Bioinformatics for Geneticists. Edited by Michael R. Barnes and Ian C. Gray

ISBNs: 0-470-84393-4 (HB); 0-470-84394-2 (PB)

CHAPTER 14

Amino Acid Properties and Consequences of Substitutions

MATTHEW J. BETTS[1] and ROBERT B. RUSSELL[2]

[1]*Bioinformatics*
deCODE genetics, Sturlugötu 8
101 Reykjavık, Iceland
[2]*Structural & Computational Biology Programme*
EMBL, Meyerhofstrasse 1
69117 Heidelberg, Germany

14.1 INTRODUCTION

Since the earliest protein sequences and structures were determined, it has been clear that the positioning and properties of amino acids are key to understanding many biological processes. For example, the first protein structure, haemoglobin provided a molecular explanation for the genetic disease sickle cell anaemia. A single nucleotide mutation leads to a substitution of glutamate in normal individuals with valine in those who suffer the disease. The substitution leads to a lower solubility of the deoxygenated form of haemoglobin and it is thought that this causes the molecules to form long fibres within blood cells which leads to the unusual sickle-shaped cells that give the disease its name.

Haemoglobin is just one of many examples now known where single mutations can have drastic consequences for protein structure, function and associated phenotype. The current availability of thousands or even millions of DNA and protein sequences means that we now have knowledge of many mutations, either naturally occurring or synthetic. Mutations can occur within one species, or between species at a wide variety

of evolutionary distances. Whether mutations cause diseases or have subtle or drastic effects on protein function is often unknown.

The aim of this chapter is to give some guidance as to how to interpret mutations that occur within genes that encode for proteins. Both authors of this chapter have been approached previously by geneticists who want help interpreting mutations through the use of protein sequence and structure information. This chapter is an attempt to summarize our thought processes when giving such help. Specifically, we discuss the nature of mutations and the properties of amino acids in a variety of different protein contexts. The hope is that this discussion will help in anticipating or interpreting the effect that a particular amino acid change will have on protein structure and function. We will first highlight features of proteins that are relevant to considering mutations: cellular environments, three-dimensional structure and evolution. Then we will discuss classifications of the amino acids based on evolutionary, chemical or structural principles, and the role for amino acids of different classes in protein structure and function in different contexts. Last, we will review several studies of mutations, including naturally-occurring variations, SNPs, site-directed mutations, mutations that allow adaptive evolution and post-translational modification.

14.2 PROTEIN FEATURES RELEVANT TO AMINO ACID BEHAVIOUR

It is beyond the scope of this chapter to discuss the basic principles of proteins, since this can be gleaned from any introductory biochemistry text-book. However, a number of general principles of proteins are important to place any mutation in the correct context.

14.2.1 Protein Environments

A feature of key importance is cellular location. Different parts of cells can have very different chemical environments with the consequence that many amino acids behave differently. The biggest difference is between *soluble* proteins and *membrane* proteins. Whereas soluble proteins tend to be surrounded by water molecules, membrane proteins are surrounded by lipids. Roughly speaking this means that these two classes behave in an 'inside-out' fashion relative to each other. Soluble proteins tend to have polar or hydrophilic residues on their surfaces, whereas membrane proteins tend to have hydrophobic residues on the surface that interact with the membrane.

Soluble proteins also come in several flavours. The biggest difference is between those that are *extracellular* and those that are *cytosolic* (or *intracellular*). The cytosol is quite different from the more aqueous environment outside the cell; the density of proteins and other molecules effects the behaviour of some amino acids quite drastically, the foremost among these being cysteine. Outside the cell, cysteines in proximity to one another can be *oxidized* to form disulphide bonds, sulphur–sulphur covalent linkages that are important for protein folding and stability. However, the reducing environment inside the cell makes the formation of these bonds very difficult; in fact they are so rare as to warrant special attention.

Cells also contain numerous compartments, the organelles, which can also have slightly different environments from each other. Proteins in the nucleus often interact with DNA, meaning they contain different preferences for amino acids on their surfaces (e.g. positive amino acids or those containing amides most suitable for interacting with the negatively charged phosphate backbone). Some organelles such as mitochondria or chloroplasts are

quite similar to the cytosol, while others, such as lysosomes or Golgi apparati are more akin to the extracellular environment. It is important to consider the likely cellular location of any protein before considering the consequences of amino acid substitutions.

A detailed hierarchical description of cellular location is one of the three main branches of the classification provided by the Gene Ontology Consortium (Ashburner *et al*., 2000), the others being 'molecular function' and 'biological process'. The widespread adoption of this vocabulary by sequence databases and others should enable more sophisticated investigation of the factors governing the various roles of proteins.

14.2.2 Protein Structure

Proteins themselves also contain different microenvironments. For soluble proteins, the surface lies at the interface with water and thus tends to contain more polar or charged amino acids than one finds in the core of the protein, which is more likely to comprise hydrophobic amino acids. Proteins also contain regions that are directly involved in protein function, such as active sites or binding sites, in addition to regions that are less critical to the protein function and where mutations are likely to have fewer consequences. We will discuss many specific roles for particular amino acids in protein structures in the sections below, but it is important to remember that the context of any amino acid can vary greatly depending on its location in the protein structure.

14.2.3 Protein Evolution

Proteins are nearly always members of homologous families. Knowledge about the family a protein belongs in will generally give insights into the possible function, but several things should be considered. Two processes can give rise to homologous protein families: *speciation* or *duplication*. Proteins related by speciation only are referred to as *orthologues*, and as the name suggests, these proteins have the same function in different species. Proteins related by duplications are referred to as *paralogues*. Successive rounds of speciation and intra-genomic duplication can lead to confusing situations where it becomes difficult to say whether paralogy or orthology applies.

To be maintained in a genome over time, paralogous proteins are likely to evolve different functions (or have a dominant negative phenotype and so resist decay by point mutation (Gibson and Spring, 1998)). Differences in function can range from subtle differences in substrate (e.g. malate versus lactate dehydrogenases), to only weak similarities in molecular function (e.g. hydrolases) to complete differences in cellular location and function (e.g. an intracellular signalling domain homologous to a secreted growth factor (Schoorlemmer and Goldfarb, 2001)). At the other extreme, the molecular function may be identical, but the cellular function may be altered, as in the case of enzymes with differing tissue specificities.

Similarity in molecular function generally correlates with sequence identity. Mouse and human proteins with sequence identities in excess of 85% are likely to be orthologues, provided there are no other proteins with higher sequence identity in either organism. Orthology between more distantly related species (e.g. human and yeast) is harder to assess, since the evolutionary distance between organisms can make it virtually impossible to distinguish orthologues from paralogues using simple measures of sequence similarity. An operational definition of orthology can sometimes be used, for example if the two proteins are each other's best match in their respective genomes. However there is no substitute for constructing a phylogenetic tree of the protein family, to identify

which sequences are related by speciation events. Assignment of orthology and paralogy is perhaps the best way of determining likely equivalences of function. Unfortunately, complete genomes are unavailable for most organisms. Some rough rules of thumb can be used: function is often conserved down to 40% protein sequence identity, with the broad functional class being conserved to 25% identity (Wilson *et al*., 2000).

When considering a mutation, it is important to consider how conserved the position is within other homologous proteins. Conservation across all homologues (paralogues and orthologues) should be considered carefully. These amino acids are likely to play key structural roles or a role in a common functional theme (i.e. catalytic mechanism). Other amino acids may play key roles only in the particular orthologous group (i.e. they may confer specificity to a substrate), thus meaning they vary when considering all homologues.

14.2.4 Protein Function

Protein function is key to any understanding of the consequences of amino acid substitution. Enzymes, such as trypsin (Figure 14.1), tend to have highly conserved active sites involving a handful of polar residues. In contrast, proteins that function primarily only to interact with other proteins, such as fibroblast growth factors (Figure 14.2), interact over a large surface, with virtually any amino acid being important in mediating the interaction (Plotnikov *et al*., 1999). In other cases, multiple functions make the situation even more confusing, for example a protein kinase (Hanks *et al*., 1988) can both catalyse a phosphorylation event and bind specifically to another protein, such as cyclin (Jeffrey *et al*., 1995).

