Understanding the Relationship Between Enzyme Structure and Catalysis

Alex Gutteridge*

Darwin College, Cambridge

This dissertation is submitted for the degree of Doctor of Philosophy

October 15, 2005

*EBI, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SD.

Preface

This dissertation is the result of my own work, and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text.

The length of this dissertation does not exceed the word limit specified by the Graduate School of Biological, Medical and Veterinary Sciences.

Abstract

The three dimensional structure of an enzyme is of fundamental importance to almost every function it performs. In this thesis, we aim to analyse the structures of many enzymes in order to gain insights into how they perform catalysis, and the role of structure in enzyme function. As well as improving our understanding of these important biological molecules, these insights will help in developing new tools for annotating enzyme structures of unknown function and for designing novel enzymes.

Predicting the location of the active site, and the identity of the catalytic residues, is an important first step for annotating an enzyme of unknown function. We have developed a neural network trained to distinguish catalytic and non-catalytic residues based on a mixture of sequence and structural parameters. We find that the correct location of the active site can be predicted in ~70% of cases. We also find that the most important factor in making a prediction is the conservation score of each residue. However, including structural data does improve the predictions that are made over those made on the basis of conservation alone.

Our first analysis of enzyme structure aims to measure the extent of conformational change undergone upon substrate binding. We find that most enzymes do not undergo large scale conformational change, and in many cases the catalytic residues are isolated from changes that do occur. One new theory of enzyme action suggests that certain, very small, conformational changes (deriving from changes in the enzyme dynamics) are important in catalysis. In a separate study, we look for these changes and find hints that they may occur in some enzymes, although they are at the limit of what can be reliably observed in crystal structures.

Having established that catalytic site is generally preformed, we next look at the interactions between catalytic residues. We analyse a large set of interactions and the roles they play in catalysis. We find that many catalytic residues do not require direct interactions with other catalytic residues to be active, although we do predict that many previously unidentified interactions will prove to be functionally important. Those residues that do commonly interact are often important in the pK_a modulation of other residues. We examine in greater detail several important interactions, and observe a preference for certain catalytic interactions to adopt different distributions of geometries to non-catalytic interactions.

In conclusion, it is clear that structure plays an important, though sometimes subtle, role in catalysis. We have seen that conformational changes in enzymes are often small, but that even small changes could be significant. We have also found that even the smallest chemical differences between groups are often exploited by enzyme mechanisms. We can use this information on how Nature uses these different groups in catalysis to improve our ability to predict and design new enzyme functions.

Acknowledgements

First and foremost, thanks must go to my supervisor Prof Thornton. Without her support, insight and encouragement, this thesis would be a lot less interesting and contain a lot more spelling mistaks.

Thanks and acknowledgement must also go to many other past and present members of the Thornton group both at the EBI and UCL. Firstly, thanks to Dr Craig Porter for two summer jobs, lots of pizza and more enzyme annotation than I could ever wish for. Also, thanks to Dr Roman 'Romanager' Laskowski, who despite his failing powers in recent years, has remained an inspiration both in the lab and on the football pitch.

At the EBI, thanks first to Dr Gail Bartlett, for guidance and inspiration in the formative months of my Phd. Her work, and the thesis she wrote from it, has been a template for (all the good parts) of this dissertation. Other special mentions go to Gareth Stockwell and Dr Gordon Whammond (and Dr Kevin Murray - where ever he might be) for giving me somewhere to live (even if it was in the middle of nowhere) in my first year. And to Gareth (again), Gail (again), Prof Barbara Brodsky, James Torrance and Dr Jonathan Barker for putting up with my (occasionally) smelly football kits, and for providing all the office fun and games I've (usually) enjoyed over the last three years

Outside of work, thanks go to Sheena Patel for all her support and friendship,

everyone at Drayton Park for whatever they've been doing, my Mum and Dad for trying to understand the various bits of pieces of what I've been doing for the last three years and last but not least, Sowmiya Moorthie, for making Cambridge a fun place to live, even for a big city boy.

Actually, last thanks to all the hard-working crystallographers and experimentalists who have done the work to generate the data from which I've built this thesis.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Enzymes as Catalysts

1.1.1 The Importance of Enzymes For Life

One of the fundamental properties of all living organisms is the process of metabolism, whereby organic compounds are synthesised (anabolism) and broken down (catabolism). The metabolism (derived from the Greek ‘metabolismos’, meaning ‘change’) of a whole cell is an extremely complex system, but it can be broken down into subsystems and pathways, which are comprised of individual reactions that change one compound into another. Figure 1.1 shows the metabolism for the bacteria *Escherichia coli*, a single pathway within the metabolic network and a single reaction within that pathway. It is this final level of organisation, the individual reactions themselves, that this thesis concerns itself.

Metabolic reactions are chemical reactions like any other, there are substrates consumed and products produced, and a change in the free energy between the start and the end of the reaction. Each reaction proceeds spontaneously in the direction in which this free energy change (ΔG) is negative. This is formalised in Equation

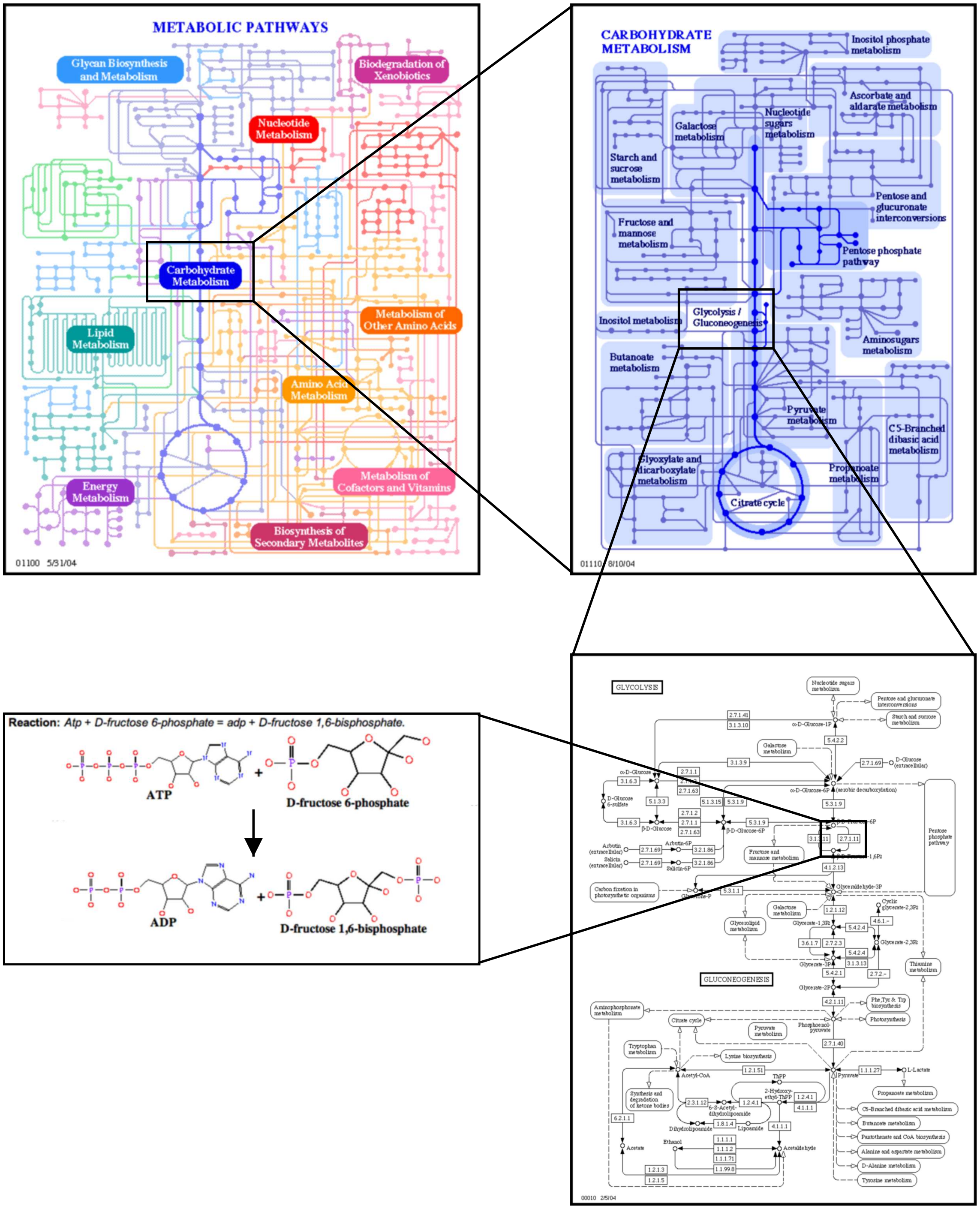

METABOLIC PATHWAYS
Glycan Biosynthesis and Metabolism
Biodegradation of Xenobiotics
Nucleotide Metabolism
Carbohydrate Metabolism
Lipid Metabolism
Metabolism of Other Amino Acids
Amino Acid Metabolism
Energy Metabolism
Metabolism of Cofactors and Vitamins
Biosynthesis of Secondary Metabolites
01100 5/31/04
CARBOHYDRATE METABOLISM
Inositol phosphate metabolism
Ascorbate and aldarate metabolism
Nucleotide sugars metabolism
Galactose metabolism
Starch and sucrose metabolism
Pentose and glucuronate interconversions
Fructose and mannose metabolism
Pentose phosphate pathway
Glycolysis / Gluconeogenesis
Inositol metabolism
Aminosugars metabolism
Butanoate metabolism
Pyruvate metabolism
C5-Branched dibasic acid metabolism
Glyoxylate and dicarboxylate metabolism
Propanoate metabolism
Citrate cycle
01110 8/10/04
GLYCOLYSIS
GLUCONEOGENESIS
00010 2/5/04
Reaction: Atp + D-fructose 6-phosphate = adp + D-fructose 1,6-bisphosphate.
ATP
D-fructose 6-phosphate
ADP
D-fructose 1,6-bisphosphate

Figure 1.1: Clockwise from top left: All the metabolic pathways annotated in the KEGG[1] database, those pathways associated with carbohydrate metabolism, the glycolysis and gluconeogenesis pathways, the reaction catalysed by the enzyme 6-phosphofructokinase within glycolysis.

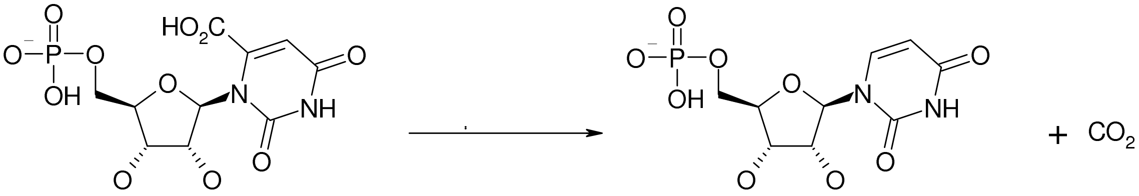


Figure 1.2: The decarboxylation of orotic acid as catalysed by orotidine 5'-phosphate decarboxylase. The reaction is speeded up by a factor of 10^{17} by the enzyme.

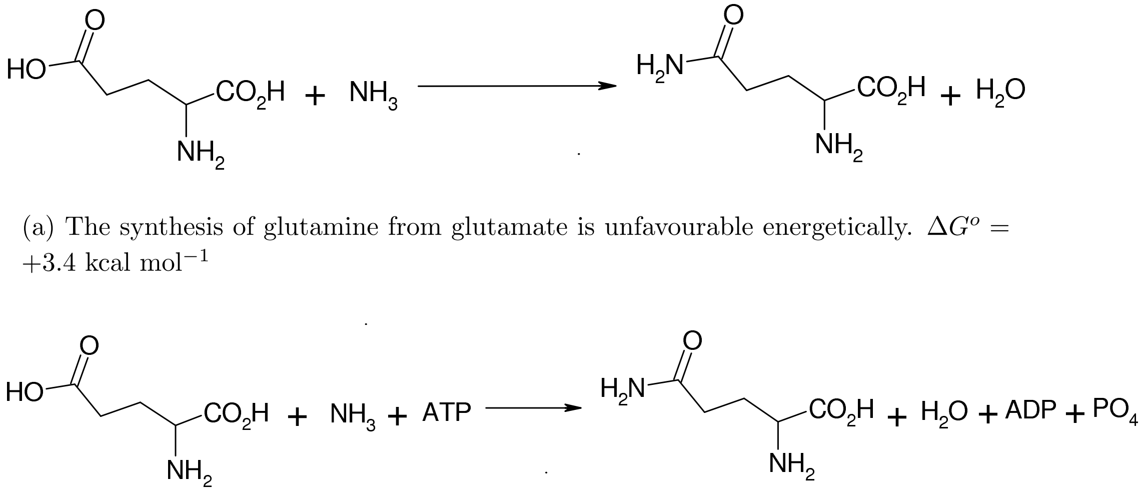
1.1 which links ΔG^o (the free energy change at standard state) to the equilibrium constant of the reaction K, the gas constant R and the absolute temperature T.

$$\Delta G^o = -RTlnK \tag{1.1}$$

Although ΔG determines the overall direction a reaction proceeds in, the rate of a reaction depends on another quantity known as the activation energy. If the activation energy is too high, the reaction will proceed very slowly even if the ΔG is favourable. For instance, the spontaneous decarboxylation of orotic acid (shown in Figure 1.2) has a half life of 78 million years at room temperature[2]. This is many orders of magnitude too slow to be useful for a biological process, yet this reaction is the last step in the *de novo* synthesis of uridine monophosphate (UMP), an essential cellular metabolite.

Organisms can use such slow reactions in their metabolism because they speed them up using biological catalysts called enzymes. In the example given above, the enzyme orotidine 5'-phosphate decarboxylase is used, which enhances the rate of the reaction by 10^{17} times, such that the half life of the enzyme catalysed reaction is less than $1/40^{th}$ of a second. Enzymes allow many reactions to take place in cellular metabolism, which, without catalysis, would proceed too slowly to be used otherwise.

As well as rate enhancement, enzymes also make it possible to 'reverse' reactions.



(a) The synthesis of glutamine from glutamate is unfavourable energetically. $\Delta G^o = +3.4$ kcal mol^{-1}

(b) By coupling the synthesis of glutamine to the dephosphorylation of ATP the overall reaction is favourable energetically. $\Delta G^o = -3.9$ kcal mol^{-1}

Figure 1.3: The synthesis of glutamine from glutamate is coupled by the enzyme glutamine synthetase, to the dephosphorylation of ATP to make it energetically favourable.

Enzymes achieve this by coupling two different reactions together energetically. For instance, the synthesis of glutamine from glutamate shown in Figure 1.3(a) has a ΔG^o of +3.4 kcal mol^{-1}. However, the enzyme responsible for this reaction, glutamine synthetase, couples it to a second reaction: the dephosphorylation of adenosine triphosphate (ATP) to adenosine diphosphate (ADP) and inorganic phosphate (PO_4). This dephosphorylation reaction has a ΔG^o of -7.3 kcal mol^{-1}, so when these two reactions are combined to get the complete enzyme catalysed reaction shown in Figure 1.3(b) the overall ΔG^o is -3.9 kcal mol^{-1} (+3.4 kcal mol^{-1} + -7.3 kcal mol^{-1}) and so, provided there is a supply of ATP, glutamine can be synthesised from glutamate.

The free energy changes undergone in a reaction are usually represented using a diagram such as that shown in Figure 1.4. The free energy of the whole system is shown on the vertical co-ordinate, and the progress of the reaction is shown on the horizontal axis. The peak in free energy in the centre of the reaction represents the transition state (TS) that is formed between the substrate (S) and the product (P). The transition state is usually highly unstable because of its high free energy, but it must be formed for the reaction to proceed.

The rate of a reaction depends on the difference in free energy between the substrate and the transition state. This is the activation energy, and is shown as E_A in Figure 1.4. All catalysts, including enzymes, speed up reactions by lowering the activation energy as shown in Figure 1.5(a). However, enzymes are more complex than simple chemical catalysts, partly because they have to bind to their substrates prior to catalysis, but also because the enzyme catalysed reactions can often be broken down into several smaller steps with stable intermediates along the reaction co-ordinate as shown in Figure 1.5(b).

In this thesis, we aim to understand better how enzymes perform catalysis on the molecular level. The key to this understanding is the three dimensional structure of each enzyme, which determines almost every aspect of an enzyme's function: substrate binding, chemical catalysis, regulation and stability. By drawing together large numbers of enzyme structures for analysis we try to determine both features that are common to all enzymes, and features which are specific to enzymes of particular catalytic or cellular functions. Understanding how enzymes work, both generally and specifically, is important for our understanding of cellular metabolism as a whole, our ability to explain the effects of and treat genetic diseases, which might disrupt enzyme function, how to predict enzyme function and how we can predict the effect of drugs or toxins on metabolism.

