrophobic solvent to water. This procedure underlies a number of empirical methods that use surface-dependent estimates for the solvation entropy and solve the Poisson equation, or use distance-dependent coulombic energies to estimate the solvation enthalpies. Conformational enthalpies and entropies are well estimated using molecular mechanics potentials. Nevertheless, difficulties remain in calculating $\Delta\Delta G_f$ accurately.

Double Mutant Cycles as a Test of Additivity

If two mutations are spatially well separated, their effects on $\Delta\Delta G_f$ are generally additive: this means that the $\Delta\Delta G_f$ of the double mutant can be approximated by the sum of the $\Delta\Delta G_f$ of each single mutant. The approximation is most accurate when all of the wild-type and mutant residues are noncharged. However, charged residues can have long-range interaction with one another and lead to nonadditivity.

The double mutant cycle (**Figure 5**) is a useful tool for dissecting the components of $\Delta\Delta G_f$. Consider an interior salt bridge of Asp^- and Lys^+. We can design mutants with each salt bridge partner mutated to a hydrophobic residue with similar shape and size (e.g. Asp to Leu and Lys to Met), or with both mutated. Folding free energy

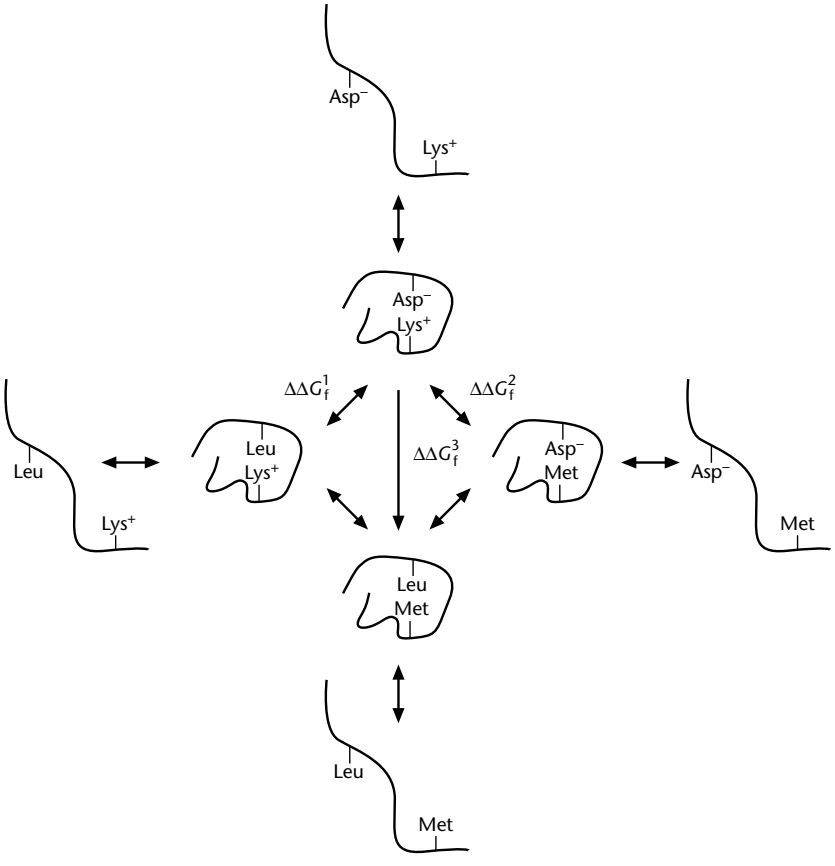

Figure 5 Differential free energy changes in double mutant cycles.

differences between states of such a mutation cycle mainly reflect solvation components of the folding free energy, if we assume that conformation enthalpy and entropy are not affected by such equal-volumic substitutions. Generally speaking, $\Delta\Delta G_f^1$ and $\Delta\Delta G_f^2$ of the single mutants can be both highly positive, reflecting the unfavourable state of unpaired charged residues in the interior of a protein. However, often $\Delta\Delta G_f^3$ of the double mutant can be negative, which indicates that two hydrophobic residues are more favourable than a salt bridge. This might seem counterintuitive, since the coulombic energy between two charges at $\approx 3\,\text{Å}$ (the average distance between two salt bridge partners) can, with a dielectric constant of 2, be as large as $210\,\text{kJ mol}^{-1}$. In fact, the penalty of desolvating the two charges can be larger than the coulombic energy and thus the net contribution of a salt bridge to protein stability can be unfavourable.

Summary

Mutations can now be introduced into proteins as we wish, specifically or randomly, and at multiple positions simultaneously. The array of new sequences thus generated can be used to improve our understanding of protein stability and activity and the relation between them. Such experimental results, coupled with new computational methods have been especially fruitful in developing a predictive understanding of structure.

References

Gassner NC, Baase WA and Matthews BW (1996) A test of the "jigsaw puzzle" model for protein folding by multiple methionine substitutions within the core of T4 lysozyme. *Proceedings of the National Academy of Sciences of the USA* **93**: 12155–12158.

Sauer RT (1996) Protein folding from a combinatorial perspective. *Folding and Design* **1**: 27–30.

Vajda S, Weng Z, Rosenfeld R and DeLisi C (1994) Effect of conformational flexibility and solvation on receptor ligand binding free energies. *Biochemistry* **33**: 13977.

Weng Z, Vajda S and DeLisi C (1996) Rigid body docking with semi empirical free energy functions. *Protein Science* **5**: 614–626.

Further Reading

Branden C and Tooze J (1999) *Introduction to Protein Structure*. New York: Garland Publishing.