It is not possible to discuss all of the possible functional themes here, but we emphasize that functional information, if known, should be considered whenever studying the effects of substitution.

14.2.5 Post-translational Modification

Although there are only 20 possible types of amino acid that can be incorporated into a protein sequence upon translation of DNA, there are many more variations that can occur

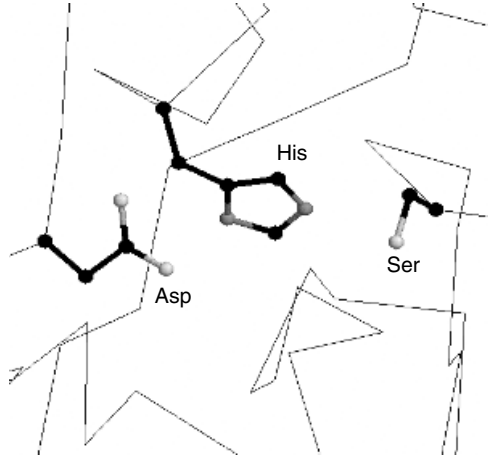

Figure 14.1 RasMol (Sayle and Milner-White, 1995) figure showing the catalytic Asp-His-Ser triad in trypsin (PDB code 1mct; Berman *et al*., 2000).

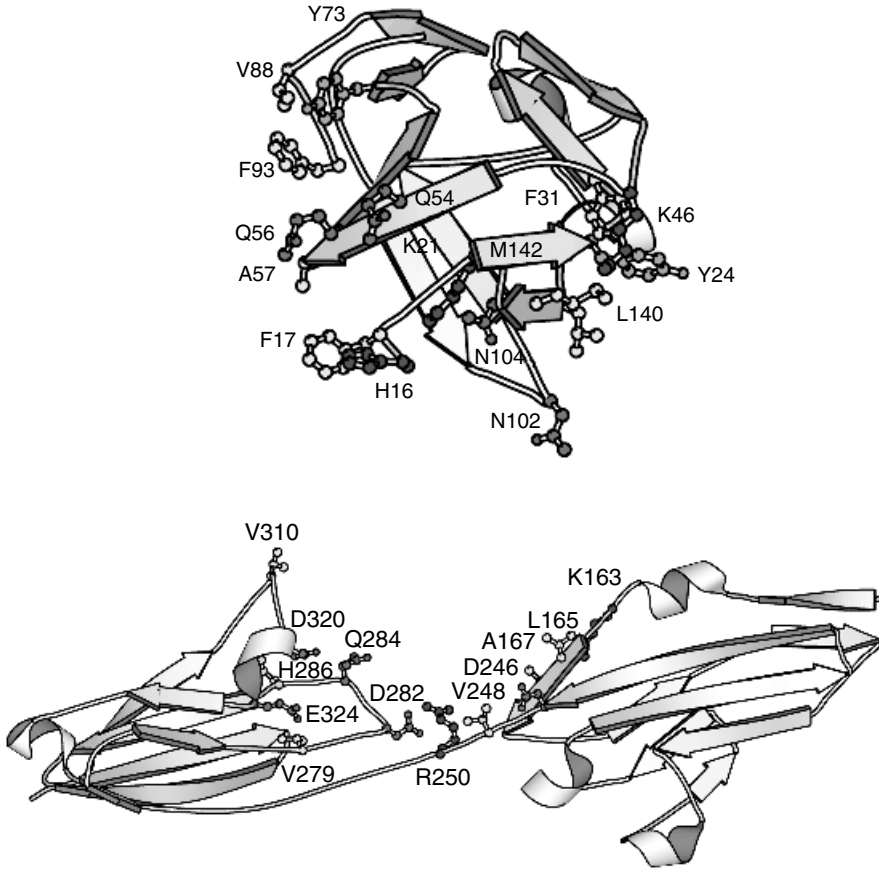

Figure 14.2 Molscript (Kraulis, 1991) figure showing fibroblast growth factor interaction with its receptor (code 1cvs; Plotnikov *et al*., 1999). Residues at the interface are labelled. The two molecules have been pulled apart for clarity.

through subsequent modification. In addition, the gene-specified protein sequence can be shortened by proteolysis, or lengthened by addition of amino acids at either terminus.

Two common modifications, phosphorylation and glycosylation, are discussed in the context of the amino acids where they most often occur (tyrosine, serine, threonine and asparagine; see below). We direct the reader to the review by Krishna for more information on many other known types and specific examples (Krishna and Wold, 1993). The main conclusion is that modifications are highly specific, with specificity provided by primary, secondary and tertiary protein structure, although with detailed mechanisms being obscure. The biological function of the modified proteins is also summarized, from the reversible phosphorylation of serine, threonine and tyrosine residues that occurs in signalling through to the formation of disulphide bridges and other cross-links that stabilize tertiary structure, and on to the covalent attachment of lipids that allows anchorage to cell membranes. More detail on biological effects is given by Parekh and Rohlff (1997), especially where it concerns possible therapeutic applications. Many diseases arise by

abnormalities in post-translational modification, and these are not necessarily apparent from genetic information alone.

14.3 AMINO ACID CLASSIFICATIONS

Humans have a natural tendency to classify, as it makes the world around us easier to understand. As amino acids often share common properties, several classifications have been proposed. This is useful, but a little bit dangerous if over-interpreted. Always remember that, for the reasons discussed above, it is very difficult to put all amino acids of the same type into an invariant group. A substitution in one context can be disastrous in another. For example, a cysteine involved in a disulphide bond would not be expected to be mutatable to any other amino acid (i.e. it is in a group on its own), one involved in binding to zinc could likely be substituted by histidine (group of two) and one buried in an intracellular protein core could probably mutate to any other hydrophobic amino acid (a group of 10 or more). We will discuss other examples below.

14.3.1 Mutation Matrices

One means of classifiying amino acids is a mutation matrix (or substitution or exchange matrix). This is a set of numbers that describe the propensities of exchanging one amino acid for another (for a comprehensive review and explanation see Durbin *et al*., 1998). These are derived from large sets of aligned sequences by counting the number of times that a particular substitution occurs and comparing this to what would be expected by chance. High values indicate that a substitution is seen often in nature and so is favourable, and vice versa. The values in the matrix are usually calculated using some model of evolutionary time, to account for the fact that different pairs of sequences are at different evolutionary distances. Probably the best known matrices are the Point Accepted Mutation (PAM) matrices of Dayhoff *et al*. (Dayhoff *et al*., 1978) and BLOSUM matrices (Henikoff and Henikoff, 1992).

Mutation matrices are very useful as rough guides for how good or bad a particular change will be. Another useful feature is that they can be calculated for different datasets to account for some of the protein features that effect amino acid properties, such as cellular locations (Jones *et al*., 1994) or different evolutionary distances (e.g. orthologues or paralogues; Henikoff and Henikoff, 1992). Several mutation matrices are reproduced in Appendix II.

14.3.2 Classification by Physical, Chemical and Structural Properties

Although mutation matrices are very useful for protein sequence alignments, especially in the absence of known three-dimensional structures, they do not precisely describe the likelihood and effects of particular substitutions at particular sites in the sequence. Position-specific substitution matrices can be generated for the family of interest, such as the profile-HMM models generated by HMMER (Eddy, 1998) and provided by Pfam (Bateman *et al*., 2000), and those generated by PSI-BLAST (Altschul *et al*., 1997). However, these are automatic methods suited to database searching and identification of new members of a family, and as such do not really give any qualitative information about the chemistry involved at particular sites.