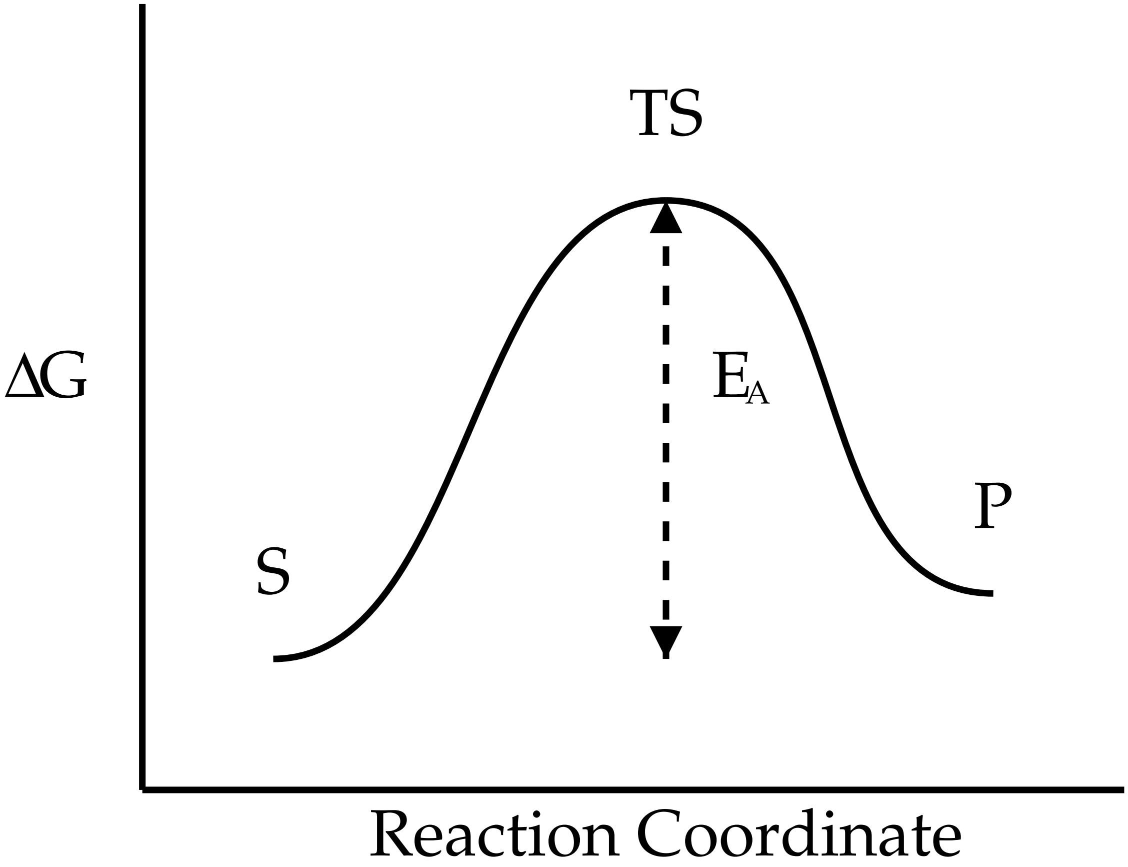
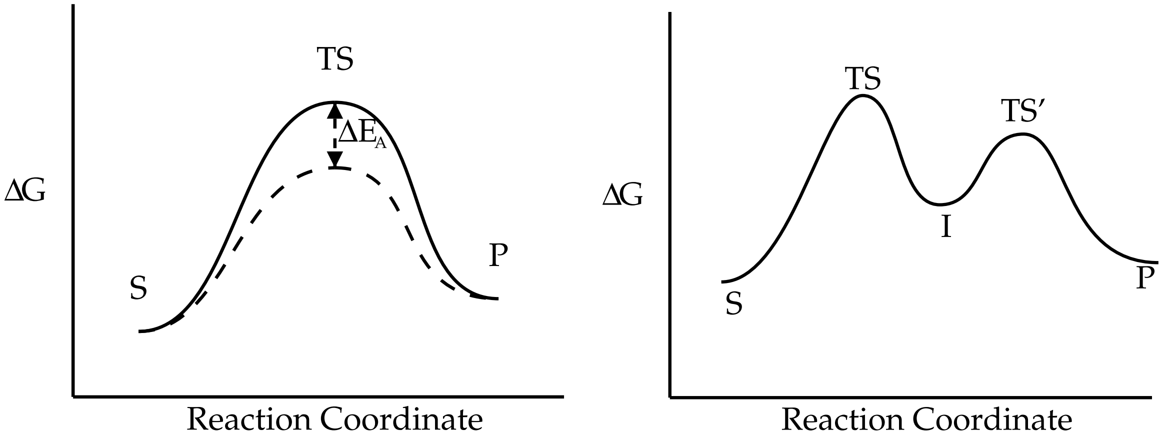


Figure 1.4: The free energy of an uncatalysed reaction.



(a) Free energy profile of a catalysed reaction.

(b) Free energy profile of an enzyme catalysed reaction with an intermediate (I).

Figure 1.5: A catalyst lowers the free energy of the transition state (TS). The efficiency of catalysis depends on ΔE_A, the difference between the free energies of the catalysed and uncatalysed transition state.An enzyme lowers the free energy of the transition state, but often also modifies the reaction that takes place so that stable intermediates (I) are formed.

1.1.2 Historical Perspective on Enzyme Catalysis

The basic principles of enzyme catalysis began to be revealed in the last decade of the 19^{th} and first half of the 20^{th} centuries. The earliest studies were made on fermentation in yeast, and it was from observing that different enzymes were required for the hydrolysis of polysaccharides, depending on whether the polysaccharide was made from α-D-glucose or β-D-glucose subunits, that, in 1894, Emil Fischer developed the 'lock and key' hypothesis of enzyme action. Fischer proposed that enzymes have a defined three-dimensional structure that fits (is complementary to) the structure of the substrate, just as a lock fits a key. This explained the precise selectivity of enzymes for their substrates that Fischer observed, and with the demonstration in 1897, by Eduard Buchner, that fermentation continued in cell free extracts, showed that enzyme catalysis was not just a property of living cells as had been previously believed. Despite these advances, it was not until 1926 that James Summer proved that biological catalysis was carried out by a chemical species, and it was several years after this that it was widely accepted that enzymes were proteins[3].

The study of enzyme kinetics was also developed around this time, with Michaelis and Menten introducing their eponymous equation (Equation 1.2) in 1913. This successfully modelled enzyme kinetics by relating the rate of the reaction (v) to: k_{cat}, the first order rate constant for the slowest step in the reaction, representing the maximum number of substrate molecules which can be consumed per enzyme molecule per unit time, K_M, a measure of the enzymes ability to bind substrate, the concentration of the enzyme ($[E]$) and the concentration of the substrate ($[S]$).

In 1925, Briggs and Haldane rationalised the Michaelis-Menten equation by applying the steady-state approximation to the model of enzyme action shown in Equation 1.3. This model defines three states: the free enzyme and substrate in solution ($E+S$), the substrate bound to the enzyme (ES) and the free enzyme and product ($E+P$). K_M, k_{cat}, and k_{cat}/K_M (the specificity constant which provides a measure of the overall efficiency of the enzyme) remain the standard parameters for describing the kinetic properties of an enzyme.

$$v = \frac{k_{cat}[E][S]}{K_m + [S]} \tag{1.2}$$

$$E + S \rightleftharpoons ES \rightarrow E + P \tag{1.3}$$

Fischer's original conception of the enzyme fitting the substrate as a lock fits a key was slowly modified during the early 20^{th} century. In 1930, Haldane modified the theory by proposing that 'the key does not fit the lock perfectly but exercises a certain strain on it'. In this conception, the enzyme distorts the substrate such that it is more likely to react and hence the reaction is speeded up. This theory of strained binding was developed by Pauling into the theory of transition state stabilisation, where the enzyme is complementary to the transition state rather than the substrate[3]. The enzyme, by binding better to the transition state than the substrate, reduces the free energy of the transition state and so, as shown in Figure 1.5(a) reduces the energy of activation.

Throughout the second half of the 20^{th} century the theory of transition state stabilisation has developed to try and fully explain the extraordinarily high rate increases that enzymes are capable of. In 1962, it was shown that the placement of catalytic groups around the substrate leads to an extremely high 'effective concentration' of reactants leading to a similarly high rate of reaction[3]. More recently, the role of dynamics within the enzyme structure in controlling quantum tunnelling by hydrogen atoms[4] and the forcing of substrates into 'near attack conformers' suitable for reaction[5] have been proposed to try to explain the rate increases achieved by some enzymes.

1.1.3 Chemical Catalysis

The studies described above describe the general mechanisms by which all enzymes are understood to operate: substrates are bound and converted via stabilised transition states into products. However, the precise chemical ways in which this is achieved vary from one enzyme to another: depending on the nature of the reaction to be catalysed, and the vagaries of evolution that have shaped the enzyme. In this section we summarise some of the chemical mechanisms used in enzyme catalysis.

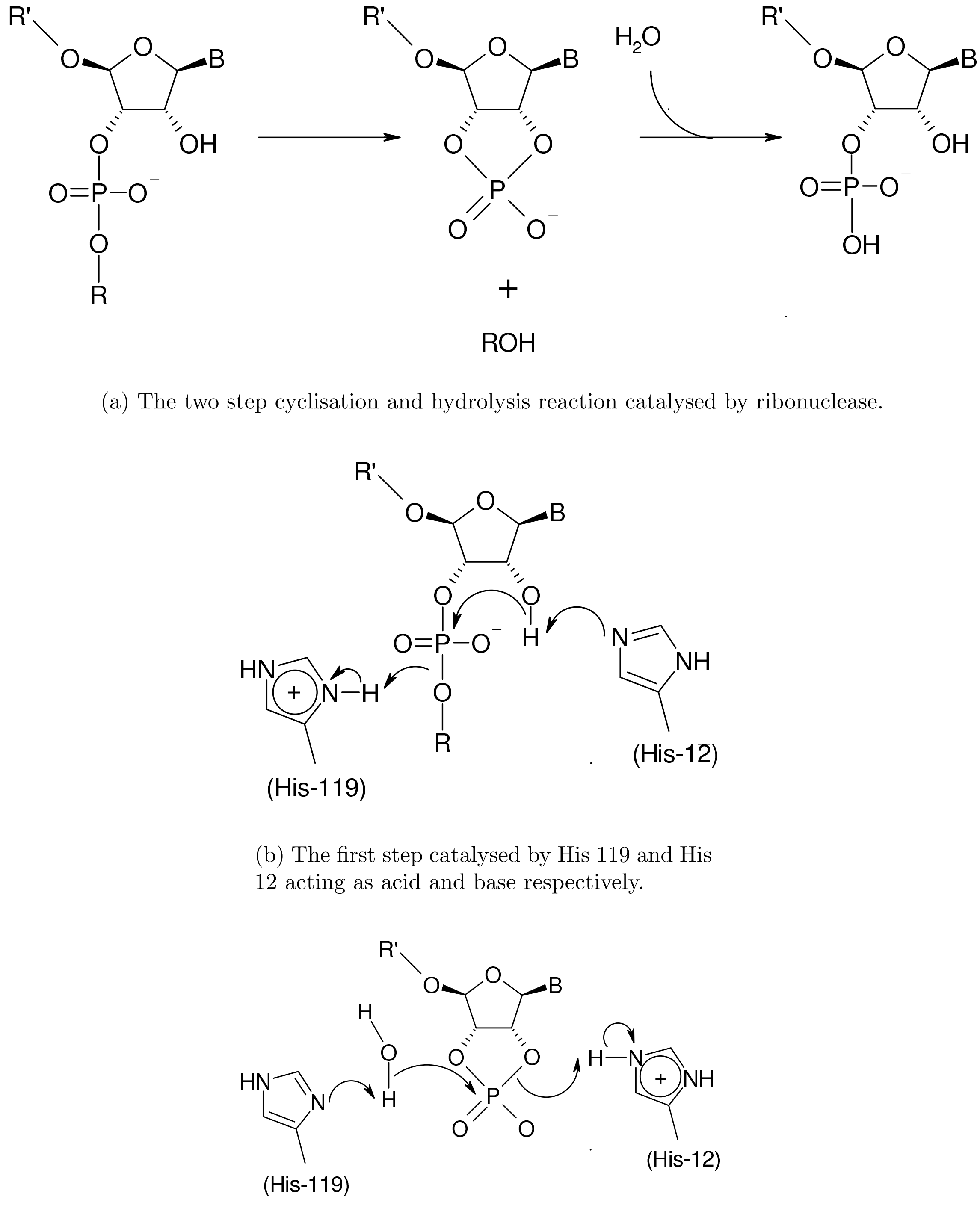
Acid/base Catalysis

Much of the instability of a typical transition state is due to the build up of unfavourable charges within the molecule as bonds form and break. Stabilising these charges is an important function of catalytic groups within the enzyme.

One mechanism for charge stabilisation is through acid/base catalysis. This process stabilises the transition state, by abstracting or donating protons from the substrate to groups on the enzyme. Ribonuclease is a classic example of acid/base catalysis, the structure of the enzyme was solved in 1967[6] and confirmed a mechanism proposed in 1961[7] based on the pH activity curve.

Ribonuclease catalyses the hydrolysis of RNA in a two-step process during which a cyclic intermediate is formed as shown in Figure 1.6(a). The uncatalysed reaction involves the unfavourable development of positive and negative charges on the 2'-OH and the phosphate oxygen respectively during the first step of the reaction, and the reverse charges in the second step. The enzyme catalysed reaction avoids these charges by providing two histidine residues that are capable of donating and accepting protons that neutralise these charges as shown in Figures 1.6(b) and 1.6(c).

Acid/base catalysis depends strongly on the basic or acidic strength of the catalytic group and its ionization state, which is determined by its pK_a and the pH the reaction takes place in. A strongly basic or acidic group is potentially a powerful catalyst, because it will tend to strongly abstract or donate protons from other groups, but since its pK_a value will be far from neutral, it will rarely be in the correct protonation state (protonated for acids and unprotonated for bases). The best acid/base catalysts are often the weaker acid/bases, such as histidine with a pK_a of around seven, as they are more likely to be in the correct ionization state.



(a) The two step cyclisation and hydrolysis reaction catalysed by ribonuclease.

(b) The first step catalysed by His 119 and His 12 acting as acid and base respectively.

(c) The second hydrolysis step catalysed by His 119 and His 12 acting as a base and an acid respectively.

Figure 1.6: Acid/base catalysis as performed by ribonuclease.

Electrostatic Catalysis

Just as some enzymes, like ribonuclease, use protons to neutralise charges that develop in transition states, other enzymes use complementary charged groups on the enzyme to stabilise charged transition states. These charges can be provided by many different groups: dipoles such as backbone amide groups are sufficient in serine proteases[8], whilst other enzymes use fully charged groups such as the arginines used in Staphylococcal nuclease[9] and the Zn^{2+} ion found in carboxypeptidase[10]. In each case the charges stabilise the transition state and so enhance catalysis.

Covalent Catalysis

Many enzymes alter the mechanism by which a reaction proceeds by forming covalent bonds to the substrate(s). The chemical modification of the substrate serves to activate it and so the reaction is speeded up. One common example is the formation of a Schiff base by the condensation of an amine with a carbonyl. This leads to the formation of a carbon-nitrogen double bond, with the nitrogen then able to take on a positive charge and act as an electron sink. The Schiff base activates either the carbonyl carbon for nucleophilic attack by another group, or one of the other groups bound to the carbon, by stabilising the development of negative charges on them. This type of catalysis is often performed by the cofactors pyridoxal phosphate[11] and thiamine pyrophosphate[12].

The Schiff base is an electrophilic group (the positively charged nitrogen attracts electrons), but covalent catalysis is also performed by nucleophiles. Enzymes can use nucleophiles to form covalently attached intermediates such as the tetrahedral intermediate found in serine proteases shown in Figure 1.7. This aids catalysis because the formation of this intermediate and its subsequent breakdown is easier to achieve than the uncatalysed reaction.

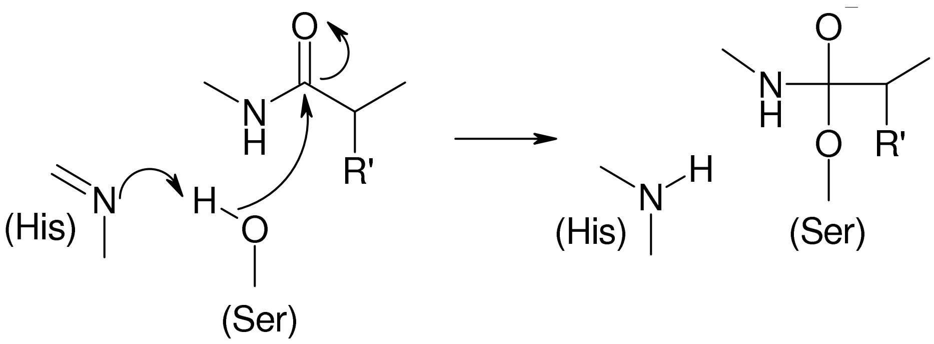


Figure 1.7: The serine proteases use nucleophilic catalysis to form a tetrahedral intermediate.

Entropy

The change in free energy (ΔG) of a reaction has two components shown in Equation 1.4: the enthalpy change (ΔH), representing the energy changes due to the making and breaking of bonds, and the entropy change (ΔS).

$$\Delta G = \Delta H - T\Delta S \tag{1.4}$$

For the purposes of this discussion we consider the entropy of a system to be equivalent to the freedom of movement (rotationally, translationally and vibrationally) of the molecules in that system. A process that restricts the motion of these molecules will reduce the entropy of the system and thus increase the free energy; making such a process unfavourable. In enzyme catalysis, the complete system we must consider comprises the enzyme itself, the substrates and the bulk solvent which surrounds them. Changes in any of these three components will affect the entropy of the system.

Joining two molecules together (for instance, as an enzyme binds a substrate) leads to a loss of entropy due to the restriction of rotational and translational movement. In enzymes, this entropic cost is partly offset by the entropic benefit of freeing water that was previously bound in the active site and the enthalpic benefits of forming bonds between the substrate and the enzyme.

Once the enzyme has bound the substrate, the reaction essentially takes place intramolecularly, since we can consider the enzyme-substrate complex as one molecule. This is an important advantages for enzymes over solution catalysts, because in the enzyme, there is no entropic cost to forming the transition state (the cost can be thought of as being paid during binding). In contrast, in the same reaction catalysed in solution, the reactants entropy is reduced as the transition state is formed, thus raising the free energy of the transition state. This entropic benefit of an intramolecular reaction over a bimolecular reaction has been estimated to be up to 190 J/deg/mol[13]; equivalent to a rate enhancement of 6×10^9. Thus we can see that entropy is one of the crucial aspects of enzyme catalysis that provides much of an enzyme's power.

1.2 The Structure and Evolution of Enzymes

1.2.1 Structural Studies of Enzymes

An enzyme's function is intrinsically linked to its three dimensional structure, determining how it performs substrate binding, catalysis and regulation. X-ray crystallography has been the most important technique in the development of our understanding of enzyme structure and hence enzyme function. Nuclear magnetic resonance (NMR) has also been used successfully to study many structures, but crystallography remains the principle technique for structure elucidation. The first enzyme to be crystallised and have its structure successfully solved was chicken egg lysozyme in 1965[14, 15]. Importantly, as well as the structure of the free enzyme, it was possible to crystallise lysozyme with a substrate analog bound in the active site. This structure, allowed the proposal of a chemical mechanism for the enzyme, based on positioning of groups around the site of substrate cleavage. The use of crystal structures with bound substrate and transition state analogs has helped to reveal the catalytic mechanisms of countless enzymes since.

The structure of lysozyme was solved to a resolution of 2Å; at this resolution it is possible to accurately place the residue side chains and the plane of each peptide bond. However, individual atoms are not generally well resolved. ‘Atomic’ resolution (1.2Å resolution or higher) allows the placement of atoms with fewer geometrical restraints and so gives a better picture of the ‘true’ protein structure. In recent years, advances in X-ray sources and cryocrystallography have led to increasing numbers of structures solved at these high resolutions[16].

Any structure, no matter what the resolution, contains a certain amount of error. Quite substantial errors, that have only been identified at a later date[17], have been found in some published protein structures. It has also been found that different areas of a crystal structure can have quite different amounts of error. Part of the reason for this lies in the physical nature of the crystal, some parts, particularly the loop regions, may be naturally flexible and so do not lie in a single conformation. This leads to these regions diffracting poorly and so the atoms within them are placed less accurately. Disorder in the crystal is measured by the temperature factor associated with each atom that specifies the positional variability of that atom. It has also been found that some important parts of protein structures, such as ligands bound within an enzyme, are prone to error, because the geometrical restraints used in refining these regions are of a lower quality compared to those used to refine the protein itself[18].