Taylor presented a classification that explains mutation data through correlation with the physical, chemical and structural properties of amino acids (Taylor, 1986). The major

factor is the size of the side chain, closely followed by its hydrophobicity. Effects of differerent amino acids on protein structure can account for mutation data when these physico-chemical properties do not. For example, hydrophobicity and size differ widely between glycine, proline, aspartic acid and glutamic acid. However, they are still closely related in mutation matrices because they prefer sharply turning regions on the surface of the protein; the phi and psi bonds of glycine are unconstrained by any side chain, proline forces a sharp turn because its side chain is bonded to the backbone nitrogen as well as to carbon, and aspartate and glutamate prefer to expose their charged side chains to solvent.

The Taylor classification is normally displayed as a Venn diagram (Figure 14.3). The amino acids were positioned on this by multidimensional scaling of Dayhoff's mutation matrix, and then grouped by common physico-chemical properties. Size is subcategorized into small and tiny (with large included by implication). Affinity for water is described by several sets: polar and hydrophobic, which overlap, and charged, which is divided into positive and negative. Sets of aromatic and aliphatic amino acids are also marked. These properties were enough to distinguish between most amino acids. However, properties such as hydrogen-bonding ability and the previously mentioned propensity for sharply turning regions are not described well. Although these factors are less important on average, and would confuse the effects of more important properties if included on the diagram, the dangers of relying on simple classifications are apparent. This can be overcome somewhat by listing all amino acids which belong to each subset (defined as an intersection or union of the sets) in the diagram, for example 'small and non-polar', and including extra subsets to describe important additional properties. These subsets can be used to give qualitative descriptions of each position in a multiple alignment, by associating the positions with the smallest subset that includes all the amino acids found at that position.

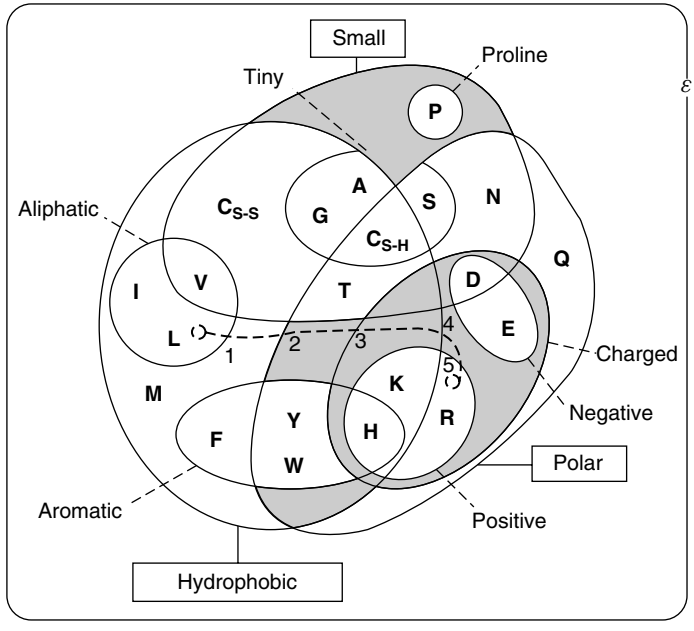

Figure 14.3 Venn diagram illustrating the properties of amino acids.

This may suggest alternative amino acids that could be engineered into the protein at each position.

14.4 PROPERTIES OF THE AMINO ACIDS

The sections that follow will first consider several major properties that are often used to group amino acids together. Note that amino acids can be in more than one group, and that sometimes properties as different as 'hydrophobic' and 'hydrophilic' can be applied to the same amino acids.

14.4.1 Hydrophobic Amino Acids

Probably the most common broad division of amino acids is into those that prefer to be in an aqueous environment (hydrophilic) and those that do not (hydrophobic). The latter can be divided according to whether they have *aliphatic* or *aromatic* side chains.

14.4.1.1 Aliphatic Side Chains

Strictly speaking aliphatic means that the side chain contains only hydrogen and carbon atoms. By this strict definition, the amino acids with aliphatic side chains are *alanine, isoleucine, leucine, proline* and *valine*. Alanine's side chain, being very short, means that it is not particularly hydrophobic and proline has an unusual geometry that gives it special roles in proteins as we shall discuss below. Although it also contains a sulphur atom, it is often convenient to consider *methionine* in the same category as isoleucine, leucine and valine. The unifying theme is that they contain largely non-reactive and flexible side chains that are ideally suited for packing in the protein interior.

Aliphatic side chains are very non-reactive, and are thus rarely involved directly in protein function, although they can play a role in substrate recognition. In particular, hydrophobic amino acids can be involved in binding/recognition of hydrophobic ligands such as lipids.

Several other amino acids also contain aliphatic regions. For example, arginine, lysine, glutamate and glutamine are *amphipathic*, meaning that they contain hydrophobic and polar areas. All contain two or more aliphatic carbons that connect the protein backbone to the non-aliphatic portion of the side chain. In some instances it is possible for such amino acids to play a dual role, with part of the side chain being buried in the protein and another being exposed to water.

14.4.1.2 Aromatic Side Chains

A side chain is aromatic when it contains an aromatic ring system. The strict definition has to do with the number of electrons contained within the ring. Generally, aromatic ring systems are planar and electrons are shared over the whole ring structure. *Phenylalanine and tryptophan* have very hydrophobic aromatic side chains, whereas *tyrosine* and *histidine* are less so. The latter two can often be found in positions that are somewhere between buried and exposed. The hydrophobic aromatic amino acids can sometimes substitute for aliphatic residues of a similar size, for example phenylalanine to leucine, but not tryptophan to valine.

Aromatic residues have also been proposed to participate in 'stacking' interactions (Hunter *et al*., 1991) (Figure 14.4). Here, numerous aromatic rings are thought to stack

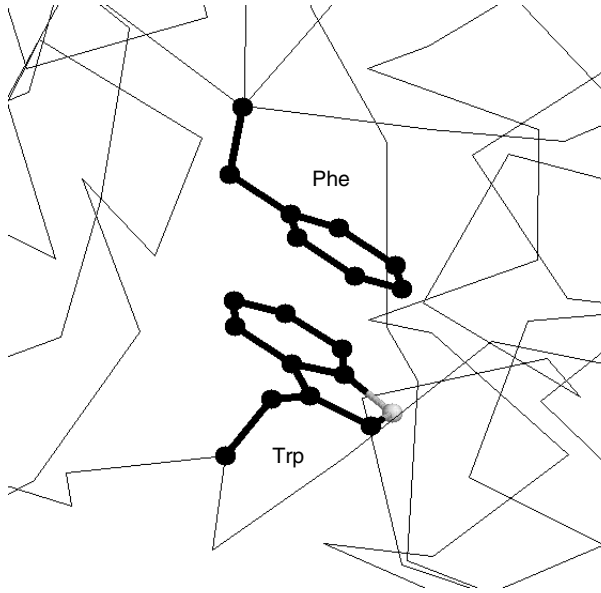

Figure 14.4 Example of aromatic stacking.

on top of each other such that their PI electron clouds are aligned. They can also play a role in binding to specific amino acids, such as proline. SH3 and WW domains, for example, use these residues to bind to their polyproline-containing interaction partners (Macias *et al*., 2002). Owing to its unique chemical nature, histidine is frequently found in protein active sites as we shall see below.

14.4.2 Polar Amino Acids

Polar amino acids prefer to be surrounded by water. Those that are buried within the protein usually participate in hydrogen bonds with other side chains or the protein main-chain that essentially replace the water. Some of these carry a charge at typical biological pHs: *aspartate* and *glutamate* are negatively charged; *lysine* and *arginine* are positively charged. Other polar amino acids, *histidine, asparagine, glutamine, serine, threonine* and *tyrosine*, are neutral.

14.4.3 Small Amino Acids

The amino acids *alanine*, *cysteine*, *glycine*, *proline*, *serine* and *threonine* are often grouped together for the simple reason that they are all small in size. In some protein structural contexts, substitution of a small side chain for a large one can be disastrous.