Given the errors that all crystal structures are subject to, methods for validating protein structures are of great importance. There are two types of validation: val-

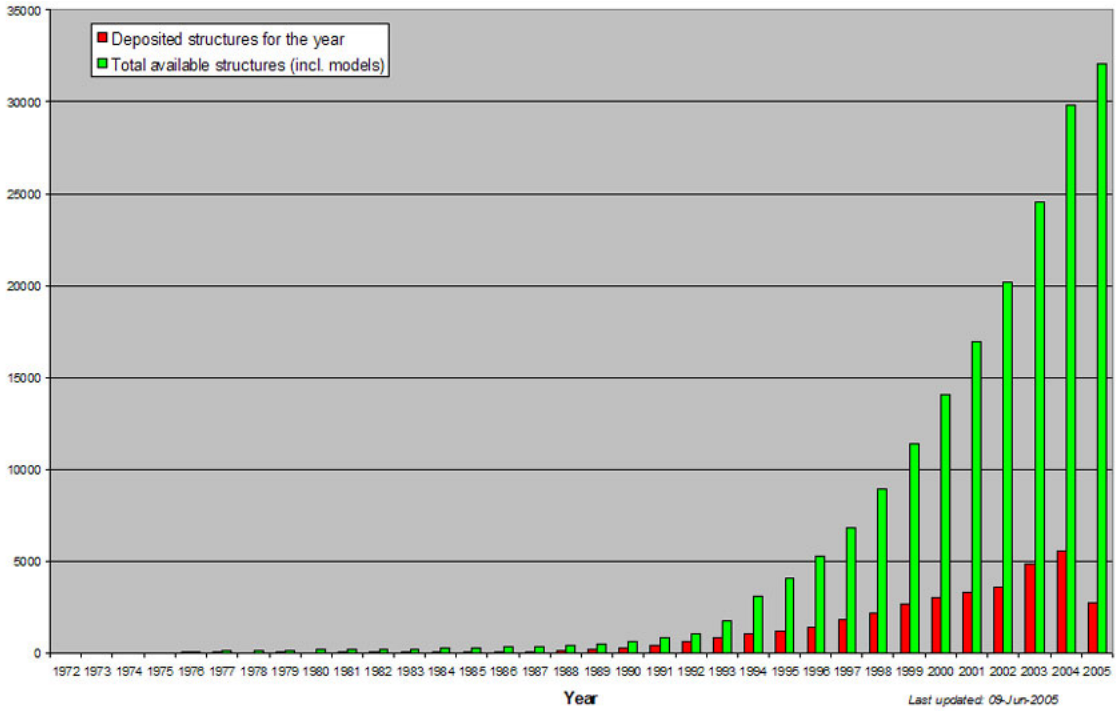


Figure 1.8: The growth of the PDB.

idating the model given the experimental data collected, and validating the model against 'typical' structures that have already been solved. A full discussion of these techniques is beyond the scope of this introduction, but both types of validation are reviewed elsewhere[19].

The Protein Data Bank (PDB)[20] was set up in 1971 to provide a central repository of solved protein structures. At this time, structure determination by protein crystallography was an extremely time consuming process, and only a few structures were solved each year. However, since 1990 the rate of structure determination has steadily increased, as shown in Figure 1.8. Over 5000 new structures were deposited in 2004 and, as of June 2005, the PDB contained over 30,000 entries, though many of these structures are different forms of the same protein.

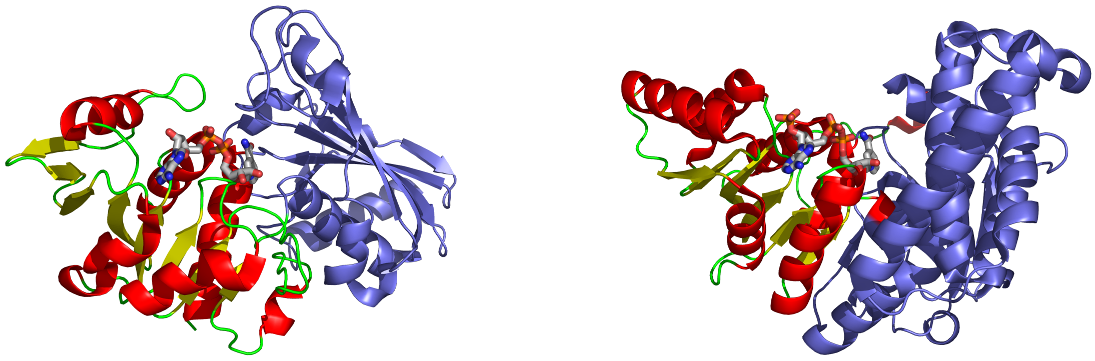
1.2.2 Domains

Larger proteins tend to fold into a series of smaller domains, each of which forms a self contained structural unit. These domains are often described as the units of evolution because they can often be swapped between proteins without disturbing the folding of other parts of the protein and thus novel functions can be created by novel combinations of domains within a single protein[13].

In enzymes, certain functions are often contained within a domain. For instance, the nucleotide binding Rossmann domain[21] is found combined with a diverse range of separate catalytic domains, allowing each enzyme to bind similar nucleotide cofactors such as nicotinamide adenine dinucleotide (NADH), nicotinamide adenine dinucleotide phosphate (NADPH) and flavin mono-nucleotide (FMN), but perform quite different chemistry. Figure 1.9 shows two different Rossmann domain containing enzymes: glyceraldehyde-3-phosphate dehydrogenase (GAPDH)[22] (EC: 1.2.1.12) and 1-deoxy-d-xylulose-5-phosphate reductoisomerase (DXR)[23] (EC: 1.1.1.267). Both enzymes contain the Rossmann domain with a common 3 parallel strand β sheet flanked by α helices. This sheet binds to the cofactor NAD in the case of GAPDH and NADP in the case of DXR. The remainder of the enzymes structure are completely unrelated, and contain quite different catalytic residues which allow them to catalyse their different reactions.

1.2.3 Active Sites and Clefts

Although enzymes are often large molecules comprising many hundreds of amino acids, the functional regions of an enzyme are generally restricted to clefts on the surface that comprise only a small part of the enzyme's overall volume. The most important of these regions is the active site - the pocket or cleft in which the enzyme



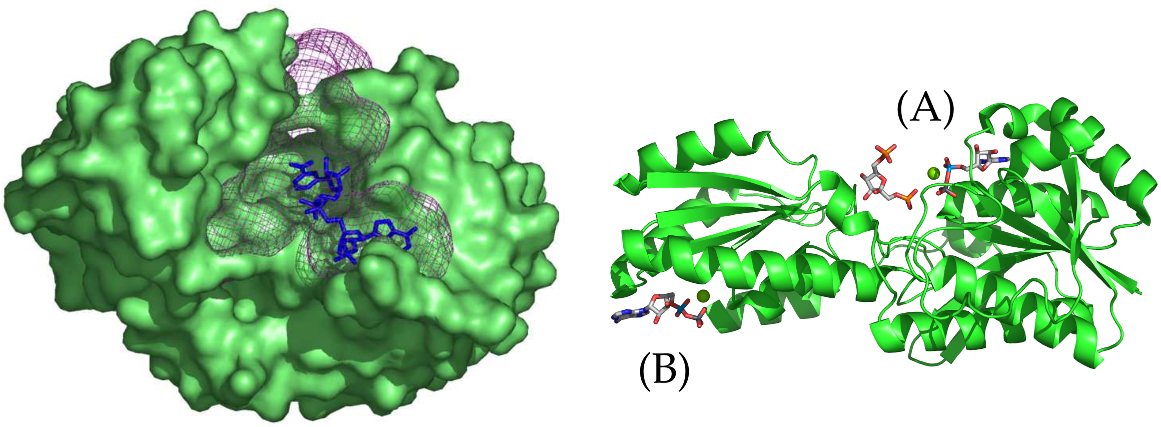
(a) GAPDH

(b) DXR

Figure 1.9: The common Rossmann domains of GAPDH and DXR are shown in red helices and yellow sheets. The NAD(P) cofactors are shown in sticks. The remainder of each protein, which is unrelated, is coloured blue.

binds the substrate and in which the catalytic chemistry of the enzyme is performed. Analysis of enzyme structures have shown that active sites tend to be formed from the largest cleft on the surface of the protein[24]. Figure 1.10(a) shows the surface of tetrahydrofolate dehydrogenase with an NADP molecule bound in the active site cleft[25].

Other clefts in the enzyme can be responsible for binding regulatory molecules. Phosphofructokinase, shown in Figure 1.10(b), catalyses the phosphorylation of D-fructose 6-phosphate, converting ATP into ADP in the process. It is regulated by binding of ATP to an allosteric site, quite distinct from the active site, that inhibits the enzyme[13]. These regulatory clefts as well as being able to bind regulatory molecules, also require the ability to transmit binding information from themselves to the active site, so that catalytic activity can be regulated.



(a) Tetrahydrofolate dehydrogenase with the largest cleft marked and NADP

(b) Phosphofructokinase with the active site (A) and allosteric site (B) marked. Substrate and inhibitor molecules are also shown.

Figure 1.10: The active site and other functional sites are usually formed from clefts in the enzyme surface.

1.2.4 Determining the Function of Enzymes

To fully appreciate an enzyme's role in biology, as well as the structure, it is also necessary to understand its chemistry, kinetics and thermodynamics. The chemistry of an enzyme is largely defined by those protein residues involved in the catalytic mechanism, and can be studied by a number of techniques. Chemical labelling[26] can identify catalytic residues by forming covalent attachments to those residues directly involved in the mechanism of the enzyme. The catalytic machinery of the serine proteases was elucidated by the use of inhibitors such as diisopropylfluorophosphate (DFP) and tosyl-L-phenylalanine chloromethyl ketone (TPCK), which form covalent links to serine and histidine residues respectively, that are involved in the mechanism[27]. pH rate profiles and NMR experiments can also suggest the involvement of different chemical groups in the active site, by tracking protonation changes and subsequent changes in the enzymes activity.

Another important experimental technique for determining the catalytic residues of an enzyme, is site directed mutagenesis. When searching for catalytic residues the alanine scanning approach can be used, whereby each residue is replaced by alanine and the effects on the reaction is measured[28]. It has been noted however, that in many cases, the removal of an 'essential' catalytic residue does not completely abolish catalysis, but rather causes the enzyme to use an alternative (and usually slower) mechanism[29]. More in-depth understanding can come from considered mutations of certain residues. In studies of the protein tyrosine phosphatases (PTPs) for instance, the replacement of an active site arginine with a lysine was found to reduce k_{cat} 8200-fold, whilst leaving K_M unchanged. This implies that the change has disrupted catalysis, without disturbing the substrate binding ability of the enzyme. Since lysine, like arginine, provides a positive charge to the active site, but does so using an amino group rather than a guanidinium group, it was deduced that the guanidinium group of the arginine must make specific contacts to the transition state which aided catalysis[30].

Even without mutagenesis or labelling experiments, measuring the kinetic parameters K_M and k_{cat} can reveal aspects of the enzyme mechanism, including those residues which are important for catalysis, binding or both. Measuring these parameters requires methods for determining the rate of formation of products or depletion of substrates; such as spectrophotometry, spectrofluorimetry and radioactivity assays. These techniques can then be used to determine the number and sequence of intermediate processes in an enzyme mechanism. An example of this is the detection of a burst of product release observed in the catalysis of the hydrolysis of *p*-nitrophenyl acetate by chymotrypsin. This burst is due to the quick formation of an intermediate (during which time *p*-nitrophenol is released), which then breaks down relatively slowly. During the burst, the enzyme molecules in the reaction are saturated with the intermediate and subsequently turnover much more slowly[13].

1.2.5 Enzyme Evolution

Enzymes, like any biological system, are under evolutionary pressures which cause them to evolve over time. There are a number of different ways in which enzyme function can evolve:

1. Evolution of a new catalytic function: The evolution of a completely novel catalytic function is perhaps the most dramatic change in enzyme function that is possible, though paradoxically it may not require a large number of individual mutations. This is because of the nature of the active site, where only a few residues are directly involved in catalysis.

2. Evolution of substrate specificity: While keeping the same basic catalytic activity an enzyme may adapt to act upon different substrates. Serine proteases are an example of this, where the catalytic triad is used to hydrolyse many different peptides. In this case, the catalytic residues remain the same whilst the binding residues mutate.

3. Evolution of stability: Extremophile species require enzymes that are stable at extremes of temperature or pH. This can be achieved by mutations in areas other than the active site, so that the scaffold on which the active site is built is maintained.

4. Evolution of rate: Some enzyme have evolved to perform catalysis as fast as is physically possible[31]. However, not all enzymes need to perform catalysis at the highest possible rate and it may be that in many cases the evolutionary

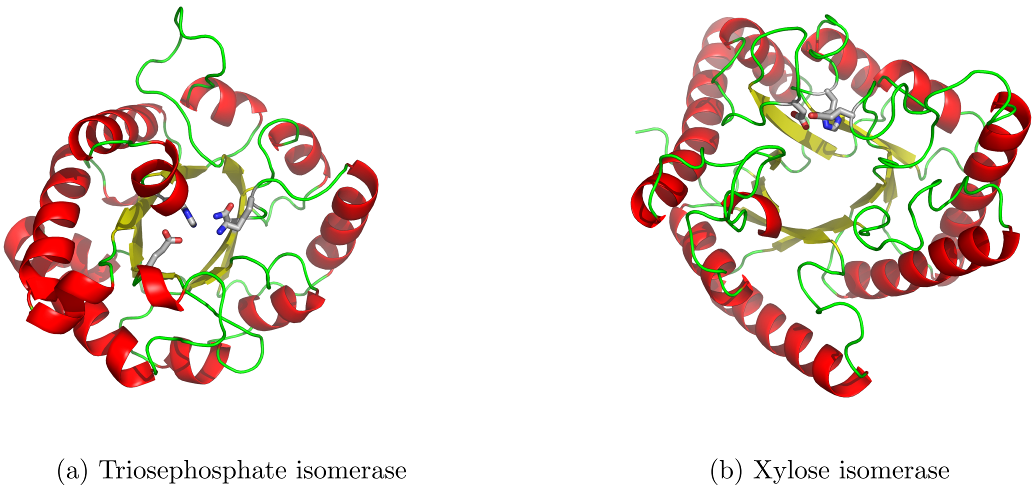
pressure on the enzyme, to increase its rate, abates before the enzyme reaches the fastest rates possible.

5. Evolution of regulatory features: Enzyme activity must be regulated, so that the metabolism of an organism can adapt to changing surroundings. There are a host of regulatory mechanisms such as post-translational modification and allosteric control that modulate the rate at which an enzyme performs catalysis.

6. Evolutionary drift: Even without evolutionary pressure, random mutations will cause a slow divergence between two enzymes. Evolution may maintain the structure of the crucial functional regions in these cases, whilst the remainder of the protein changes.

There are a number of different mechanisms by which these functions can evolve, varying from simple point mutations, to gene duplication and fusion events which may combine whole domains in novel configurations. These evolutionary events usually lead to divergent evolution whereby enzymes with a common ancestor gradually differentiate in their functions. Some families of enzymes are particularly 'functionally promiscuous' with members catalysing a wide range of different reactions[32, 33].

In a study of 31 structural superfamilies, Todd *et al*[34] found they performed almost 200 different functions. The types of changes which these families undergo were also found to vary. The non-heme di-iron carboxylate proteins, for instance, tended to undergo large scale changes: domain enlargements, domain re-organisations and oligomerisation changes; whilst, the α/β barrel proteins, of which triose phosphate isomerase (TIM) is the canonical example, varied only through the catalytic residues in the active site, while the scaffold of protein structure around the active site remained constant. Figure 1.11 shows an example of this with the TIM barrel enzymes


(a) Triosephosphate isomerase

(b) Xylose isomerase

Figure 1.11: The TIM barrels of triosephosphate isomerase and xylose isomerase. The catalytic residues shown in sticks are different in each case but similarly placed at one end of the barrel.

triose phosphate isomerase (5.3.1.1) and xylose isomerase (EC: 5.3.1.5).

The opposite of divergent evolution, which leads to enzymes with evolutionary relationships performing different functions, is convergent evolution, whereby unrelated enzymes converge on the same solution to a catalytic ‘problem’. The best known example of this being the Ser-His-Asp catalytic triad, which appears to have separately evolved on a number of occasions.

1.2.6 Structural and Functional Enzyme Classification

The wealth of structural information that emerged in the 1990’s has led to the need for schemes for the classification of protein structures. Although these schemes are designed to be used for all protein structures, they have proved extremely useful for classifying enzymes, and tracing the evolution of different enzyme functions.

Both manual and automatic methods for structural classification have been developed. Automatic schemes, such as FSSP[35] and SSM[36] have the advantage of easily staying up to date and having a clearly defined methodology. However, those systems with at least some manual curation still add value by identifying groups of structures, and distinctions between groups, that are missed by the automatic schemes.

The CATH[37] and SCOP[38] classifications are both hierarchical systems that require some manual intervention. The highest level (first digit) of the CATH classification describes the class of the protein according to the secondary structure content: mainly alpha, mainly beta, mixed alpha/beta and no secondary structure. The next level, the architecture, describes the shape of the domain as determined by the orientation of the secondary structures, but not the connectivity. The next level: topology, describes the connectivity of the secondary structures. The SSAP structural comparison program[39] is used to find structural similarities at this level. The final homologous superfamily level groups proteins known to have a common ancestor, using both structural and sequence comparisons. Some example structures from the first three levels of CATH are shown in Figure 1.12.

SCOP uses a similar classification scheme with a top level class division according to secondary structure content, a fold level which describes structural similarities without implying homology, corresponding to the architecture and topology levels of CATH, and a superfamily level which groups related proteins together.