14.5 AMINO ACID QUICK REFERENCE

In the sections that follow we discuss each amino acid in turn. For each we will briefly discuss general preferences for substitutions and important specific details regarding their

possible structure and functional roles. More information is found on the WWW site that accompanies this chapter (www.russell.embl-heidelberg.de/aas). This website also features amino acid substitution matrices for transmembrane, extracellular and intracellular proteins. These can be used to numerically score an amino acid substitution, where unpreferred mutations are given negative scores, preferred substitutions are given positive scores and neutral substitutions are given zero scores.

14.5.1 Alanine (Ala, A)

14.5.1.1 Substitutions

Alanine can be substituted by other small amino acids.

14.5.1.2 Structure

Alanine is probably the dullest amino acid. It is not particularly hydrophobic and is non-polar. However, it contains a normal Cβ carbon, meaning that it is generally as hindered as other amino acids with respect to the conformations that the backbone can adopt. For this reason, it is not surprising to see alanine present in just about all non-critical protein contexts.

14.5.1.3 Function

The alanine side chain is very non-reactive, and is thus rarely directly involved in protein function, but it can play a role in substrate recognition or specificity, particularly in interactions with other non-reactive atoms such as carbon.

14.5.2 Isoleucine (Ile, I)

14.5.2.1 Substitutions

Isoleucine can be substituted by other hydrophobic, particularly aliphatic, amino acids.

14.5.2.2 Structure

Being hydrophobic, isoleucine prefers to be buried in protein hydrophobic cores. However, isoleucine has an additional property that is frequently overlooked. Like valine and threonine it is Cβ branched. Whereas most amino acids contain only one non-hydrogen substituent attached to their Cβ carbon, these three amino acids contain two. This means that there is a lot more bulkiness near to the protein backbone and this means that these amino acids are more restricted in the conformations the main chain can adopt. Perhaps the most pronounced effect of this is that it is more difficult for these amino acids to adopt an α-helical conformation, although it is easy and even preferred for them to lie within β-sheets.

14.5.2.3 Function

The isoleucine side chain is very non-reactive and is thus rarely directly involved in protein functions like catalysis, although it can play a role in substrate recognition. In particular, hydrophobic amino acids can be involved in binding/recognition of hydrophobic ligands such as lipids.

14.5.3 Leucine (Leu, L)

14.5.3.1 Substitutions

See Isoleucine.

14.5.3.2 Structure

Being hydrophobic, leucine prefers to be buried in protein hydrophobic cores. It also shows a preference for being within alpha helices more so than in beta strands.

14.5.3.3 Function

See Isoleucine.

14.5.4 Valine (Val, V)

14.5.4.1 Substitutions

See Isoleucine.

14.5.4.2 Structure

Being hydrophobic, valine prefers to be buried in protein hydrophobic cores. However, valine is also Cβ branched (see Isoleucine).

14.5.4.3 Function

See Isoleucine.

14.5.5 Methionine (Met, M)

14.5.5.1 Substitutions

See Isoleucine.

14.5.5.2 Structure

See Isoleucine.

14.5.5.3 Function

The methionine side chain is fairly non-reactive, and is thus rarely directly involved in protein function. Like other hydrophobic amino acids, it can play a role in binding/recognition of hydrophobic ligands such as lipids. However, unlike the proper aliphatic amino acids, methionine contains a sulphur atom, that can be involved in binding to atoms such as metals. However, whereas the sulphur atom in cysteine is connected to a hydrogen atom making it quite reactive, methionine's sulphur is connected to a methyl group. This means that the roles that methionine can play in protein function are much more limited.

14.5.6 Phenylalanine (Phe, F)

14.5.6.1 Substitutions

Phenylalanine can be substituted with other aromatic or hydrophobic amino acids. It particularly prefers to exchange with tyrosine, which differs only in that it contains an hydroxyl group in place of the ortho hydrogen on the benzene ring.

14.5.6.2 Structure

Phenylalanine prefers to be buried in protein hydrophobic cores. The aromatic side chain can also mean that phenylalanine is involved in stacking (Figure 14.4) interactions with other aromatic side chains.

14.5.6.3 Function

The phenylalanine side chain is fairly non-reactive, and is thus rarely directly involved in protein function, although it can play a role in substrate recognition (see Isoleucine). Aromatic residues can also be involved in interactions with non-protein ligands that themselves contain aromatic groups via stacking interactions (see above). They are also common in polyproline binding sites, for example in SH3 and WW domains (Macias *et al*., 2002).

14.5.7. Tryptophan (Trp, W)

14.5.7.1 Substitutions

Trytophan can be replaced by other aromatic residues, but it is unique in terms of chemistry and size, meaning that often replacement by anything could be disastrous.

14.5.7.2 Structure

See Phenylalanine.

14.5.7.3 Function

As it contains a non-carbon atom (nitrogen) in the aromatic ring system, tryptophan is more reactive than phenylalanine although it is less reactive than tyrosine. Tryptophan can play a role in binding to non-protein atoms, but such instances are rare. See also Phenylalanine.

14.5.8 Tyrosine (Tyr, Y)

14.5.8.1 Substitutions

Tyrosine can be substituted by other aromatic amino acids. See Phenylalanine.

14.5.8.2 Structure

Being partially hydrophobic, tyrosine prefers to be buried in protein hydrophobic cores. The aromatic side chain can also mean that tyrosine is involved in stacking interactions with other aromatic side chains.

14.5.8.3 Function

Unlike the very similar phenylalanine, tyrosine contains a reactive hydroxyl group, thus making it much more likely to be involved in interactions with non-carbon atoms. See also Phenylalanine.

A common role for tyrosines (and serines and threonines) within intracellular proteins is phosphorylation. Protein kinases frequently attach phosphates to these three residues as part of a signal transduction process. Note that in this context, tyrosine will rarely substitute for serine or threonine since the enzymes that catalyse the reactions (i.e. the protein kinases) are highly specific (i.e. tyrosine kinases generally do not work on serines/threonines and vice versa (Hanks *et al*., 1988)).

14.5.9 Histidine (His, H)

14.5.9.1 Substitutions

Histidine is generally considered to be a polar amino acid, however it is unique with regard to its chemical properties, which means that it does not substitute particularly well with any other amino acid.

14.5.9.2 Structure

Histidine has a pK_a near to that of physiological pH, meaning that it is relatively easy to move protons on and off of the side chain (i.e. changing the side chain from neutral to positive charge). This flexibility has two effects. The first is ambiguity about whether it prefers to be buried in the protein core or exposed to solvent. The second is that it is an ideal residue for protein functional centres (discussed below). It is false to presume that histidine is always protonated at typical pHs. The side chain has a pK_a of approximately 6.5, which means that only about 10% of molecules will be protonated. The precise pK_a depends on local environment.

14.5.9.3 Function

Histidines are the most common amino acids in protein active or binding sites. They are very common in metal binding sites (e.g. zinc), often acting together with cysteines or other amino acids (Figure 14.5; Wolfe *et al*., 2001). In this context, it is common to see histidine replaced by cysteine.

The ease with which protons can be transferred on and off of histidines makes them ideal for charge relay systems such as those found within catalytic triads and in many cysteine and serine proteases (Figure 14.1). In this context, it is rare to see histidine exchanged for any amino acid at all.

14.5.10 Arginine (Arg, R)

14.5.10.1 Substitutions

Arginine is a positively-charged, polar amino acid. It thus most prefers to substitute for the other positively-charged amino acid, lysine, although in some circumstances it will also tolerate a change to other polar amino acids. Note that a change from arginine to lysine is not always neutral. In certain structural or functional contexts, such a mutation can be devastating to function (see below).

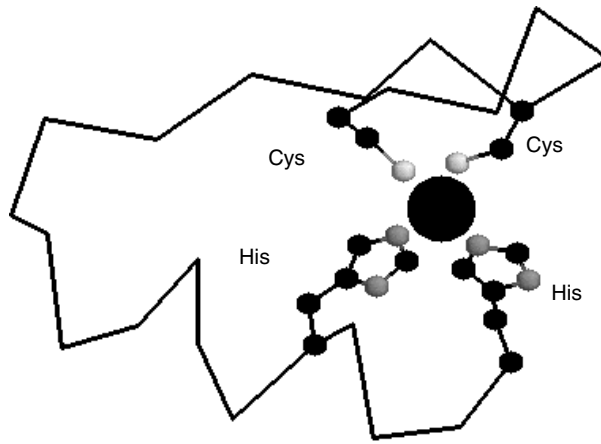

Figure 14.5 Example of a metal binding site coordinated by cysteine and histidine residues (code 1g2f; Wolfe *et al*., 2001).

14.5.10.2 Structure

Arginine generally prefers to be on the surface of the protein, but its amphipathic nature can mean that part of the side chain is buried. Arginines are also frequently involved in salt-bridges where they pair with a negatively charged aspartate or glutamate to create stabilizing hydrogen bonds that can be important for protein stability (Figure 14.6).

14.5.10.3 Function

Arginines are quite frequent in protein active or binding sites. The positive charge means that they can interact with negatively-charged non-protein atoms (e.g. anions or carboxylate groups). Arginine contains a complex guanidinium group on its side chain that has a geometry and charge distribution that is ideal for binding negatively-charged groups on phosphates (it is able to form multiple hydrogen bonds). A good example can be found in the src homology 2 (SH2) domains (Figure 14.7; Waksman *et al*., 1992). The two arginines shown in the figure make multiple hydrogen bonds with the phosphate. In this context arginine is not easily replaced by lysine. Although lysine can interact with phosphates, it contains only a single amino group, meaning it is more limited in the number of hydrogen bonds it can form. A change from arginine to lysine in some contexts can thus be disastrous (Copley and Barton, 1994).

14.5.11 Lysine (Lys, K)

14.5.11.1 Substitutions

Lysine can be substituted by arginine or other polar amino acids.

14.5.11.2 Structure

Lysine frequently plays an important role in structure. First, it can be considered to be somewhat amphipathic as the part of the side chain nearest to the backbone is long, carbon-containing and hydrophobic, whereas the end of the side chain is positively charged. For

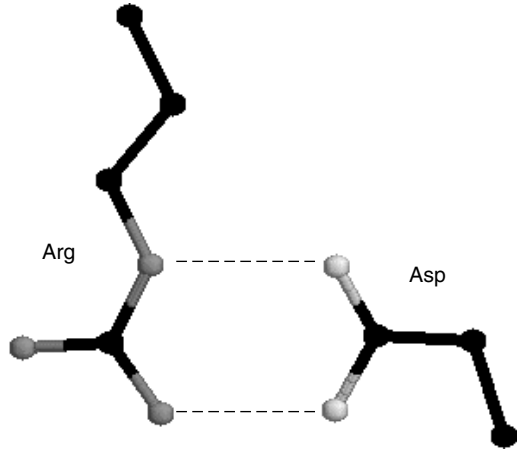

Figure 14.6 Example of a salt-bridge (code 1xel).

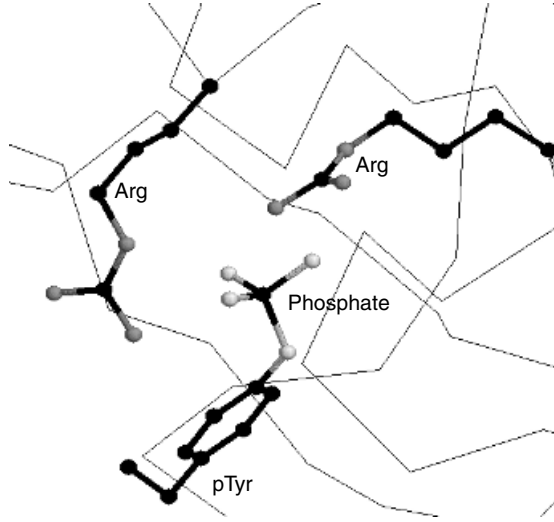

Figure 14.7 Interaction of arginine residues with phosphotyrosine in an SH2 domain (code 1sha; Waksman *et al*., 1992).

this reason, one can find lysines where part of the side chain is buried and only the charged portion is on the outside of the protein. However, this is by no means always the case and generally lysines prefer to be on the outside of proteins. Lysines are also frequently involved in salt-bridges (see Arginine).

14.5.11.3 Function

Lysines are quite frequent in protein active or binding sites. Lysine contains a positively-charged amino group on its side chain that is sometimes involved in forming hydrogen bonds with negatively-charged non-protein atoms (e.g. anions or carboxylate groups).

14.5.12 Aspartate (Asp, D)

14.5.12.1 Substitutions

Aspartate can be substituted by glutamate or other polar amino acids, particularly asparagine, which differs only in that it contains an amino group in place of one of the oxygens found in aspartate (and thus also lacks a negative charge).

14.5.12.2 Structure

Being charged and polar, aspartates generally prefer to be on the surface of proteins, exposed to an aqueous environment. Aspartates (and glutamates) are frequently involved in salt-bridges (see Arginine).

14.5.12.3 Function

Aspartates are quite frequently involved in protein active or binding sites. The negative charge means that they can interact with positively-charged non-protein atoms, such as

cations like zinc. Aspartate has a shorter side chain than the very similar glutamate meaning that is slightly more rigid within protein structures. This gives it a slightly stronger preference to be involved in protein active sites. Probably the most famous example of aspartate being involved in an active site is found within serine proteases such as trypsin, where it functions in the classical Asp-His-Ser catalytic triad (Figure 14.1). In this context, it is quite rare to see aspartate exchange for glutamate, although it is possible for glutamate to play a similar role.

14.5.13 Glutamate (Glu, E)

14.5.13.1 Substitutions

Substitution can be by aspartate or other polar amino acids, in particular glutamine, which is to glutamate what asparagine is to aspartate (see above).

14.5.13.2 Structure

See Aspartate.

14.5.13.3 Function

Glutamate, like aspartate, is quite frequently involved in protein active or binding sites. In certain cases, they can also perform a similar role to aspartate in the catalytic site of proteins such as proteases or lipases.

14.5.14 Asparagine (Asn, N)

14.5.14.1 Substitutions

Asparagine can be substituted by other polar amino acids, especially aspartate (see above).

14.5.14.2 Structure

Being polar asparagine prefers generally to be on the surface of proteins, exposed to an aqueous environment.

14.5.14.3 Function

Asparagines are quite frequently involved in protein active or binding sites. The polar side chain is good for interactions with other polar or charged atoms. Asparagine can play a similar role to aspartate in some proteins. Probably the best example is found in certain cysteine proteases, where it forms part of the Asn-His-Cys catalytic triad. In this context, it is quite rare to see asparagine exchange for glutamine.

Asparagine, when occurring in a particular motif (Asn-X-Ser/Thr) can be *N*-glycosylated (Gavel and von Heijne, 1990). Thus in this context it is impossible to substitute it with any amino acid at all.

14.5.15 Glutamine (Gln, Q)

14.5.15.1 Substitutions

Glutamine can be substituted by other polar amino acids, especially glutamate (see above).

14.5.15.2 Structure

See Asparagine.

14.5.15.3 Function

Glutamines are quite frequently involved in protein active or binding sites. The polar side chain is good for interactions with other polar or charged atoms.

14.5.16 Serine (Ser, S)

14.5.16.1 Substitutions

Serine can be substituted by other polar or small amino acids in particular threonine which differs only in that it has a methyl group in place of a hydrogen group found in serine.

14.5.16.2 Structure

Being a fairly indifferent amino acid, serine can reside both within the interior of a protein, or on the protein surface. Its small size means that it is relatively common within tight turns on the protein surface, where it is possible for the serine side chain hydroxyl oxygen to form a hydrogen bond with the protein backbone, effectively mimicking proline.

14.5.16.3 Function

Serines are quite common in protein functional centres. The hydroxyl group is fairly reactive, being able to form hydrogen bonds with a variety of polar substrates.