In addition to the structural classifications which, at the lower levels, reflect evolutionary relationships, there is also a need for a classification that describes enzyme function, since enzymes with the same common ancestor can evolve to catalyse quite different reactions. The Enzyme Commission (EC) classification and the ENZYME database[40] describe an enzyme reaction using a four level classification. The first level describes the general class of the reaction: oxidoreductase, transferase, hydrolase, lyase, isomerase or ligase. The second and third levels then describe different

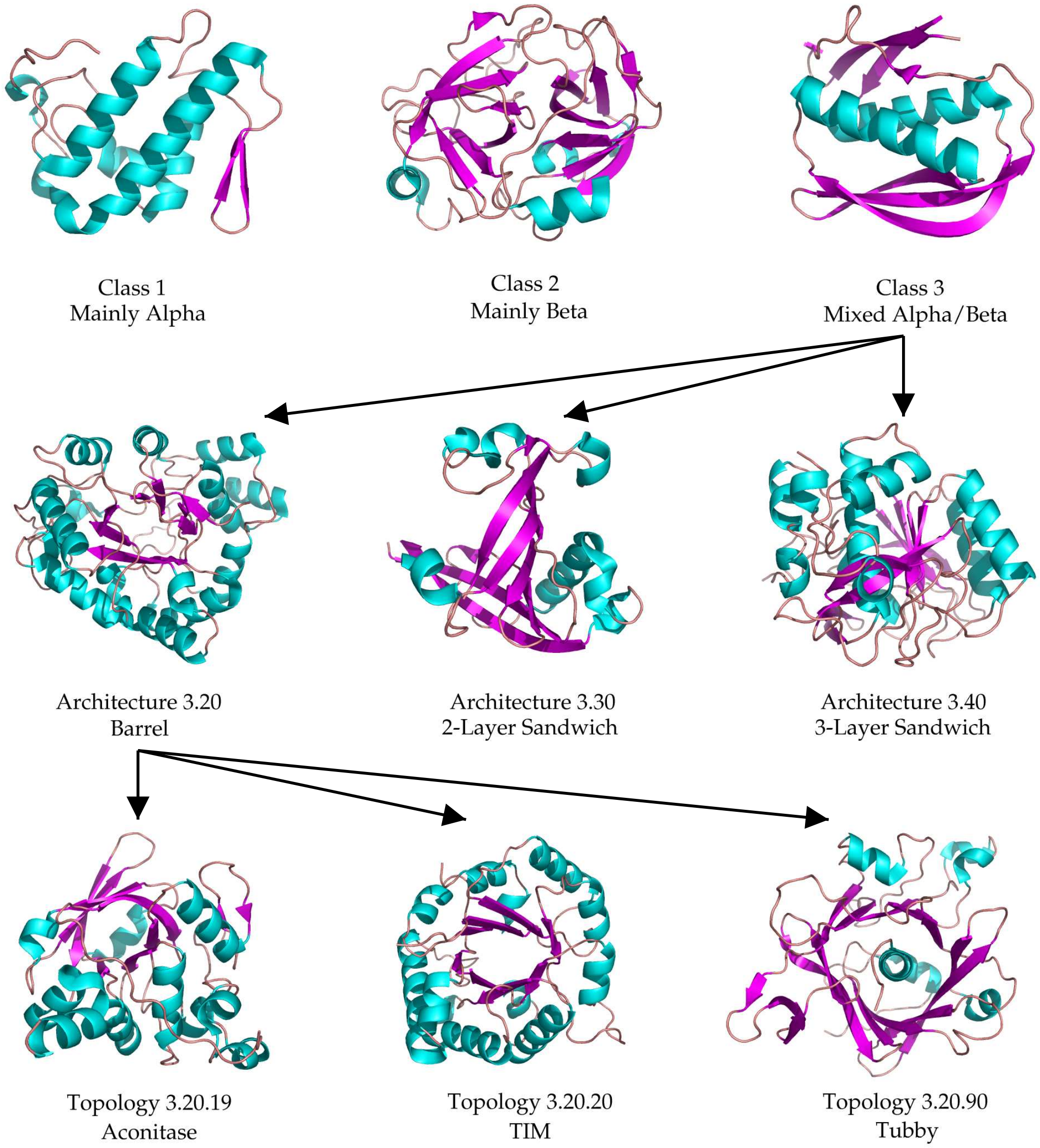


Figure 1.12: The top level shows examples of the three main classes of structures in CATH (class 4 has few secondary structures). The middle level shows three different architectures within the mixed alpha-beta class. The bottom level shows three different topologies within the barrel architecture. Each topology is then broken down into homologous superfamilies that represent the final structure based level of CATH.

properties of the reaction depending on the top level class. The fourth level generally describes the substrate specificity of the enzyme, so that each unique EC number describes a single enzyme catalysed reaction.

The difficulty in using the EC classification for studying enzymes, is that it does not specify the direction of the reaction catalysed or the mechanism that the enzyme uses to catalyse the reaction. This means that just as two related enzymes in the same CATH superfamily can catalyse quite different reactions, it is also possible for two enzymes to share the same EC code without being related or even using a similar reaction mechanism. Some of the beta-lactamases (EC: 3.5.2.6) for instance, that are responsible for the hydrolysis of beta-lactamase antibiotics such as penicillin, use a zinc dependent mechanism, whilst others use a serine hydrolase type mechanism[41, 42].

1.3 Serine Proteases

The serine proteases are one of the most well studied enzyme families, and are repeatedly invoked as archetypal examples of enzyme mechanisms. In this section we briefly summarise the features of these important enzymes.

1.3.1 Evolution and Families

Serine proteases catalyse the hydrolysis of peptides using a nucleophilic serine residue at the active site. Of the 10 different evolutionary clans, shown in Table 1.1, 5 use a serine-histidine-carboxylate triad, and it is these enzymes that we focus on. The fact that the Ser-His-Asp/Glu catalytic triad has evolved on at least five separate occasions for the hydrolysis of peptides seems to imply that this is an extremely effective catalytic mechanism for peptide bond cleavage.

Clan	Members	Example	Catalytic Residues
PA(S)	301	chymotrypsin	His-Asp-Ser
SB	91	subtilisin	Asp-His-Ser
SC	64	carboxypeptidase Y	Ser-Asp-His
SK	14	Clp protease	Ser-His-Asp
SN	4	dipeptidase	Ser-His-Glu
SE	16	D-Ala-D-Ala carboxypeptidase A	Ser-Lys
SF	24	signal peptidase I	Ser-Lys/His
SH	7	cytomegalovirus assemblin	His-Ser-His
SM	7	C-terminal processing protease	Ser-Lys
PB(S)	4	penicillin amidohydrolase precursor	N-terminal ser

Table 1.1: The 10 different evolutionary clans of serine proteases and the catalytic residues used in each case. Table taken from Hedstrom *et al*[27].

1.3.2 Structural Studies

The three dimensional structure of bovine α-chymotrypsin was solved to 2Å resolution by David Blow *et al*[43] in 1967, the second enzyme after lysozyme to have its structure solved. As with lysozyme, these studies revealed important details about the mechanism of the protein, in particular the triad of residues Ser 195, His 57 and Asp 102 were identified as crucial, along with two main chain amides that form an oxyanion hole stabilising the transition state. In 1969, Joseph Kraut solved the structure of the unrelated bacterial subtilisin enzyme and found that the same catalytic triad was found in an almost identical conformation, though the two structures showed no other structural similarities[44]. Figure 1.13 shows the overall fold of trypsin and subtilisin and the catalytic triad in each case.

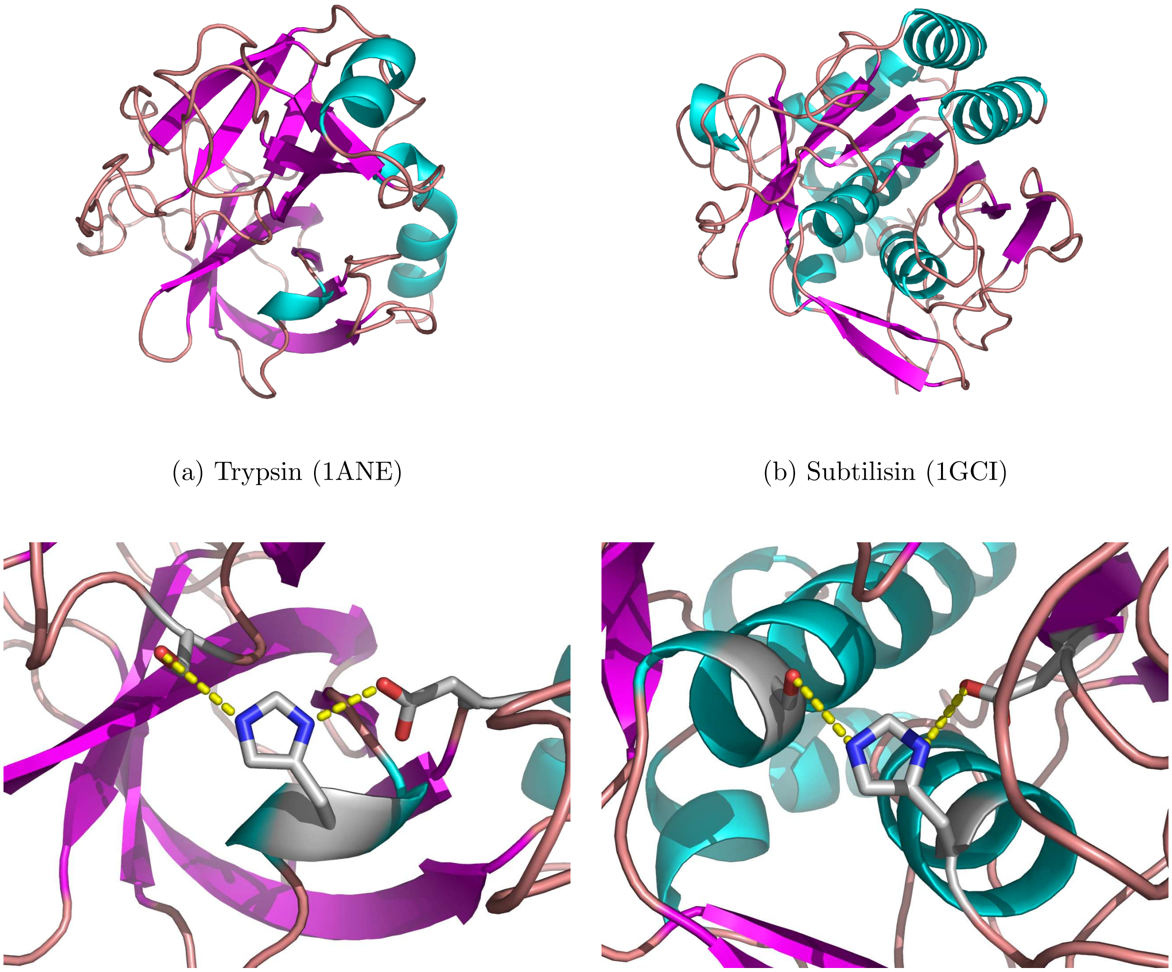
Even with this structural information, there was originally some dispute as to the precise roles of the residues in the catalytic triad. In particular, the protonation state of the histidine and aspartate residues was unclear. Protons only weakly diffract X-rays so very high resolution studies are required to locate them. In fact neutron diffraction was used to show the protonation state of the key catalytic residues[45], which revealed that the Nδ atom of the histidine remains protonated throughout the reaction, and does not transfer the proton to the aspartate. Some controversy over the mechanism still remains, with the role of a potential ‘low-barrier’ hydrogen bond between the aspartate and histidine of particular interest[46, 47].

1.3.3 Mechanism

The mechanism of serine proteases encapsulates much of the theory detailed above, using as it does electrostatic, acid/base and nucleophilic catalysis as well as transition state stabilisation.

Each step in the mechanism is shown in Figure 1.14. In the first step (a) (once a peptide has bound) the oxygen of the serine hydroxyl attacks the carbonyl carbon of the substrate with concomitant removal of the hydroxyl proton by histidine. The protonated form of the histidine is stabilised by its interaction with aspartate. The transition state that leads to the intermediate shown in (b) is also stabilised by interaction with the main chain amide protons that form the oxyanion hole. The tetrahedral intermediate (b) is unstable because of the negative charge on the carbonyl oxygen, and breaks down by expelling the amine which is protonated by the histidine.

The acyl enzyme linked intermediate (c) requires essentially the same reaction to break it down. Instead of serine, water is used as the nucleophile, again activated by the histidine acting as a base. Again a tetrahedral transition state and subsequent intermediate (d) is stabilised using the main chain amides, and the breakdown (e) is assisted by the donation of the proton from the histidine back to the serine which breaks the bond connecting it to the peptide.



(a) Trypsin (1ANE)

(b) Subtilisin (1GCI)

(c) Trypsin catalytic triad

(d) Subtilisin catalytic triad

Figure 1.13: Subtilisin and trypsin are unrelated, and have totally different global structures. However, the conformation of the catalytic triad is the same in each case.

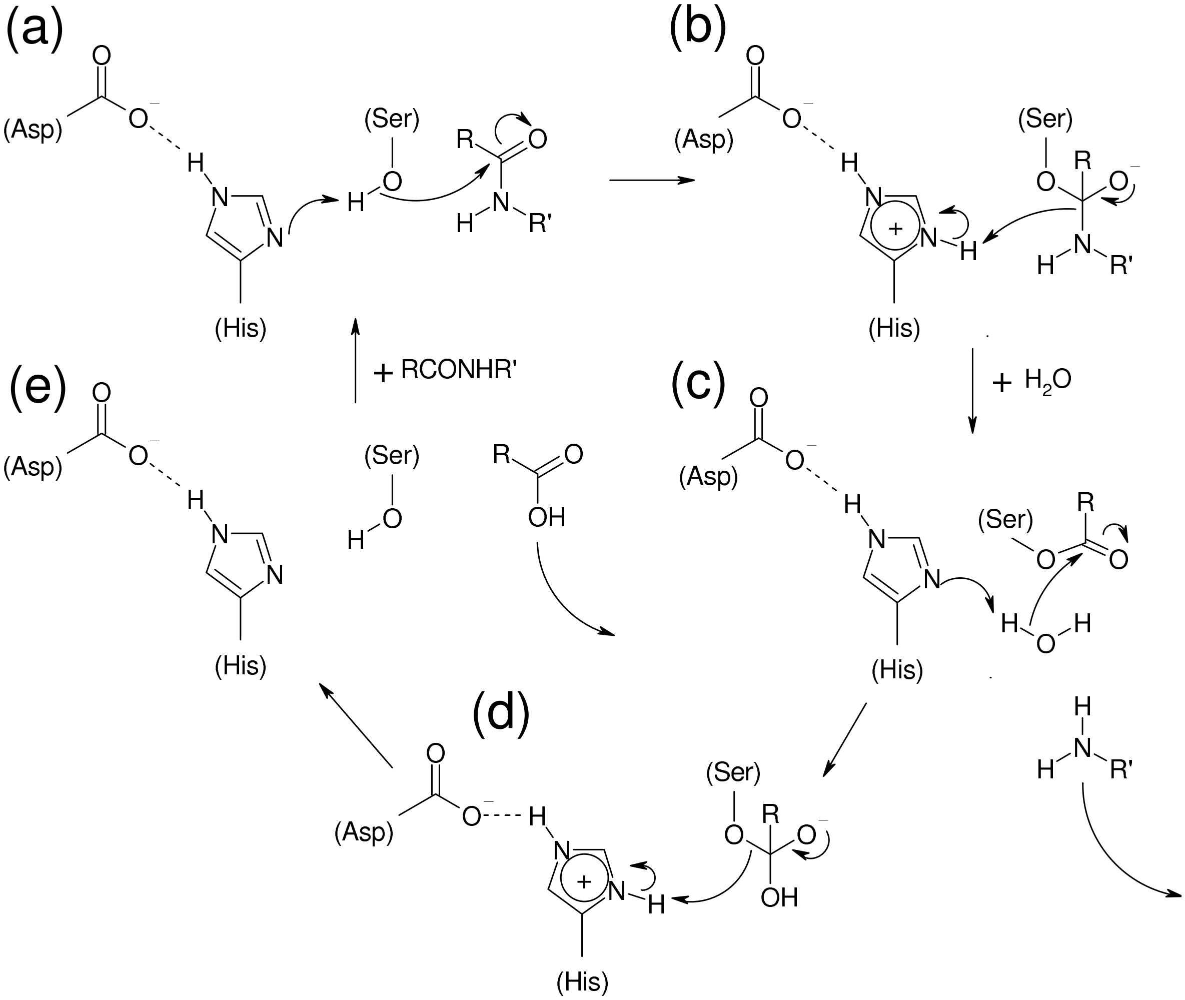


Figure 1.14: The mechanism of the serine proteases. The amide groups that form the oxyanion hole are not shown.

1.4 Structural Genomics and Predicting Enzyme Function

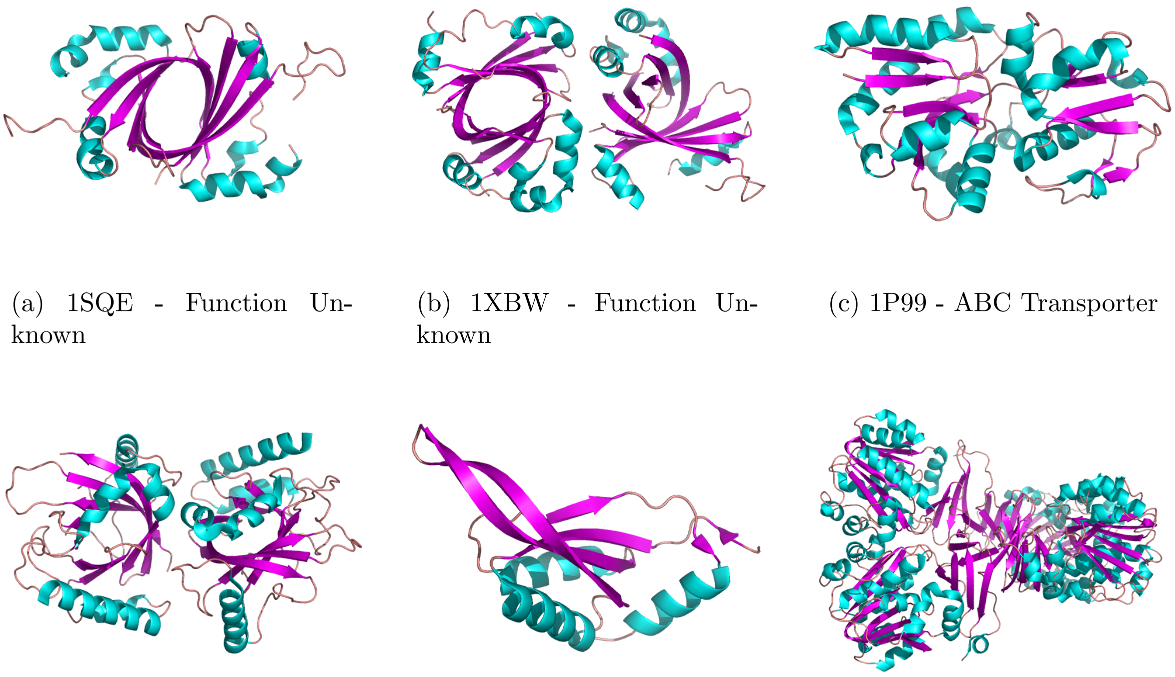
In recent years there have been a number of 'structural genomics' efforts aiming to automate structure solving methods (principally crystallography) in order to rapidly increase the number of known protein structures[48, 49]. The goals of each project differs: for instance, some projects aim to solve the structures of medically important human and pathogen proteins whilst others focus on solving structures that are predicted to have new folds.

The effect of these projects is to change the way protein (including enzyme) function is studied. Previously, because of the difficulty of determining X-ray structures, extensive biochemical characterisation of a protein would usually be performed before any structural studies, and so the function of the protein would at least be broadly known when the structure was revealed. With structural genomics, the structure of proteins that are relatively unstudied is revealed. The problem then becomes how can we use the structure to reveal information about the function of a protein?

We have already seen how a structure with bound ligands can provide important clues to an enzymes function, but this usually requires some prior knowledge of the function, so that the ligand can be placed in the crystallisation medium. Techniques for predicting function purely from structure are of increasing interest therefore, and we summarise some here.

1.4.1 Predicting Function From Sequence

Prior to structure determination, the function of an unannotated protein is usually derived from sequence analysis. The commonest method is usually to find homolo-



(a) 1SQE - Function Unknown

(b) 1XBW - Function Unknown

(c) 1P99 - ABC Transporter

(d) 1NG5 - Transpeptidase (new fold)

(e) 1KR4 - Cation tolerance protein

(f) 1INL - Spermidine Synthase (new fold)

Figure 1.15: Some example structures from the Midwest Center for Structural Genomics (MCSG) project. Those structures with a novel, previously unseen, fold are marked.

gous proteins using sequence database searching methods such as PSI-BLAST[50]. If a significant sequence similarity is found between the unannotated protein and an annotated protein then the annotation can be transferred from the annotated protein to the unannotated. However, several studies have shown the difficulty of using these kinds of methods, and there are a number of examples of proteins with high sequence similarity, but quite different functions[51, 52, 53].

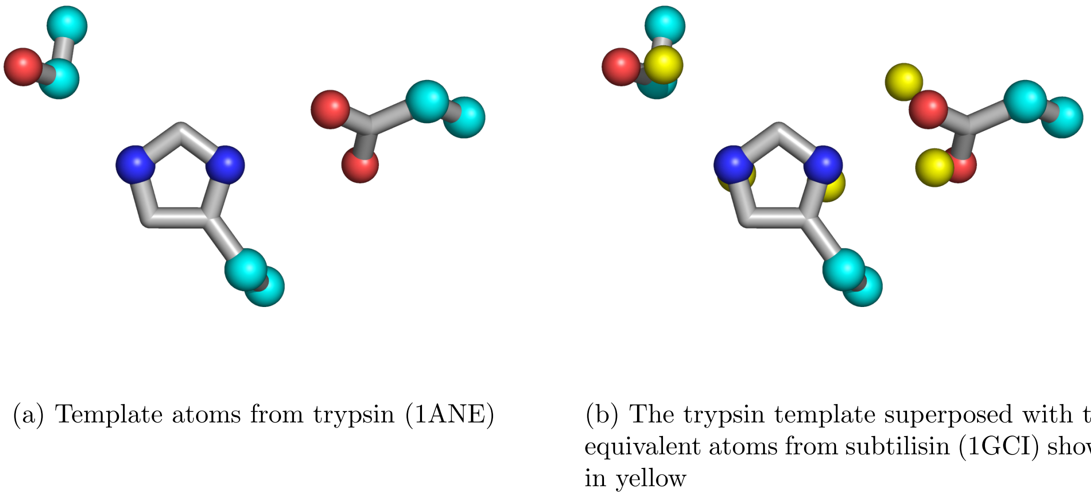
Alternative sequence searching methods include pattern matching such as Prosite[54], and domain matching such as Pfam[55]. These techniques can find more distant evolutionary relationships, but this makes transferring functions even more difficult, since proteins that have diverged so much that their relationship is only detectable at this level are more likely to have also diverged functionally.

1.4.2 Predicting Function From Structure

3D Pattern Matching

As we have seen, totally unrelated enzymes such as trypsin and subtilisin, can have very similar active site architectures and mechanisms. It is also possible for a pair of homologous enzymes to be so far diverged that their relationship is undetectable by conventional means, leaving the functional regions of the proteins the only conserved areas. 3D templates aim to exploit both these ideas to find similarities between enzymes that are otherwise unrelated, or so distantly related that similarities are extremely difficult to detect through other methods[56].

Templates are formed by selecting a few functional residues or atoms whose relative 3D positions define the template. A template made from the active site of trypsin is shown in Figure 1.16(a). Methods such as Jess[57], TESS[56], SPASM[58] and FFF[59] are then used to search for similar occurrences of the template in other


(a) Template atoms from trypsin (1ANE)

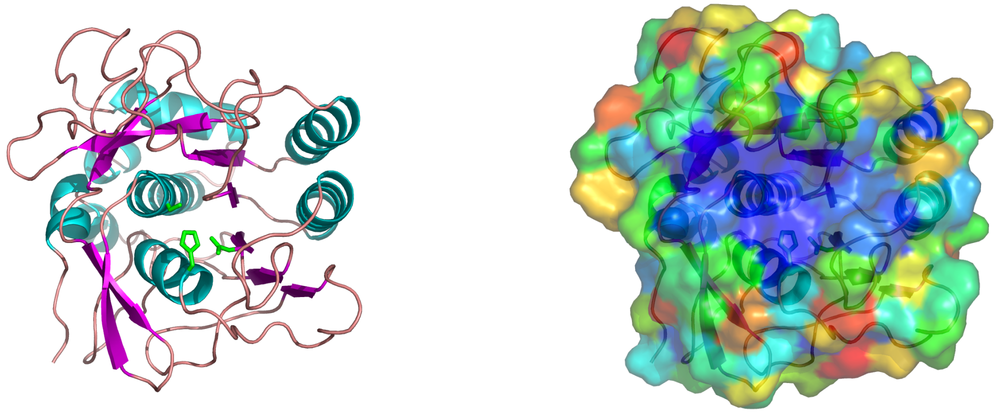
(b) The trypsin template superposed with the equivalent atoms from subtilisin (1GCI) shown in yellow

Figure 1.16: The use of templates to find and annotate enzyme structures. The RMSD is 1.036Å between the template from trypsin (1ANE) and the matched atoms in subtilisin (1GCI). Functional atoms are shown in blue and red for nitrogen and oxygen respectively, Cα and Cβ atoms are shown in cyan. The matched atoms from subtilisin are shown in yellow.

structures. We would expect the trypsin template to find a match in subtilisin for instance, whose active site atoms are shown overlaid with the template in Figure 1.16(b). A match is often scored by measuring the root mean squared deviation (RMSD) between the template and the matched atoms. Templates developed from the active sites of a large selection of enzymes have been tested to try and find examples of convergent or highly divergent enzymes[60]. Template creation is usually done by hand, though automatic methods such as SiteSeer[61] have also been developed.

Conservation Approaches

As a pair of enzymes diverge, assuming their functions remain the same, the functional regions, such as the catalytic residues and binding sites, will evolve at the



(a) The active site of subtilisin

(b) The surface of subtilisin coloured by conservation score. Blue represents highly conserved regions and red/yellow highly variable regions

Figure 1.17: Mapping the conservation scores produced by ConSurf onto the structure of subtilisin. The central conserved patch is the active site and peptide binding pockets.

slowest rate, since it is assumed that mutations here are more likely to be detrimental to the enzyme's function. This assumption has been successfully used to predict the location of functional regions in proteins of unknown function. Most such methods begin by performing sequence searches to find homologous proteins which are then aligned in a multiple alignment. The multiple alignment is then analysed to find those residues which are especially conserved. Some methods, such as ConSurf[62] and Evolutionary Trace[63, 64], also create phylogenetic trees from the alignments. However it is derived, the conservation score is then mapped onto the known structure and clustering techniques are used to try and locate regions within the protein that are especially conserved and hence likely to be functional.

Figure 1.17 shows the conservation scores from the ConSurf program mapped to the structure of subtilisin. The central cleft which contains the catalytic residues and the peptide binding residues is clearly identifiable by the high conservation scores.

Ab initio Methods

Conservation methods rely on having a large number of close, but non-identical homologues to the structure of interest. Without these sequences, the multiple alignments generated do not contain enough information to make reliable predictions. In these cases, purely structural approaches to finding functional sites can be used. Techniques such as THEMATICS[65], or that developed by Elcock[66] use *in silico* calculations to find residues with unusual ionisation or electrostatic properties. These residues are predicted to be functional because the catalytic mechanism of many enzymes rely on residues with unusual electrostatic properties or ionisation states. Testing these methods, and the conservation based methods described above, have been problematic, as there are few large, reliable datasets which define the position of functional sites and residues, against which the methods can be tested.

The ProFunc server[67] aims to combine many of these different analyses into a simple interface, without addressing the difficult question of making a consensus prediction itself. Instead, the human expert operator is left to make their own conclusions based on the results presented. ProFunc uses sequence searching methods such as PSI-BLAST and Pfam as well as global structure searching methods such as SSM. These methods suggest potential evolutionary relationships, whilst conservation mapping and structure analyses such as cleft analysis, DNA binding motifs searching, nest analysis and template searching are used to suggest functional sites.

1.5 Learning the Principles of Catalysis and Unresolved Issues

Given that so much is known about both the chemistry performed by enzymes and the structure of enzymes and their active sites, what questions remain to be answered about enzyme catalysis and the structural basis of catalysis?

There are ongoing efforts to explore the effects of quantum mechanics[68] and dynamics[69] on enzyme catalysis. However, in addition to these detailed studies, which usually concentrate on a single enzyme or system, there is still a need for large scale analyses of enzyme structures and mechanisms. There have been many studies that have looked at large sets of structures to gain new insights into protein folding[70, 71, 72, 73, 74]. There is now sufficient information for a similar approach to studying enzymes and enzyme catalysis.

1.5.1 The Catalytic Site Atlas and MACiE

To perform a large scale analysis of enzyme catalysis, one requires a large dataset of enzymes with known structure that are well annotated. Although the PDB provides all the structural information required, there is little annotation as to the location of the active site and the identity of the catalytic residues. The Catalytic Site Atlas (CSA)[75] is a database which aims to add this annotation to enzyme structures in the PDB.

The first important step of the CSA is to develop a fixed definition of what comprises a catalytic residue. Other potential sources of annotation, such as the SITE record in the PDB file format, do not specify any particular definition for what can be placed in them, so they are not suitable for a rigorous analysis. The CSA defines a catalytic residue as a residue that fulfils any of the following tests:

1. Direct involvement in the catalytic mechanism - e.g. as a nucleophile.
2. Exerting an effect on another residue or water molecule, which is directly involved in the catalytic mechanism, which aids catalysis (e.g. by electrostatic or acid-base action).
3. Stabilisation of a proposed transition-state intermediate.
4. Exerting an effect on a substrate or cofactor which aids catalysis, e.g. by polarising a bond which is to be broken. Includes steric and electrostatic effects.

Note that residues that bind substrate, cofactor or metal ions are not included, unless they also perform one of the functions listed above.

The first analysis of the CSA focused on an original core dataset of 178 enzymes and examined the propensity of different residues to be used in catalysis and measured some general structural parameters for catalytic residues[76]. As expected, certain residues, such as histidine, were found to play catalytic roles far more often than others, and catalytic residues were also found to be highly conserved and slightly buried within the enzyme structure. Subsequent work examined the way in which related pairs of enzymes and enzyme/nonenzymes evolved to catalyse different reactions and perform different functions[77]. By looking in detail at reaction mechanisms, this work also showed that different enzymes catalysing quite different overall reactions can share mechanistic similarities.

The MACiE database complements the CSA by providing a chemistry orientated approach and is particularly geared towards studying the evolution and use of different mechanistic tools in catalysis. However, the information required to annotate enzymes with fully detailed mechanisms is considerably more than that required to simply annotate the catalytic residues and the information itself is harder to deal with computationally. MACiE now comprises over 150 different enzyme mechanisms and is beginning to provide an insight into how enzymes perform different chemical functions.

1.6 Outline of Thesis

The goal of this thesis is to understand how enzymes perform catalysis by an analysis of the structural properties of enzyme structure, and in particular the active site. The reason for studying this is two-fold: firstly catalysis is a vital part of cellular metabolism and so impacts on all aspects of molecular biology, secondly, there are practical efforts to see whether, on the basis of this understanding, we can firstly predict functions in enzymes of unknown structure and secondly help in efforts to design enzymes with novel functions[78].

In Chapter 2, we use a neural network trained using the initial core dataset of the CSA, to try and predict the location of active sites based on the structural and sequence parameters we derive for each residue.

The next task was to enlarge the set of enzymes for which annotation was available and Chapter 3 describes the ongoing efforts to achieve this. In subsequent chapters we use this enlarged dataset to study various structural aspects of enzyme catalysis.

The first part of an enzymes action is the binding of substrates. The induced fit theory suggests that enzymes undergo changes in structure upon substrate binding. However, despite numerous known examples of this, the extent to which this is commonly observed is not clear. This question is of importance in understanding the processes by which enzymes bind substrates and also, in practical terms, the importance of distinguishing between substrate bound and apo structures in subsequent analyses. In Chapter 4 we set out to measure the conformational changes undergone by a large set of enzymes both upon substrate binding and during the remainder of the catalytic cycle.

The next step in enzyme catalysis, after substrate binding, is catalysis itself. Enzymes use a range of different mechanisms, though the number of different chemical tools used is relatively small. In Chapter 5 we aim to show how different combinations of residues are used to provide different functions used in catalysis. One of the most important catalytic tools is the imidazole ring of histidine which can be used in the acid/base chemistry so important to enzymes. In Chapter 6 we specifically look at the way in which different residues are used to prime the imidazole for action and the specific geometries which are preferred for these interactions.

As well as the specific chemical tools investigated in Chapters 5 and 6; in chapter 7, we look at a new hypothesis, that suggests that entropy-enthalpy compensation is important in enzyme catalysis. The hypothesis predicts that entropy-enthalpy compensation leads to a significant shrinkage of the weak, polar interactions that are found throughout enzyme structures. We examine these interactions in order to find evidence to test this theory.

Chapter 8 provides a short discussion of the results from this work and the implications for future studies of enzymes and catalysis.

Chapter 2

Using a Neural Network to Predict the Location of Active Sites

2.1 Introduction

The huge increase in the rate of DNA sequencing, and the use of gene prediction technologies, such as Genscan [79] and Genewise [80], have flooded protein databases with new sequence data. The various Structural Genomics Initiatives (SGI), now aim to produce a similar increase in the amount of structural information [48]. One of the most important tasks in biology today is to use these data to provide functional annotation that leads to biologically useful knowledge [49].

One type of information that structural data can provide is the location and nature of the functional regions of a protein, such as protein-protein interaction sites and ligand binding pockets. Knowing the location of the functional sites within a protein allows the study of targeted mutants, structure-based drug design, and functional annotation of the protein by comparison with other characterised proteins. Traditional molecular biology techniques for finding functional sites, such as mutagenesis [28], pH dependence [81] and chemical labelling [26] are generally time consuming, and rely on some prior knowledge of the function of the protein to allow it to be assayed.

In silico methods for finding and annotating functional sites would clearly be of great help in annotating novel protein structures from structural genomics. In this chapter, we describe a new method for *de novo* prediction of functional sites specific for the active sites of enzymes. Instead of searching for clusters of conserved or electrostatically unusual residues as previous methods have done[63, 64, 82, 83, 84, 85, 66, 65], a neural network is used to score the residues of a protein structure by the likelihood that they are catalytic. By searching for clusters of high ranking residues the algorithm determines the most likely active site. The neural network is trained using a dataset of proteins for which the catalytic residues have been confidently located by experiment. Structural parameters such as the solvent accessibility, type of secondary structure, depth, and cleft that the residue lies in, as well as the conservation score and residue type are used as inputs for the neural network.

2.2 Materials and Methods

2.2.1 Protein Test Set

The test set contains 159 proteins from the core data set of the CSA, containing no homologous pairs and covering all 6 top level EC numbers. This data set contains approximately 55000 non-catalytic residues and 550 catalytic residues available for training the network.

2.2.2 Compilation of Data

The catalytic residues for each enzyme were extracted from the CSA. Many studies have used the SITE records defined in PDB files as the basis for defining functional residues and sites. Unfortunately SITE records are not a homogeneous data set, and there are no fixed rules on what may or may not be included in a SITE entry. Only 13 of the 159 PDB files in our data set contain SITE records, less than 10%. These 13 structures contain 50 catalytic residues, as defined above and 94 SITE residues. The overlap between these two groups contains 36 residues. We find therefore, that in our data set 28% of catalytic residues are not found in the SITE records and only 38% of SITE residues are catalytic.

The following parameters were derived for each residue (catalytic and non-catalytic) in all 159 proteins:

- **Conservation:** The sequence of each chain in the protein was used to initiate a PSI-BLAST search of the NCBI Non-Redundant Data Base (NRDB) with an E-value cut-off of 10^{-20} for inclusion in the next iteration. Each PSI-BLAST search was run to convergence or a maximum of 20 iterations. The final multiple alignment generated by PSI-BLAST was then scored for conservation and Diversity Of Position Score (DOPS) as described by Valdar *et al* [86].

- **Relative Solvent Accessibility (RSA):** NACCESS [87] was used with standard parameters to calculate the RSA of each residue.

- **Secondary Structure:** DSSP [88] was used to extract the secondary structure for each residue. The DSSP classification was simplified to three categories: helix, sheet or coil/other.

- **Cleft:** Surfnet [89] was used to define in which, if any, cleft the residue lay. If

$$\overbrace{0.7}^{Cons.}\ \ \overbrace{0.9}^{DOPS}\ \ \overbrace{0.3}^{Depth}\ \ \overbrace{0.15}^{RSA}\ \overbrace{0\,0\,1}^{SS}\ \overbrace{1\,0\,0\,0}^{Cleft}\ \overbrace{0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0}^{Residue\,type:Serine} \tag{2.1}$$

Figure 2.1: An example of the neural network input encoding.

a residue lay in two or more clefts only the largest was recorded.

- **Depth:** The depth of a residue within the protein structure is defined as the average minimum distance between each of its atoms and the closest solvent accessible atom in the structure. NACCESS was used to define solvent accessibility.

2.2.3 Encoding and generation of Data Sets

Conservation, as calculated above, is already encoded as a suitably scaled factor between 0 and 1 (0 for no conservation and 1 for perfect conservation) and so is passed to the network as is. The RSA is a percentage and is scaled to between 0 and 1 before presentation to the network. Depth is scaled so that the deepest residue in each structure is scored 1 and surface residues 0.