Perhaps the best known role for serine in protein active sites is exemplified by the classical Asp-His-Ser catalytic triad found in many hydrolases (e.g. proteases and lipases; Figure 14.1). Here, a serine, aided by a histidine and an aspartate act as a nucleophile to hydrolyse (effectively cut) other molecules. This three-dimensional 'motif' is found in many non-homologous (i.e. unrelated) proteins and is a classic example of molecular convergent evolution (Russell, 1998). In this context, it is rare for serine to exchange with threonine, but in some cases, the reactive serine can be replaced by cysteine, which can fulfil a similar role.

Intracellular serines can also be phosphorylated (see Tyrosine). Extracellular serines can also be *O*-glycosylated where a carbohydrate is attached to the side chain hydroxyl group (Gupta *et al.*, 1999).

14.5.17 Threonine (Thr, T)

14.5.17.1 Substitutions

Threonine can be substituted with other polar amino acids, particularly serine (see above).

14.5.17.2 Structure

Being a fairly indifferent amino acid, threonine can reside both within the interior of a protein or on the protein surface. Threonine is also Cβ branched (see Isoleucine).

14.5.17.3 Function

Threonines are quite common in protein functional centres. The hydroxyl group is fairly reactive, being able to form hydrogen bonds with a variety of polar substrates. Intracellular

threonines can also be phosphorylated (see Tyrosine) and in the extracellular environment they can be *O*-glycosylated (see Serine).

14.5.18 Cysteine (Cys, C)

14.5.18.1 Substitutions

In the case of cysteine there is no general preference for substitution with any other amino acid, although it can tolerate substitutions with other small amino acids. Cysteine has a role that is very dependent on cellular location, making substitution matrices dangerous to interpret (e.g. Barnes and Russell, 1999).

14.5.18.2 Structure

The role of cysteines in structure is very dependent on the cellular location of the protein in which they are contained. Within extracellular proteins, cysteines are frequently involved in disulphide bonds, where pairs of cysteines are oxidized to form a covalent bond. These bonds serve mostly to stabilize the protein structure and the structure of many extracellular proteins is almost entirely determined by the topology of multiple disulphide bonds (e.g. Figure 14.8).

The reducing environment inside cells makes the formation of disulphide bonds very unlikely. Indeed, instances of disulphide bonds in the intracellular environment are so rare that they almost always attract special attention. Disulphide bonds are also rare within the membrane, although membrane proteins may contain disulphide bonds within extracellular domains. Disulphide bonds are such that cysteines must be paired. If one half of a disulphide bond pair is lost, then the protein may not fold properly.

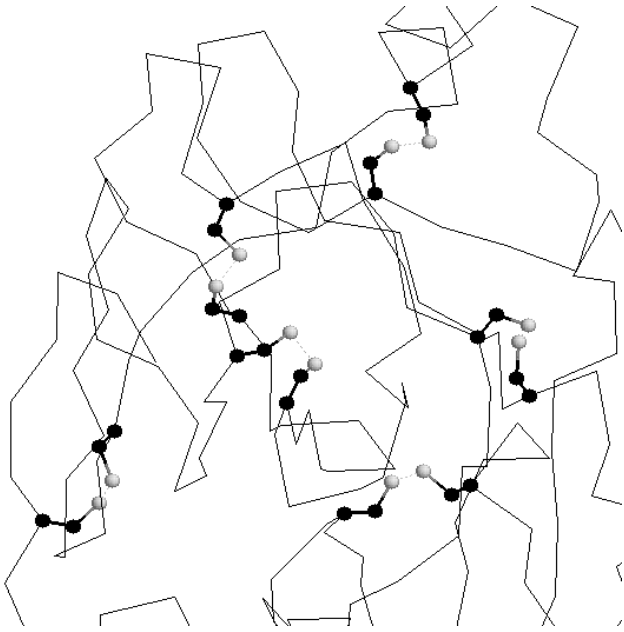

Figure 14.8 Example of a small, disulphide-rich protein (code 1tfx).

In the intracellular environment cysteines can still play a key structural role. Their sulphydryl side chain is excellent for binding to metals, such as zinc, meaning that cysteines (and other amino acids such as histidines) are very common in metal binding motifs such as zinc fingers (Figure 14.5). Outside of this context within the intracellular environment and when it is not involved in molecular function, cysteine is a neutral, small amino acid and prefers to substitute with other amino acids of the same type.

14.5.18.3 Function

Cysteines are also very common in protein active and binding sites. Binding to metals (see above) can also be important in enzymatic functions (e.g. metal proteases). Cysteine can also function as a nucleophile (i.e. the reactive centre of an enzyme). Probably the best known example of this occurs within the cysteine proteases, such as caspases or papains, where cysteine is the key catalytic residue, being helped by a histidine and an asparagine.

14.5.19 Glycine (Gly, G)

14.5.19.1 Substitutions

Glycine can be substituted by other small amino acids, but be warned that even apparently neutral mutations (e.g. to alanine) can be forbidden in certain contexts (see below).

14.5.19.2 Structure

Glycine is unique as it contains a hydrogen as its side chain (rather than a carbon as is the case for all other amino acids). This means that there is much more conformational flexibility in glycine and as a result of this it can reside in parts of protein structures that are forbidden to all other amino acids (e.g. tight turns in structures).

14.5.19.3 Function

The uniqueness of glycine also means that it can play a distinct functional role, such as using its backbone (without a side chain) to bind to phosphates (Schulze-Gahmen *et al*., 1996). This means that if one sees a conserved glycine changing to any other amino acid, the change could have a drastic impact on function.

A good example is found among the protein kinases. Figure 14.9 shows a region around the ATP binding site in a protein kinase; the ATP is shown to the right of the figure and part of the protein to the left. The glycines in this loop are part of the classic 'Gly-X-Gly-X-X-Gly' motif present in the kinases (Hanks *et al*., 1988). These three glycines are almost never mutated to other residues; only glycines can function to bind to the phosphates of the ATP molecule using their main chains.

14.5.20 Proline (Pro, P)

14.5.20.1 Substitutions

Proline can sometimes substitute for other small amino acids, although its unique properties mean that it does not often substitute well.

14.5.20.2 Structure

Proline is unique in that it is the only amino acid where the side chain is connected to the protein backbone twice, forming a five-membered ring. Strictly speaking, this makes proline an imino acid (since in its isolated form, it contains an NH^{2+} rather than an NH^{3+} group, but this is mostly just pedantic detail). This difference is very important as

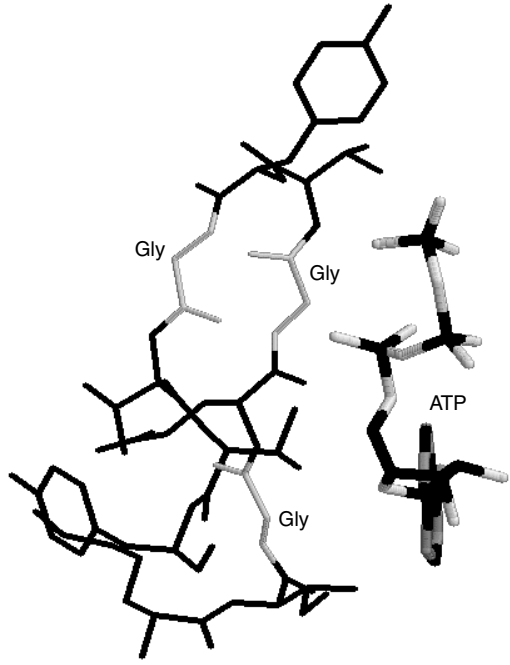

Figure 14.9 Glycine-rich phosphate binding loop in a protein kinase (code 1hck; Schulze-Gahmen *et al*., 1996).