The other parameters: residue type, secondary structure and cleft are categorical in nature, and are encoded using 1-of-C encoding. Amino acid type is encoded as an array of 20 inputs where one input is set to 1 and the rest to 0. Secondary structure is encoded by three input parameters. Cleft size is divided into four categories: no cleft, largest cleft, 2nd or 3rd largest cleft and 4th to 9th largest cleft.

An example encoding is shown in Figure 2.1 for a serine residue with conservation 0.7, DOPS score 0.9, depth 0.3, RSA 15%, in a coil region and lying in the largest cleft.

2.2.4 Training the Neural Network

The neural network software used is FFNN [90], a feed forward neural network trained using a scaled conjugate gradients algorithm. A single-layer architecture is used in all cases. In order to accurately measure the performance of the network it is trained using a 10-fold cross validation experiment. The dataset is divided into 10 equal subgroups, and then in each training run 9 of the groups are used for training, whilst the network is tested on the single remaining group. The network is run 10 times using a different subgroup as the test group each time. In this study the dataset was divided by structure rather than residue, so each subgroup contains the data for approximately 16 structures. The ratio of catalytic to non-catalytic residues is approximately 1:60 in the training set. Presenting the data in this ratio causes the net to predict every residue as non-catalytic. The best balanced training set was found to have a ratio of 1:6. Each training group is balanced by discarding a random selection of the non-catalytic residues prior to training. Training was for 100 epochs, in every case the network converged to a stable error-level before training was terminated. The number of training epochs was not optimised, and in particular the performance of the test set was not used to optimise the stopping point in any way.

2.2.5 Measuring Performance

In order to judge the neural network learning process, a suitable measure of performance is required. Total error (percentage of incorrect predictions) is not sufficient because of the highly unbalanced nature of the dataset. All of the statistics are derived from the following quantities:

p = Number of correctly classified catalytic residues.

n = Number of correctly classified non-catalytic residues.

o = Number of non-catalytic residues incorrectly predicted to be catalytic (over-predictions)

u = Number of catalytic residues incorrectly predicted to be non-catalytic (under-predictions)

t = Total residues (p + n + o + u)

The total error (Q_{Total}) is given by Equation 2.2:

$$Q_{\text{Total}} = \frac{p+n}{t} \times 100 \tag{2.2}$$

To complement this, two other measures of performance are used: $Q_{\text{Predicted}}$ measures the percentage of catalytic predictions that are correct and Q_{Observed} measures the percentage of catalytic residues that are correctly predicted. The formulae for these two parameters are shown in Equations 2.3 and 2.4:

$$Q_{\text{Predicted}} = \frac{p}{p+o} \times 100 \tag{2.3}$$

$$Q_{\text{Observed}} = \frac{p}{p+u} \times 100 \tag{2.4}$$

A measure of performance that takes both these factors into account is the Matthews Correlation Coefficient (MCC). The formula for calculating MCC is shown in Equation 2.5:

$$MCC = \frac{pn - ou}{\sqrt{(p+o)(p+u)(n+o)(n+u)}} \tag{2.5}$$

2.2.6 Ranking and Clustering

The residues in each structure are ranked by network score, and all residues scoring above a cut-off value are used in the clustering algorithm. A pair of residues are clustered together if any of their atoms lies within 4Å of each other. Each cluster is then defined as a sphere with its centre at the geometric centroid of all the Cβ atoms of the component residues (Cα for glycine) and a radius such that all the Cβ atoms lie within the sphere. The first ranking cut-off was set at 35% of the highest scoring residue. If any sphere in a structure had a radius greater than 15Å, the clustering was repeated, increasing the ranking cut-off by 1% until no sphere was greater than 15Å in radius. Single residue clusters were discarded at this stage.

The definition of the known sites is the same. Spheres were defined for each active site with centres at the centroid of the Cβ atoms and radii such that all the Cβ atoms are within the sphere. Proteins with single catalytic residues were set a radius of 3Å.

2.3 Results

2.3.1 Analysis of Parameters

A detailed analysis of the parameters is provided by Bartlett *et al* [76]. A brief summary is presented here:

Conservation was the most powerful parameter for discriminating catalytic and non-catalytic residues. However, some proteins failed to find sufficiently diverse homologues to generate meaningful conservation scores. It is hoped that in these cases the predictive power of the other parameters will be enough to allow reliable predictions to still be made.

Catalytic residues show a tendency to be buried within the structure and so have a lower RSA than other residues, particularly non-catalytic polar residues, the majority of which lie exposed on the surface of the protein. Despite this tendency to be buried, catalytic residues are often found lining a large cleft. This tendency is particularly marked for the largest cleft, and is significant for the second and third largest clefts. For clefts smaller than this (4th - 9th largest) the difference is not particularly significant.

There is a slight tendency for catalytic residues to prefer coil regions over helix or sheet regions, this could be due to the extra conformational flexibility this gives them (allowing the active site to change conformation on ligand binding)

The hydrophobic and small residue groups were found to be very rarely catalytic, presumably because they do not contain the chemical groups required for most catalytic tasks. The obvious exceptions to this are when the backbone amide and carbonyl groups perform catalytic functions. It was found that glycine is the residue most often used in this case.

2.3.2 Depth

Depth values were calculated for all the residues in the data set, and the distribution (Figure 2.2) shows that almost 40% of residues lie on, or near, the surface of the protein, with depths less than 1Å. These residues are almost completely exposed to the surface with only a few of their atoms not solvent accessible. The proportion of the total represented by each 1Å division then decreases steadily, apart from a small peak in the 4-5Å division. Presumably this second peak is due to invaginations on the protein surface, which alter the distribution from the smooth decrease one would expect given a perfectly spherical protein. The very deepest residues in this data set lie at ∼13Å. Catalytic residues show a different distribution, with only 17% lying

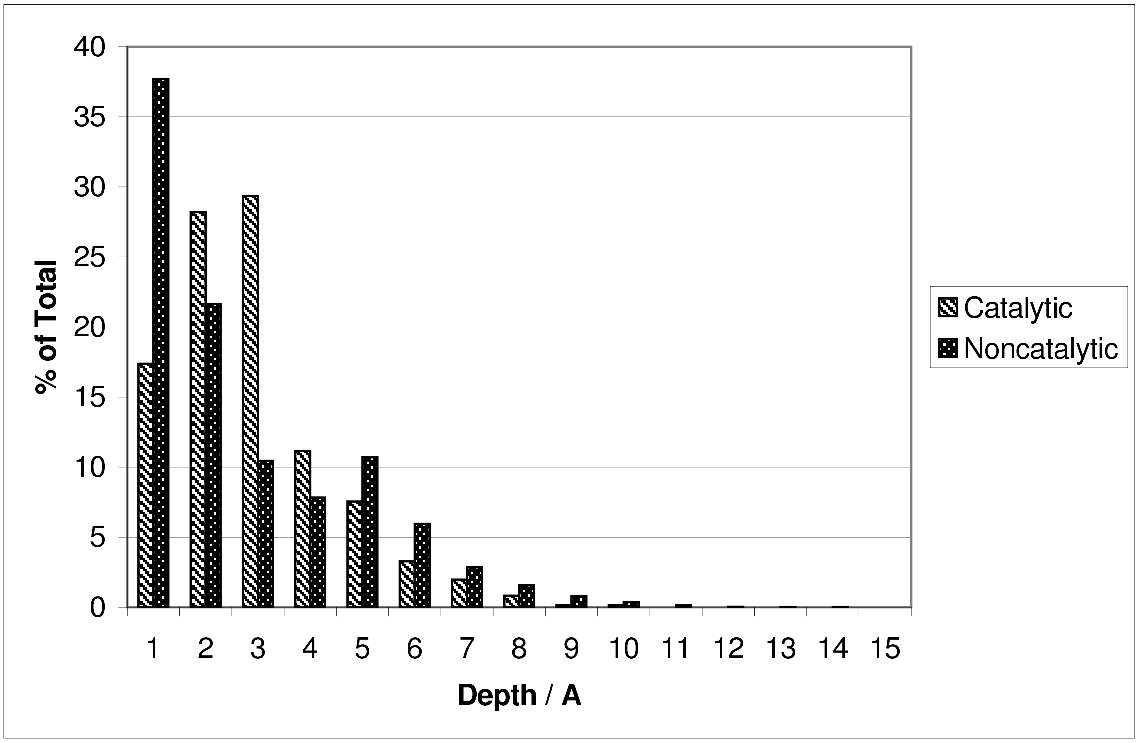


Figure 2.2: Distribution of residue depths for all residues and catalytic residues.

in the outer 1Å, the majority occupy the next partially buried layer between 2 and 4Å. This allows the catalytic residues to have some solvent accessibility (in order to interact with the substrates) whilst remaining mostly buried (to allow themselves to be correctly orientated by other residues). The catalytic residues rarely have depths greater than 5Å.

2.3.3 An Example: Quinolate Phosphoribosyltransferase

As an example of the neural network output, the scores along the 286 amino acid sequence of quinolate phosphoribosyltransferase (PDB code: 1QPR) are shown in Figure 2.3. Most residues score very low (a large majority score less than 0.01), and around 20 residues score over 0.5. The four known catalytic residues (Arg 105, Lys 140, Glu 201 and Asp 222) all score highly, though several other residues

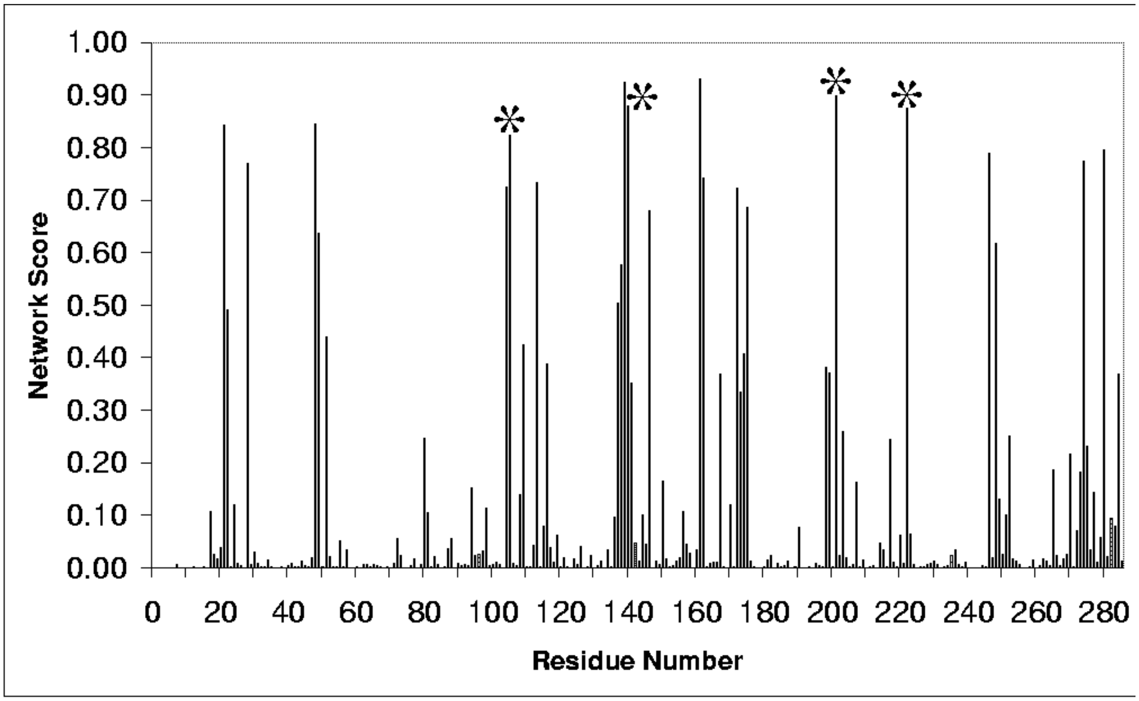


Figure 2.3: The distribution of neural network scores along the sequence of 1QPR. The true catalytic residues are highlighted.

score as high or higher. There is some grouping of the high scoring residues in the sequence, particularly around residue 140, but most high scores are isolated spikes. When the scores are mapped on to the 1QPR structure (Figure 2.4) the high scoring areas, although widely separated in the sequence, are brought together and cluster into two areas corresponding to the two active sites of the quinolate phosphoribosyltransferase homodimer.

2.3.4 Training the network

The training process is tracked by measuring the MCC after each epoch, Figures 2.5 and 2.6 show how the Matthews correlation coefficient (MCC) varies as training progresses. The variation in performance is quite considerable, with the final MCC varying between 0.35 and 0.25, reflecting the natural variation within the data set.

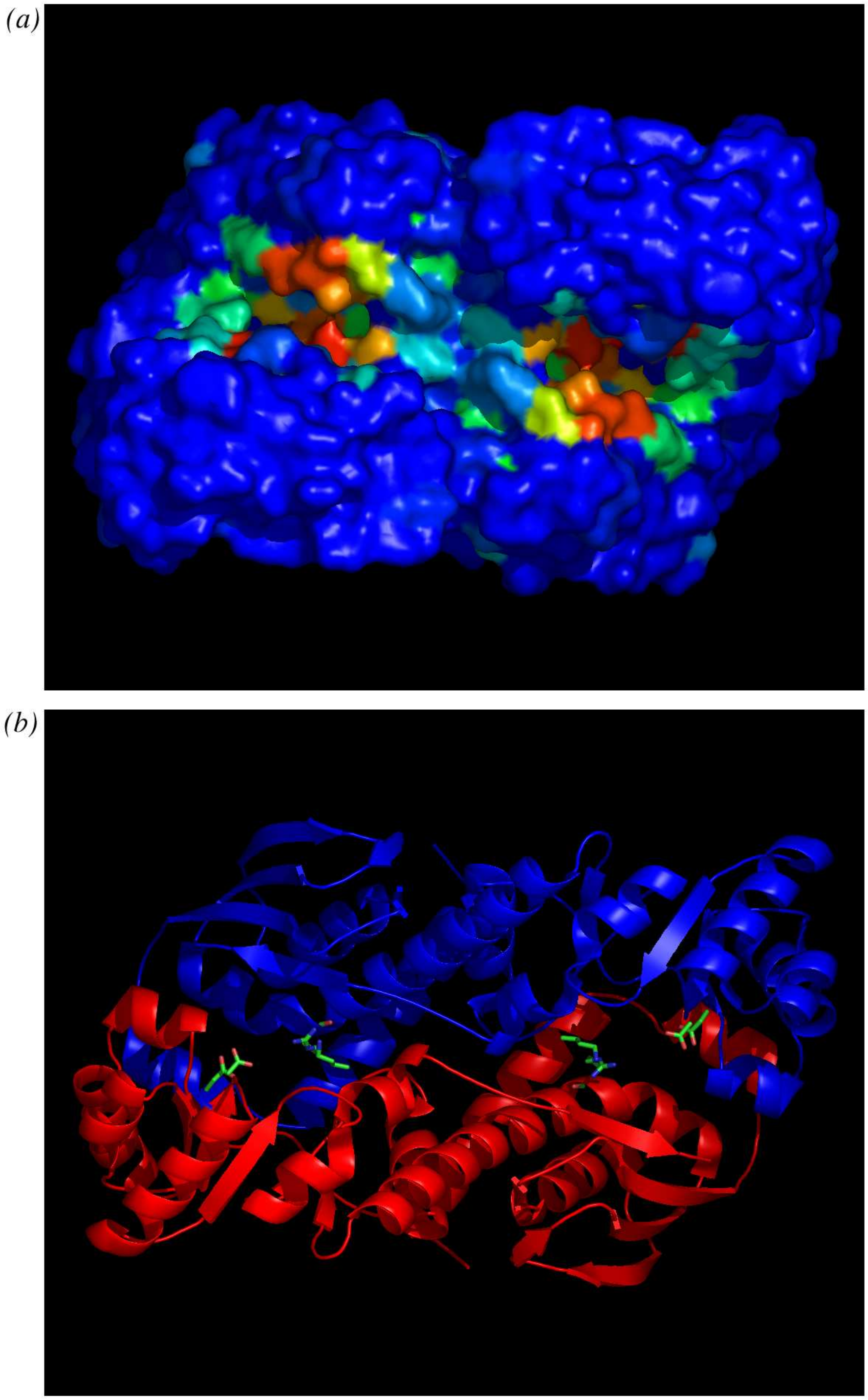


Figure 2.4: (a) Distribution of neural network scores in the 1QPR structure. Residues are coloured by network score (Red = high, blue = low). (b) The structure of the 1QPR homodimer, coloured by chain, with the known catalytic residues drawn in thick lines.

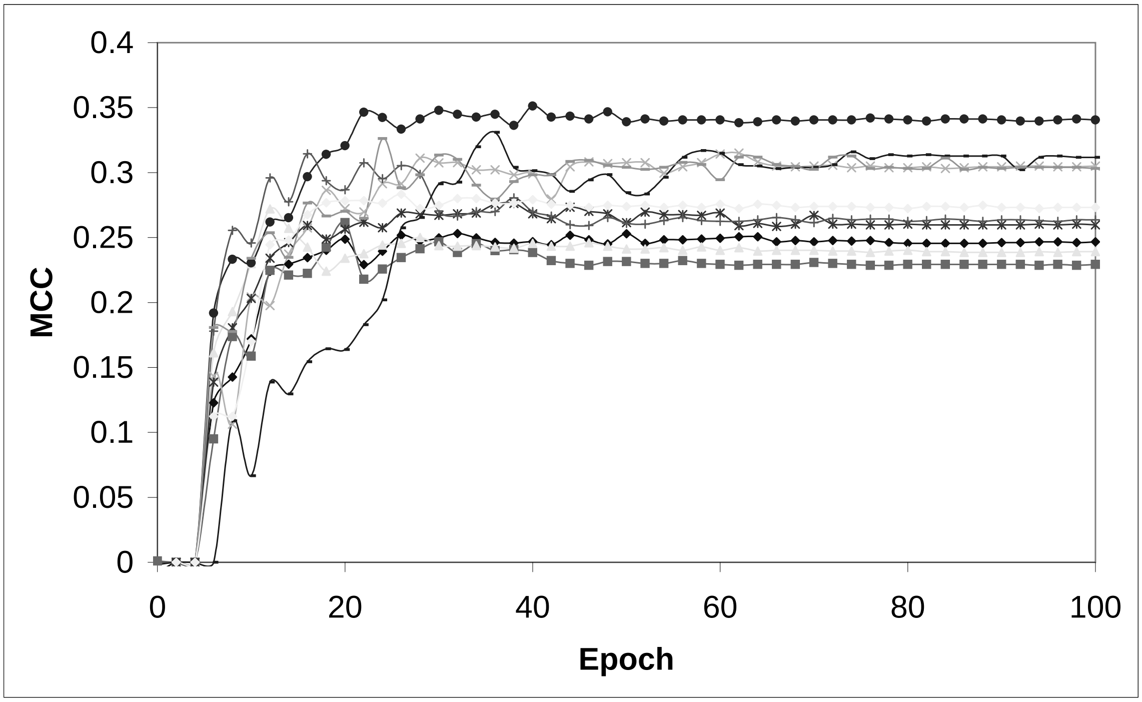


Figure 2.5: Training the neural network, each line represents one of the ten cross validation runs.

Figure 2.6 shows the MCC varying with each epoch averaged over all ten runs. The network reaches its best MCC after only 30 epochs or so, levelling off at an average MCC of around 0.28. There is no evidence of over-fitting in the results, as the MCC does not fall significantly once it has plateaued.

2.3.5 Network Weights

The relative strength of the weights that the network converges to are shown in Figure 2.7. Conservation and DOPS are both highly weighted. As expected the network also looks for buried residues, as RSA is given a negative weighting. The cleft categories show that lying in a cleft, and the size of that cleft are important factors in the network score, though not important as conservation or RSA. Depth is not weighted strongly in either direction, and is not important in making a predic-

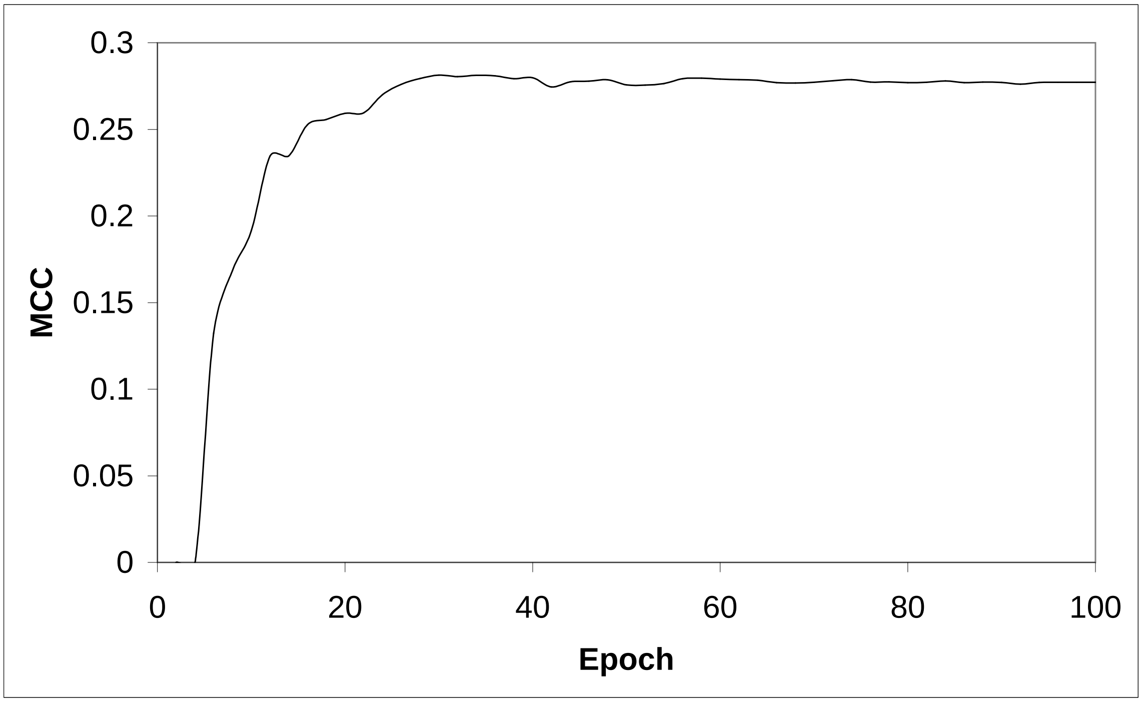


Figure 2.6: MCC averaged over all ten cross validation runs.

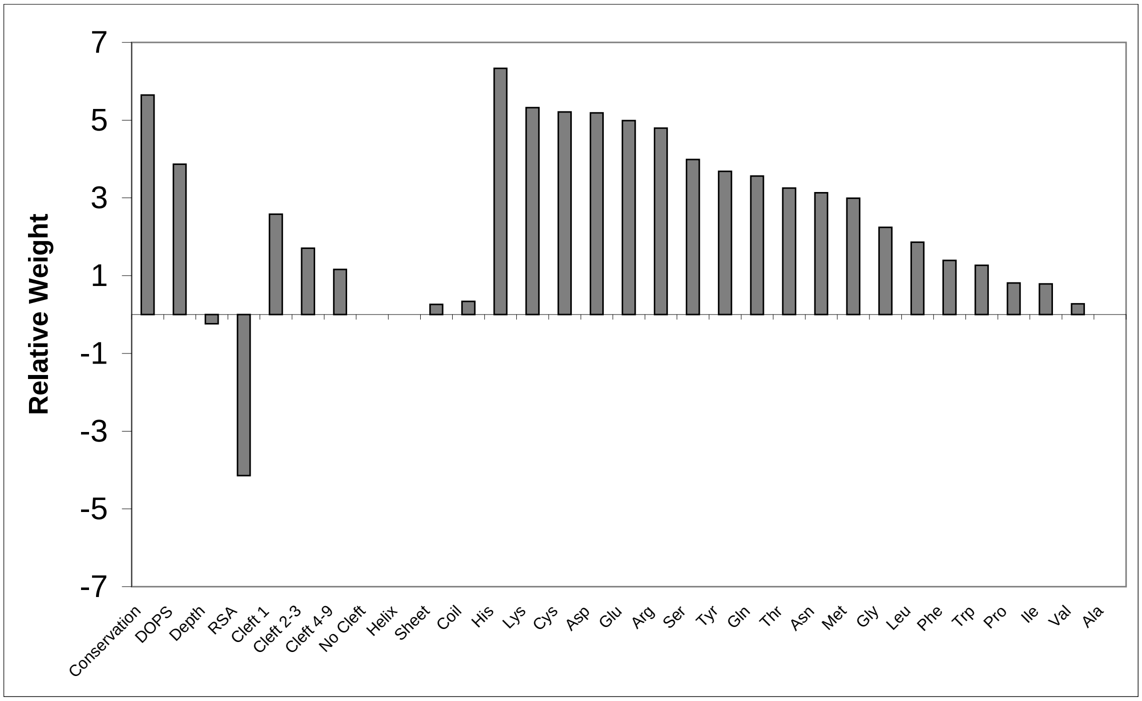


Figure 2.7: The relative strengths of the weights placed on the various parameters. Categorical parameters such as residue type are grouped, with the lowest weight set at 0.

tion. The difference for the secondary structure parameters is also small. Residue type has a very large variation with histidine, cysteine and the charged residues (aspartate, glutamate, lysine and arginine) all scoring highly, whilst the hydrophobic residues score low.

The high DOPS weighting is interesting, as it is the same for all residues within a protein chain. The only effect is to raise all the scores of all residues in chains with high DOPS and lower all the scores of all residues in chains with low DOPS. The network has learnt that when DOPS is low it is better, in terms of the overall error rate, to make no catalytic predictions at all, rather than predict everything to be catalytic. Since the clustering algorithm uses residues based on their rank rather than absolute scores, this makes no difference in the later stages.

2.3.6 Clustering

In the network scoring we consider each residue as independent of the others. However, catalytic residues are likely to cluster together in the structure. Ranking and clustering the residues allows us to use this information to improve the predictions and locate the active site. For each structure a list of possible catalytic residues is generated by ranking the residues by network score. The clustering algorithm finds distinct clusters of these residues and generates a sphere that forms the predicted active site.

1158 clusters are generated from the test set, an average of 7.2 per protein. The multimeric nature of most of the proteins means that the average number of known active sites is 2.6 per protein. The distribution of sphere sizes calculated for the known sites and all the predicted sites is shown in Figure 2.8, Figure 2.9 shows the sizes of the top scoring site in each protein.

Most predicted clusters are small and contain two or three members with a radius

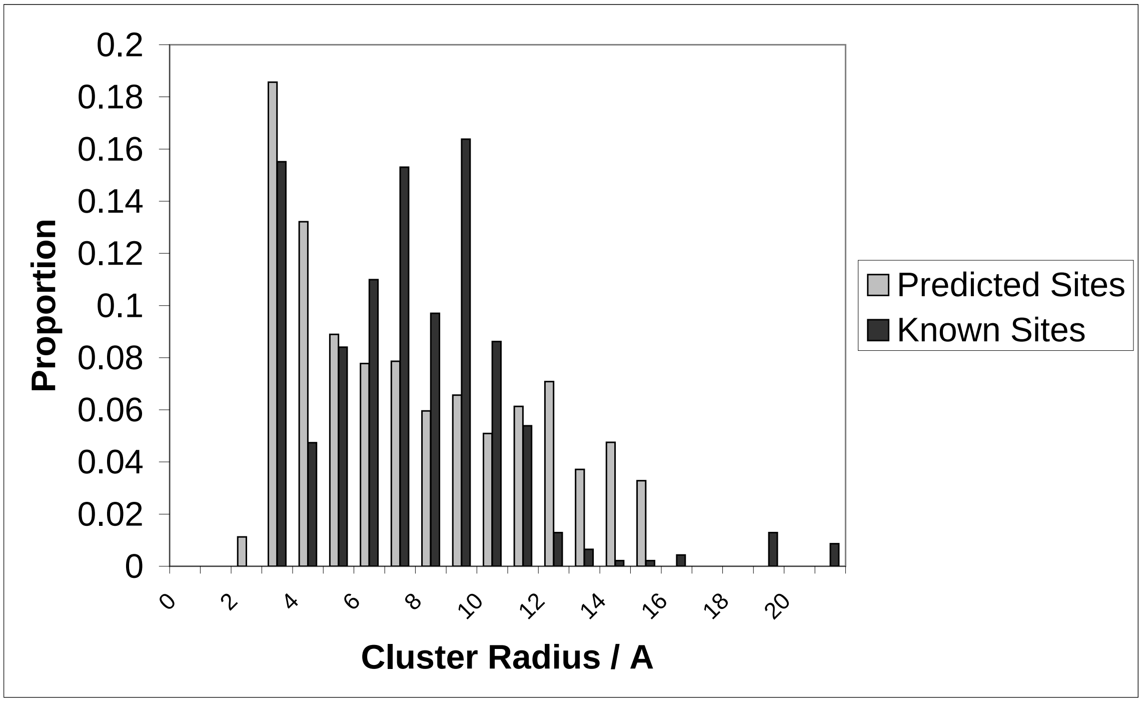


Figure 2.8: Size distribution of all the predicted sites compared to the known sites.

of 3-4Å, in contrast the top scoring predictions in each structure are generally large and lie at the upper end of the allowed size range (15Å). The known sites generate spheres with sizes between 6 and 12Å, though a significant number have a single catalytic residue and so have radii of 3Å. A few outliers have spheres larger than 20Å in radius. These cases all represent structures where the catalytic cluster is thought to come together upon substrate binding so the cluster appears very large in the unbound form.

2.3.7 Comparing the Predicted Sites to the Known Sites

To test whether a prediction is correct, the overlap between the predicted site and the closest known active site is calculated. A correct prediction occurs when the overlap is greater than 50% of the volume of the known active site, a partially correct prediction occurs when there is some overlap but less than 50%, a failure

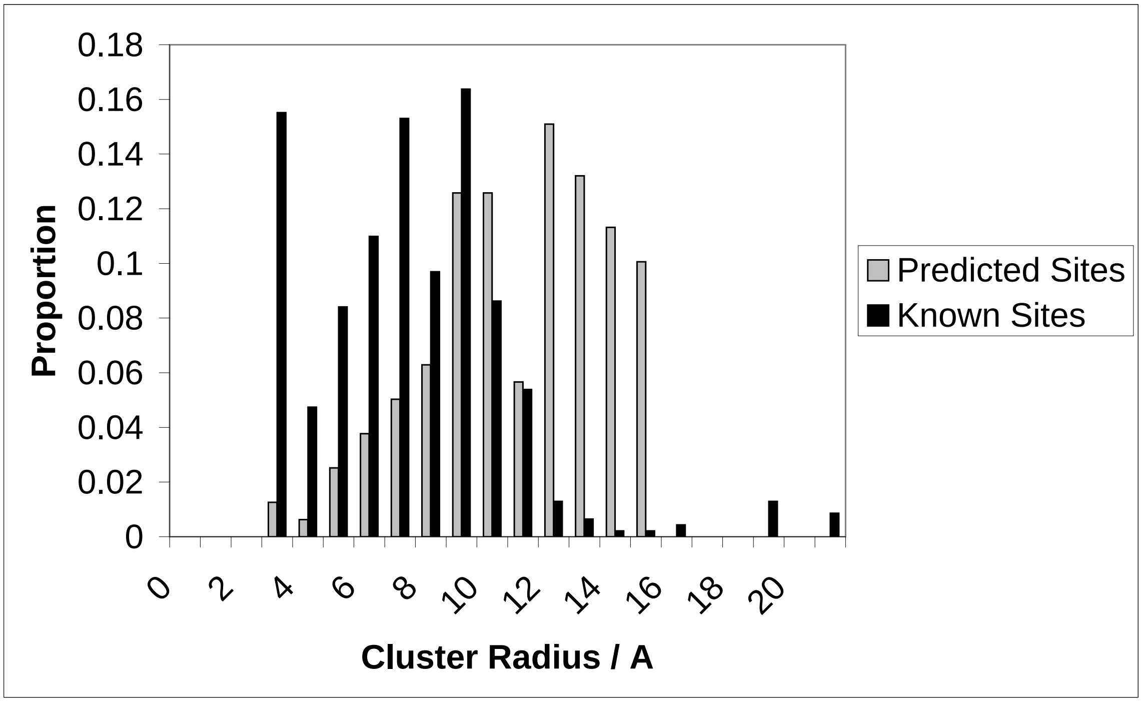


Figure 2.9: Size distribution of the top scoring predicted sites compared to the known sites.

occurs when there is no overlap between the known and predicted spheres.

For each protein in the test set, the prediction with the highest total network score was selected and compared to the known sites. The results are shown in Figure 2.10, 62% of the proteins have the active site correctly identified, and a further 22% are partially correct.

When we consider the overlap for all the sites predicted for each protein we find the results improve: 69% of the proteins have the active site correctly identified and 25% have a partially correct prediction. The increase of only 7% when all predictions are considered shows that the highest scoring cluster is very often the true active site. 11 cases were found where the top prediction was not correct, but one of the other predictions was. In 6 of these cases the correct cluster was the second or third highest scoring prediction and in 4 cases the correct cluster was the fourth or fifth highest scoring, in the final case the correct cluster was the seventh highest scoring.

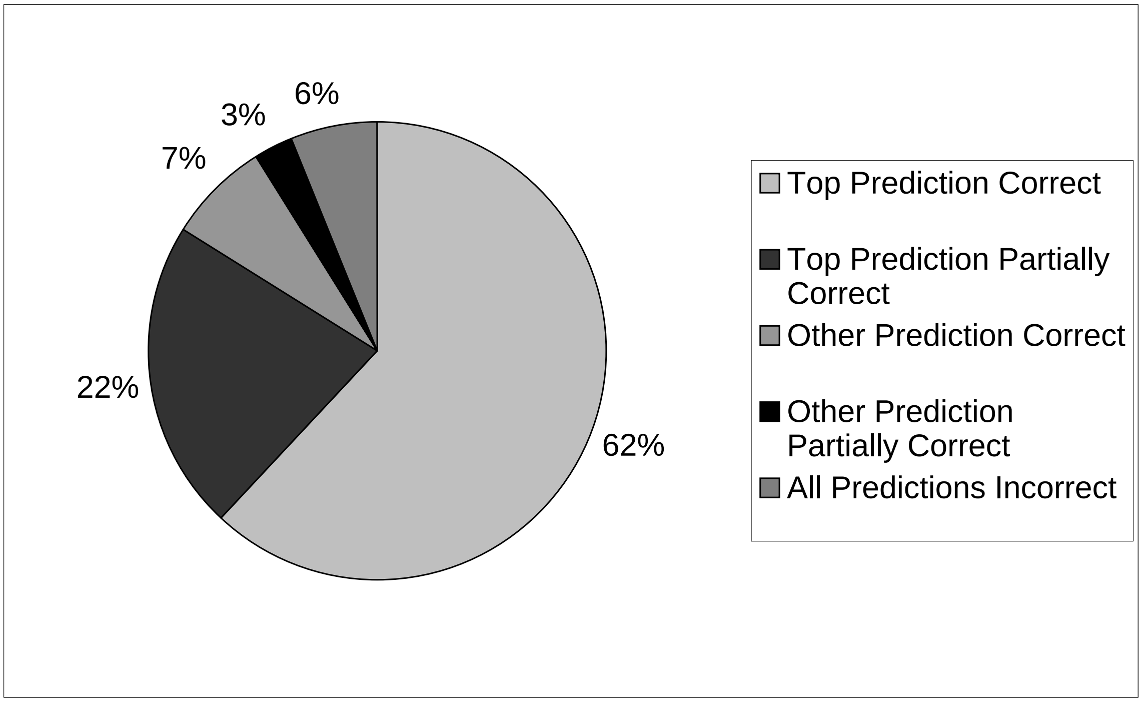


Figure 2.10: Pie chart showing the per protein accuracy when only the top prediction is considered for each protein, and when all predictions are considered.

When each of the 1158 predicted clusters is considered individually, as opposed by each protein, 25% are found to be correct and 41% are partially correct. The high number of partial hits is presumably due to the tendency of the network to find residues lying near the active site, but which aren't close enough to the true catalytic residues to score as correct. It is also possible that many of the partially correct and incorrect clusters represent secondary functional sites such as ligand binding or protein-protein interaction sites. These clusters are biologically interesting, but are considered incorrect when searching solely for active sites.

2.3.8 Significance of Results

To calculate the significance of these results, we estimate the probability ($\mathrm{P_R}$) of achieving this level of prediction by random chance. A similar method to that used by Aloy *et al* [82] is applied: To a reasonable approximation a correct hit occurs when the centre of the smaller of the two spheres lies within the volume of the larger. Since the known catalytic site is usually smaller than the largest predicted site, and assuming the prediction has an equal probability of being anywhere within the volume of the protein, $\mathrm{P_R}$ is the ratio of the volume of the predicted sphere to the volume of the protein. Since most of the proteins are multimeric this ratio is then multiplied by the number of active sites (any one of which could have overlapped with the predicted site).