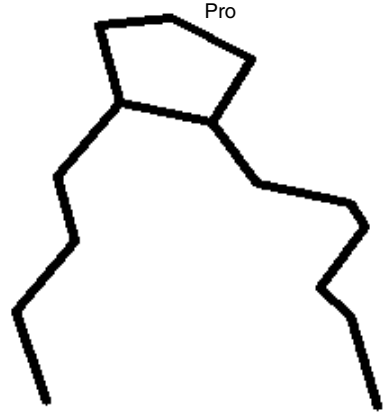

Figure 14.10 Example of proline in a tight protein turn (code 1ag6).

it means that proline is unable to occupy many of the main-chain conformations easily adopted by all other amino acids. In this sense, it can be considered to be an opposite of glycine, which can adopt many more main-chain conformations. For this reason proline is often found in very tight turns in protein structures (i.e. where the polypeptide chain must change direction; Figure 14.10). It can also function to introduce kinks into α-helices,

since it is unable to adopt a normal helical conformation. Despite being aliphatic the preference for turn structure means that prolines are usually found on the protein surface.

14.5.20.3 Function

The proline side chain is very non-reactive. This, together with its difficulty in adopting many protein main-chain conformations means that it is very rarely involved in protein active or binding sites.

14.6 STUDIES OF HOW MUTATIONS AFFECT FUNCTION

Several studies have been carried out previously in an attempt to derive general principles about the relationship between mutations, structure, function and diseases. We review some of these below.

14.6.1 Single Nucleotide Polymorphisms (SNPs)

A SNP is a point mutation that is present at a measurable frequency in human populations. They can occur either in coding or non-coding DNA. Non-coding SNPs may have effects on important mechanisms such as transcription, translation and splicing. However, the effects of coding SNPs are easier to study and are potentially more damaging, and so they have received considerably more attention. They are also more relevant to this chapter. Coding SNPs can be divided into two main categories, synonymous (where there is no change in the amino acid coded for), and non-synonymous. Non-synonymous SNPs tend to occur at lower frequencies than synonymous SNPs. Minor allele frequencies also tend to be lower in non-synonymous SNPs. This is a strong indication that these replacement polymorphisms are deleterious (Cargill *et al*., 1999).

To examine the phenotypic effects of coding SNPs, Sunyaev *et al*. (2000) studied the relationships between non-synonymous SNPs and protein structure and function. Three sets of SNP data were compared: disease causing susbtitutions, substitutions between orthologues and those represented by human alleles. Disease-causing mutations were more common in structurally and functionally important sites than were variations between orthologues, as might be expected. Allelic variations were also more common in these regions than were those between orthologues. Minor allele frequency and the level of occurrence in these regions were correlated, another indication of evolutionary selection of phenotype. The most damaging allelic variants affect protein stability, rather than binding, catalysis, allosteric response or post-translational modification (Sunyaev *et al*., 2001). The expected increase in the number of known protein structures will allow other analyses and refinement of the details of the phenotypic effects of SNPs.

Wang and Moult (2001) developed a description of the possible effects of missense SNPs on protein structure and used it to compare disease-causing missense SNPs with a set from the general population. Five general classes of effect were considered: protein stability, ligand binding, catalysis, allosteric regulation and post-translational modification. The disease and population sets of SNPs contain those that can be mapped onto known protein structures, either directly or through homologues of known structure. Of the disease-causing SNPs, 90% were explained by the description, with the majority (83%) being attributed to effects on protein stability, as reported by Sunyaev *et al*. (2001). The 10% that are not explained by the description may cause disease by effects not easily identified by structure alone. Of the SNPs from the general population, 70% were predicted

to have no effect. The remaining 30% may represent disease-causing SNPs previously unidentified as such, or molecular effects that have no significant phenotypic effect.

14.6.2 Site-directed Mutagenesis

Site-directed mutagenesis is a powerful tool for discovering the importance of an amino acid in the function of the protein. Gross changes in amino acid type can reveal sites that are important in maintaining the structure of the protein. Conversely, when investigating functionally interesting sites it is important to choose replacement residues that are unlikely to affect structure dramatically, for example by choosing ones of a similar size to the original. Peracchi (2001) reviews the use of site-directed mutagenesis to investigate mechanisms of enzyme catalysis, in particular those studies involving mutagenesis of general acids (proton donors), general bases (proton acceptors) and catalytic nucleophiles in active sites. These types of amino acid could be considered to be the most important to enzyme function as they directly participate in the formation or cleavage of covalent bonds. However, studies indicate that they are often important but not essential — rates are still higher than the uncatalysed reaction even when these residues are removed, because the protein is able to use an alternative mechanism of catalysis. Also, direct involvement in the formation and cleavage of bonds is only one of a combination of methods that an enzyme can use to catalyse a reaction. Transition states can be stabilized by complementary shape and electrostatics of the binding site of the enzyme and substrates can be precisely positioned, lowering the entropy of activation. These factors can also be studied by site-directed mutagenesis, with consideration of the physical and chemical properties of the amino acids again guiding the choice of replacements, along with knowledge of the structure of the protein.

14.6.3 Key Mutations in Evolution

Golding and Dean (1998) reviewed six studies that demonstrate the insight into molecular adaptation that is provided by combining knowledge of phylogenies, site-directed mutagenesis and protein structure. These studies emphasize the importance of protein structure when considering the effects of amino acid mutations.

Many changes can occur over many generations, with only a few being responsible for changes in function. For example, the sequences of lactate dehydrogenase (LDH) and malate dehydrogenase (MDH) from *Bacillus stearothermophillus* are only about 25% identical, but their tertiary structures are highly similar. Only one mutation, of uncharged glutamine 102 to positive arginine in the active site, is required to convert LDH into a highly specific MDH. The arginine is thought to interact with the carboxylate group which is the only difference between the substrate/products of the two enzymes (Figure 14.11; Wilks *et al*., 1988).

Thus amino acid changes that appear to be radical or conservative from their scores in mutation matrices or amino acid properties may be the opposite when their effect on protein function is considered; glutamine to arginine has a score of 0 in the PAM250 matrix, meaning that it is neutral. The importance of the mutation at position 102 in LDH and MDH could not be predicted using this information alone.

Another study showed that phylogeny and site-directed mutagenesis can identify key amino acid changes that would likely be overlooked if only structure was considered; the reconstruction of an ancestral ribonuclease showed that the mutation that causes most of the five-fold loss in activity towards double-stranded RNA is of Gly38 to Asp, more than 5 Å from the active site (Golding and Dean, 1998).

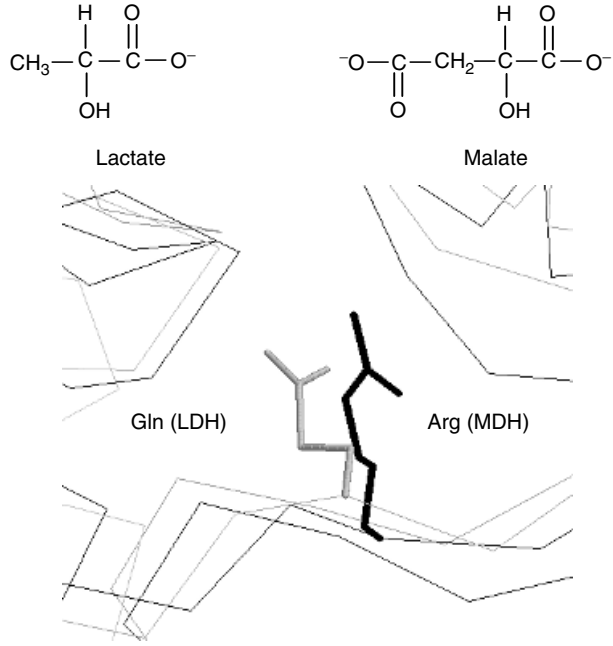

Figure 14.11 Lactate and malate dehydrogenase specificity (codes 9ltd and 2cmd; Wilks *et al*., 1988).

A third study showed that knowledge of structure can be important in understanding the effects of mutations. Two different mutations in different locations in the haemoglobin genes of the bar-headed goose and Andean goose give both species a high affinity for oxygen. Structural studies showed that both changes remove an important van der Waals contact between subunits, shifting the equilibrium of the haemoglobin tetramer towards the high-affinity state. The important point in all these studies is that no single approach, such as phylogeny alone or structural studies alone, is enough to understand the effects of all amino acid mutations.

14.7 A SUMMARY OF THE THOUGHT PROCESS

It is our hope that this chapter has given the reader some guidelines for interpreting how a particular mutation might affect the structure and function of a protein. Our suggestion would be that you ask the following questions:

First about the protein:

1. What is the cellular environment?
2. What does it do? Is anything known about the amino acids involved in its function?
3. Is there a structure known or one for a homologue?
4. What protein family does it belong to?
5. Are any post-translational modifications expected?

Then about a particular amino acid:

1. Is the position conserved across orthologues? Across paralogues?
2. If a structure is known: is the amino acid on the surface? Buried in the core of the protein?
3. Is it directly involved in function or near (in sequence or space) to other amino acids that are?
4. Is it an amino acid that is likely to be critical for function? For structure?

Once these questions have been answered it should be possible to make a rational guess or interpretation of effects seen by an amino acid substitution and select logical amino acids for mutagenesis experiments.

REFERENCES

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, *et al*. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389–3402.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, *et al*. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet* **25**: 25–29.

Barnes MR, Russell RB. (1999). A lipid-binding domain in Wnt: a case of mistaken identity? *Curr Biol* **9**: R717–R719.

Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer EL. (2000). The Pfam protein families database. *Nucleic Acids Res* **28**: 263–266.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, *et al*. (2000). The Protein Data Bank. *Nucleic Acids Res* **28**: 235–242.

Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, *et al*. (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet* **22**: 231–238.

Copley RR, Barton GJ. (1994). A structural analysis of phosphate and sulphate binding sites in proteins. Estimation of propensities for binding and conservation of phosphate binding sites. *J Mol Biol* **242**: 321–329.

Dayhoff MO, Schwartz RM, Orcutt BC. (1978). A model of evolutionary change in proteins. In Dayhoff MO. (Ed.), *Atlas of Protein Sequence and Structure*, Vol. 5. National Biomedical Research Foundation: Washington DC, pp. 345–352.

Durbin R, Eddy S, Krogh A, Mitchison G. (1998). *Biological Sequence Analysis. Probabalistic Models of Proteins and Nucleic Acids*. Cambridge University Press: Cambridge.

Eddy, SR (1998). Profile hidden Markov models. *Bioinformatics* **14**: 755–763.

Gavel Y, von Heijne G. (1990). Sequence differences between glycosylated and non-glycosylated Asn-X-Thr/Ser acceptor sites: implications for protein engineering. *Protein Eng* **3**: 433–442.

Gibson TJ, Spring J. (1998). Genetic redundancy in vertebrates: polyploidy and persistence of genes encoding multidomain proteins. *Trends Genet* **14**: 46–49; discussion 49–50.

Golding GB, Dean AM. (1998). The structural basis of molecular adaptation. *Mol Biol Evol* **15**: 355–369.

Gupta R, Birch H, Rapacki K, Brunak S, Hansen JE. (1999). O-GLYCBASE version 4.0: a revised database of O-glycosylated proteins. *Nucleic Acids Res* **27**: 370–372.

Hanks SK, Quinn AM, Hunter T. (1988). The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. *Science* **241**: 42–52.

Henikoff S, Henikoff JG. (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA* **89**: 10915–10919.

Hunter CA, Singh J, Thornton JM. (1991). Pi–pi interactions: the geometry and energetics of phenylalanine–phenylalanine interactions in proteins. *J Mol Biol* **218**: 837–846.

Jeffrey PD, Russo AA, Polyak K, Gibbs E, Hurwitz J, Massague J, *et al*. (1995). Mechanism of CDK activation revealed by the structure of a cyclinA–CDK2 complex. *Nature* **376**: 313–320.

Jones DT, Taylor WR, Thornton JM. (1994). A mutation data matrix for transmembrane proteins. *FEBS Lett* **339**: 269–275.

Kraulis PJ. (1991). MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J Appl Crys* **24**: 946–950.

Krishna RG, Wold, F. (1993). Post-translational modification of proteins. *Adv Enzymol Relat Areas Mol Biol* **67**: 265–298.

Macias MJ, Wiesner S, Sudol M. (2002). WW and SH3 domains, two different scaffolds to recognize proline-rich ligands. *FEBS Lett* **513**: 30–37.

Parekh RB, Rohlff C. (1997). Post-translational modification of proteins and the discovery of new medicine. *Curr Opin Biotechnol* **8**: 718–723.

Peracchi A. (2001). Enzyme catalysis: removing chemically 'essential' residues by site-directed mutagenesis. *Trends Biochem Sci* **26**: 497–503.

Plotnikov AN, Schlessinger J, Hubbard SR, Mohammadi M. (1999). Structural basis for FGF receptor dimerization and activation. *Cell* **98**: 641–650.

Russell RB. (1998). Detection of protein three-dimensional side chain patterns: new examples of convergent evolution. *J Mol Biol* **279**: 1211–1227.

Sayle RA, Milner-White EJ. (1995). RASMOL: biomolecular graphics for all. *Trends Biochem Sci* **20**: 374.

Schoorlemmer J, Goldfarb M. (2001). Fibroblast growth factor homologous factors are intracellular signaling proteins. *Curr Biol* **11**: 793–797.

Schulze-Gahmen U, De Bondt HL, Kim SH. (1996). High-resolution crystal structures of human cyclin-dependent kinase 2 with and without ATP: bound waters and natural ligand as guides for inhibitor design. *J Med Chem* **39**: 4540–4546.

Sunyaev S, Lathe W III, Bork P. (2001). Integration of genome data and protein structures: prediction of protein folds, protein interactions and 'molecular phenotypes' of single nucleotide polymorphisms. *Curr Opin Struct Biol* **11**: 125–130.

Sunyaev S, Ramensky V, Bork P. (2000). Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends Genet* **16**: 198–200.

Taylor WR. (1986). The classification of amino acid conservation. *J Theor Biol* **119**: 205–218.

Waksman G, Kominos D, Robertson SC, Pant N, Baltimore D, Birge RB, *et al*. (1992). Crystal structure of the phosphotyrosine recognition domain SH2 of v-src complexed with tyrosine-phosphorylated peptides. *Nature* **358**: 646–653.

Wang Z, Moult J. (2001). SNPs, protein structure, and disease. *Hum Mutat* **17**: 263–70.

Wilks HM, Hart KW, Feeney R, Dunn CR, Muirhead H, Chia WN, *et al*. (1988). A specific, highly active malate dehydrogenase by redesign of a lactate dehydrogenase framework. *Science* **242**: 1541–1544.

Wilson CA, Kreychman J, Gerstein M. (2000). Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J Mol Biol* **297**: 233–249.

Wolfe SA, Grant RA, Elrod-Erickson M, Pabo CO. (2001). Beyond the 'recognition code': structures of two Cys2His2 zinc finger/TATA box complexes. *Structure (Camb)* **9**: 717–723.

APPENDIX: TOOLS

Protein sequences

http://www.expasy.ch/
http://www.ncbi.nlm.nih.gov/

Amino acid properties

http://russell.embl-heidelberg.de/aas/

Domain assignment/sequence search tools

http://www.ebi.ac.uk/interpro/
http://www.sanger.ac.uk/Software/Pfam/
http://smart.embl-heidelberg.de/
http://www.ncbi.nlm.nih.gov/BLAST/
http://www.ncbi.nlm.nih.gov/COG/
http://www.cbs.dtu.dk/TargetP/

Protein structure

Databases of 3D structures of proteins
http://www.rcsb.org/pdb/
Structural classification of proteins
http://scop.mrc-lmb.cam.ac.uk/scop/

Protein function

http://www.geneontology.org/