The volume of each protein is estimated by drawing a sphere around all the Cβ atoms of the structure, giving an average of 510,000Å^3. Since most catalytic residues lie in the outer 5Å of the protein we shall consider the predictions restricted to only a third of this volume. The average volume of all the predicted spheres is 2632Å^3, and the average volume of the top scoring predictions is 5783Å^3. There are 7.2 predicted sites and 2.6 known sites per protein on average. A summary of the observed and expected rates of correct predictions for the three different analysis is shown in Table 2.1. We estimate the significance of the differences using equation 2.6 [82], which follows a normal distribution with mean 0 and standard deviation 1. All the results are significant to more than 10^{-15}.

$$z = \frac{P_{\mathrm{O}} - P_{\mathrm{R}}}{\sqrt{\frac{P_{\mathrm{R}}(1-P_{\mathrm{R}})}{n}}} \tag{2.6}$$

2.3.9 Comparison of the Performance of Different Networks

The neural network and clustering process incorporates a variety of different types of information: evolutionary information encoded in the conservation scores, residue propensities, structural information in the parameters and detailed structural information included in the clustering stage. To understand how these different types of information contribute to the overall performance, networks have been trained using different subsets of the parameters.

Two additional networks have been developed: Firstly a network trained solely using sequence parameters (conservation, DOPS and residue propensities), and secondly a network trained using structural parameters but excluding conservation and DOPS scores. Residue propensity is included in the structural information as the sequence of a protein would always be known given a structure. The relative performance of the different networks is shown in Figure 2.11 and in detail in Table 2.2. The performance of each network in finding the location of the active site is shown in Figure 2.12 and Table 2.3.

The performance of the technique described by Aloy *et al* [82], which uses conservation, residue propensity and clustering is also shown in Table 2.3. This study uses the same sphere based method as shown here to assess the accuracy of the predictions making comparison easy, The functional residues were based on SITE records in PDB files, which are not as well defined as the catalytic residues used in this study. Aloy *et al* analysed 106 proteins and found that 20 of them could not generate sufficiently diverse alignments to give good predictions. Since, in this study, we have included proteins with low DOPS, we include these 20 proteins as incorrect predictions when calculating performance. Once this is taken into account we find that of the 106 proteins, 68 are correctly predicted (64%), 13 partially correct (12%) and 25 are incorrect (24%), when all predictions are considered. This level of prediction is almost identical to the sequence trained network.

2.3.10 Predictions

The neural network was run on several recently published enzyme structures, which were not included in the original data, or subsequent analysis, to gauge the usefulness

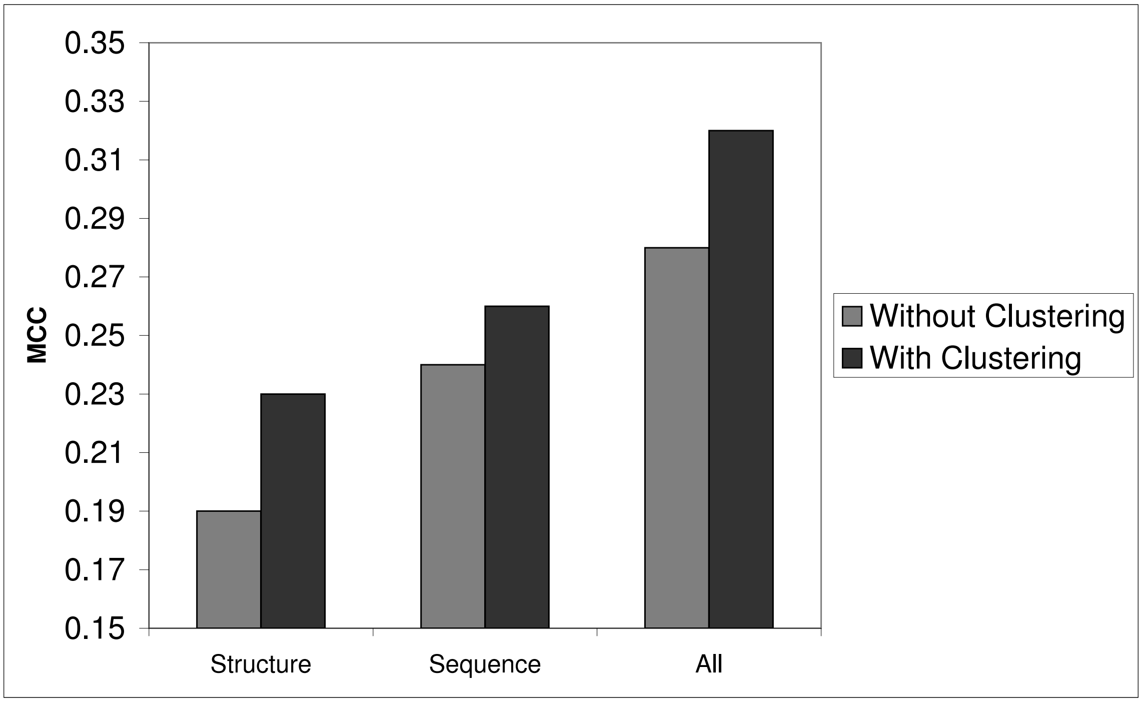


Figure 2.11: Comparison of the MCC achieved by the three different networks in predicting catalytic residues, before and after structural clustering is applied.

of the method in annotating structures.

SET Domain Histone Lysine Methyltransferases

Several recent papers [91, 92, 93, 94, 95, 96] have presented the first structures of histone lysine methyltransferase (HMTase) containing SET domains. SET domains are responsible for the methylation of specific lysine residues in histone proteins, leading to changes in chromatin regulation and gene expression. SET domains share no homology with other structurally characterised methyltransferases, and so the structure, and the functional information that the structure contains is of significant importance.

The structure of yeast protein Clr4 (PDB code: 1MVX) was used for the prediction of the functional sites. The PSI-BLAST search required 7 iterations to converge

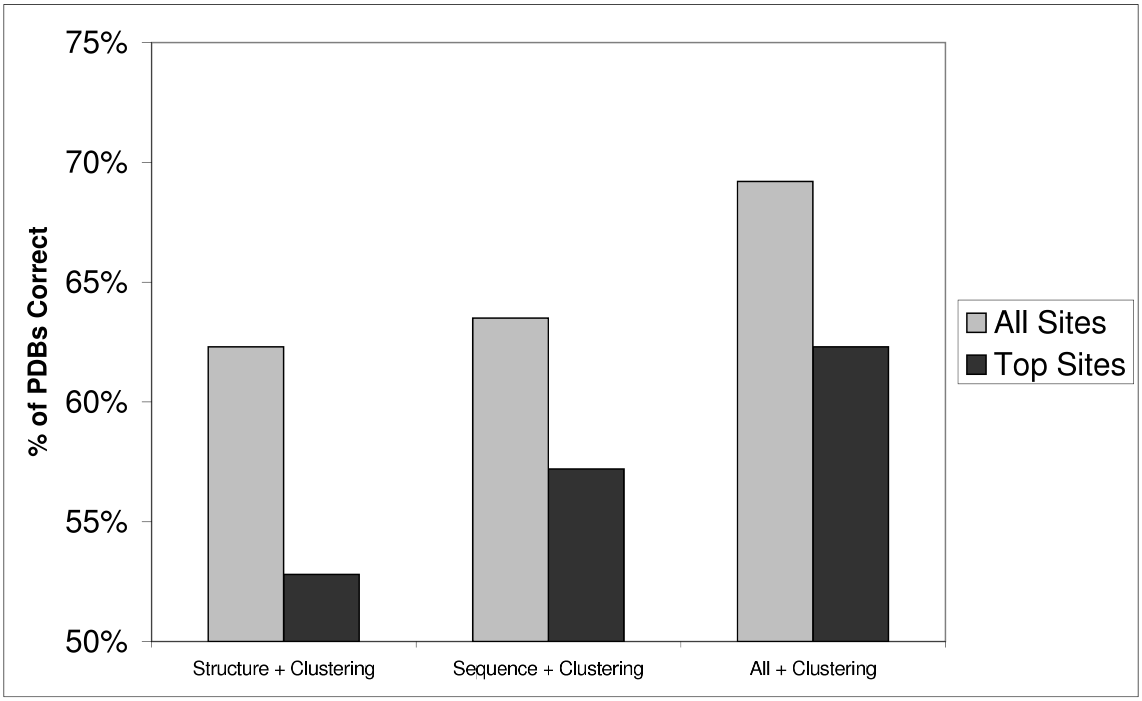


Figure 2.12: Comparison of the site prediction accuracy for the three different networks. Results are presented considering all predicted sites and the top scoring site only.

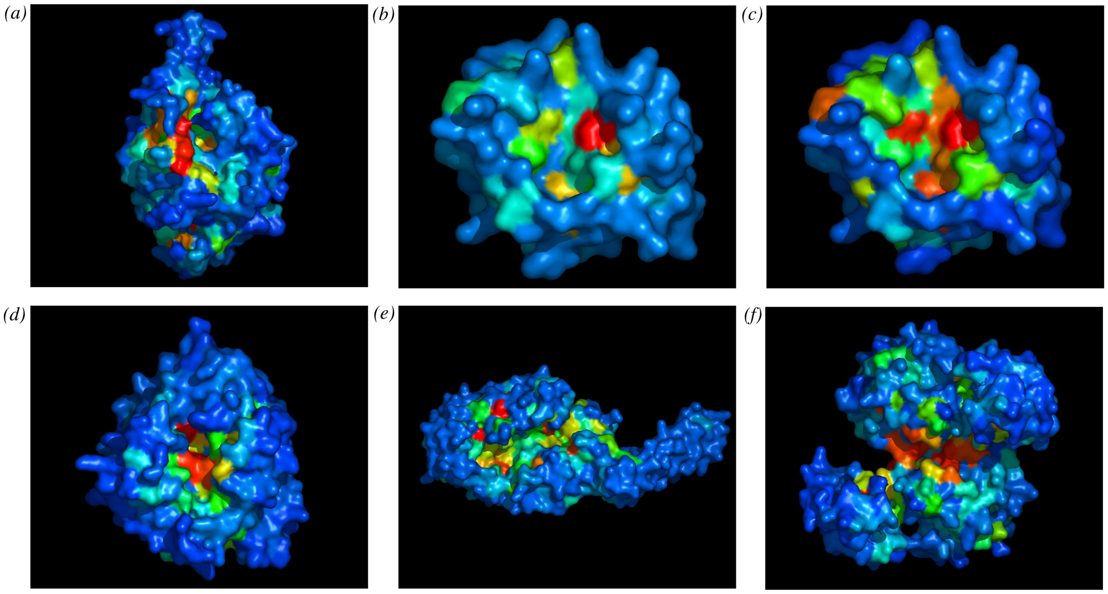


Figure 2.13: (a) The front face of the SET domain showing the large, high scoring surface patch and the Ado-HCys binding cleft. Residues His 410, Asp 450, Asn 409 and Tyr 451 form the 'L'-shaped patch in the centre, Arg 406 and Tyr 357 lie in the pocket to the right. (b) The catalytic site of I-TevI, network scores are generated without conservation data. (c) The catalytic site of I-TevI, showing the improved prediction once conservation data is included. (d) L-arabiananase, the central red patch is made of His 37 and Asp 38. Asp 158 and Glu 221 both lie close by in the same pocket. (e) FemA, the binding cleft is the long green patch in the centre of the figure. The most likely catalytic site lies at the far left end of the cleft. (f) The RlmB dimer. The subunits are stacked on top of each other running left to right. The two active sites lie in the high scoring regions in the interface between the two subunits.

(using an E-value cut-off of 10^{-20}) and found ∼150 homologues, producing a very diverse alignment. 31 residues score over the ranking cut-off and clustering reveals one large cluster, containing 19 of these residues. The output of the neural network mapped to the surface of the structure is shown in Figure 2.13. The dominant cluster forms the large 'L'-shaped patch in the centre of the structure comprising residues His 410, Asp 450, Asn 409, and Tyr 451, the other residues in the cluster extend either side of the 'L'-shape patch and into the structure.

Mutations to His 410, Cys 412, Arg 320, Glu 446 and Arg 406 have been shown to inactivate the enzyme [97, 98], though it is suggested that Arg 320 is most likely to be of structural, rather than catalytic importance. The structure with an AdoHcy cofactor bound is known for homologous SET domains. This reveals that Tyr 451, Asn 409 and His 410 make contacts to the cofactor and Tyr 357 is proposed as a possible catalytic proton source. A high resolution crystal structure of the human SET7/9 domain has been recently published [96]. This study suggests catalytic roles for the residues equivalent to Tyr 451, Tyr 419, Tyr 357 and the main chain carbonyl oxygens of Asp 403 and Phe 408. Of these functional residues, the large predicted cluster contains His 410, Glu 446, Arg 406, Tyr 357, Tyr 451 and Asp 403. The neural network identifies the correct active site and many of the known functional residues.

Intron Endonuclease I-TevI

The structure of the intron endonuclease I-TevI from bacteriophage T4 has recently been published [99]. Intron endonucleases catalyse a break in double stranded DNA, that facilitates the insertion of introns and inteins. I-TevI contains separate catalytic and DNA binding domains, the structure of the catalytic domain is analysed here (PDB code: 1LN0). Using the default PSI-BLAST parameters the sequence of 1LN0 picks up no homologues. Despite this the network still makes predictions based purely on the structure and the residue propensities. The network identifies 3 residues (His 31, His 40, Ser 42) forming the highest scoring cluster.

A putative active site is proposed based on conservation and mutagenesis data [100]. The site is located in the same cleft identified by the network. Glu 75 binds a divalent cation and is likely to be the principal functional residue. Other functional residues suggested by the authors include Tyr 17, Arg 27, His 31 and His 40. The 1LN0 structure has Arg 27 mutated to alanine, as active I-TevI cannot be produced by *Escherichia Coli*. Replacing Ala 27 by arginine in the sequence presented to PSI-BLAST, and reducing the E-value cut-off to 10^{-5}, allows conservation scores to be generated and the network to improve the prediction. 12 residues now form the largest cluster including Tyr 17, Arg 27, His 31, and His 40. However, Glu 75 still remains outside the predicted cluster.

This example demonstrates how the network can cope with structures occupying a sparsely populated region of sequence space. The prediction made only on the basis of residue propensities and structural data correctly identifies the active site and several functional residues. Once the mutated structure is corrected and conservation scores are added, the network makes improved predictions, correctly identifying the active site and many of the principal residues, though it still fails to predict the crucial Glu 75. The problems of mutated structures and limited sequence homologues highlight some of the difficulties that would be encountered in a PDB wide analysis.

α-L-Arabinanase

The structure of *Cellvibrio japonicus* arabianase has been solved recently [101] revealing a novel five-bladed β-propeller fold (PDB Code: 1GYD). Arabianase hydrolyses the arabinans polymers found in plant cell walls. The PSI-BLAST search converges after four iterations, only finding 11 homologues. However, the alignment is quite diverse and useful conservation scores are obtained. The highest scoring cluster lies centred around the high scoring pair of residues His 37 and Asp 38. The other residues in the cluster are Ser 86, Ser 112, His 92, Trp 94, Gln 316, Asp 158, Thr 58, His 291, Tyr 308, Ser 52 and Thr 53.

The authors of the paper used analogy with other enzymes [102], conservation, and mutagenesis to identify Asp 38 and Glu 221 as the likely catalytic groups.

A third carboxylate, Asp 158, is suggested to be involved in pK_a modulation or positioning of the Glu 221 side chain.

The neural network correctly identifies the three acidic residues as catalytic (all are highly ranked). However, the clustering algorithm does not link Glu 221 into the cluster containing Asp 38 and Asp 158 (even though Glu 221 is the highest scoring residue in the protein). Altering the clustering parameters to join residues separated by less than 5Å (rather than the default 4Å) allows Glu 221 to join the main cluster.

FemA

FemA is a *Staphylococcus aureus* protein identified as a member of the Fem (Factors Essential for Methcillin resistance) family, a series of antibiotic resistance genes [103, 104]. FemA is responsible for the addition of glycines to peptidoglycan molecules in the bacterial cell wall. The structure used is the first example of this important family [105] (PDB Code: 1LRZ). PSI-BLAST converges after 4 iterations finding 40 homologues and generates a diverse alignment. The network scores mapped to the structure are shown in Figures 2.13. The high scoring residues line the large cleft that runs the length of the protein. The clustering algorithm suggests a seven residue cluster comprising the high scoring residues His 106 and His 29, and five other lower scoring residues. This cluster lies at the very end of the cleft. Another five residue cluster lies approximately halfway along the cleft comprising Lys 383, Phe 382, Ser 342, Ser 314 and Thr 332, . The crystal structure does not have any ligand bound, and no mutagenesis data is available to pinpoint the actual catalytic residues.

The cleft is the only structure large enough to accommodate the peptidoglycan substrate and hence is the most likely binding site, though a conformational change on substrate binding cannot be ruled out. The network suggests several residues as potential catalytic groups and further experimentation is required to confirm which, if any, of these residues form the catalytic centre.

RlmB 23S rRNA Methyltransferase

RlmB is an *Escherichia coli* protein representing the novel Ado-Met dependent methyltransferase class, SPOUT. RlmB is responsible for the methylation of a specific guanosine group in the 23S rRNA component of the ribosome [106]. The crystal structure of the enzyme has recently been solved [107] (PDB Code: 1GZ0). PSI-BLAST converges after iteration 5, having found 100 homologues and generates a very diverse alignment. RlmB forms a homodimer in solution and the high scoring residues cluster into two almost identical sites in the dimer interface region. Each site contains residues from both chains A and F. The highest scoring residue is Arg 114 which is involved in a salt bridge with Glu 198 from the opposite chain. Surrounding this pair are His 9, Asp 117, Glu 147, Ser 148 and Gly 144 from the same chain as Arg 114 and Ser 224, Leu 225, Asn 226 and Ser 228 from the same chain as Glu 198. A secondary cluster comprised of Asp 105, His 107 and Asn 108 lies 4.3Å from this main cluster.

The authors propose a putative active site based on conservation of three previously identified motifs, found in most methyltransferases [108, 109]. Motif 1 covers residues Asn 108 to Arg 114, motif II covers Glu 198 and motif III covers Ser 224, Leu 225 and Asn 226. They also report that mutagenesis of the equivalent residue to Glu 198 in a homologue abolishes methyltransferase activity. Glu 198 and Ser 224 are suggested as possible catalytic bases. The His 9 residue, identified by the network, is implicated in RNA binding. However, several other putative RNA binding residues are not identified.

The network has correctly identified the putative catalytic centre, though again the clustering has split the site, leaving part in a small secondary cluster.

2.4 Discussion

One of the original aims, to predict catalytic residues from structures, has proven to be an extremely difficult task given the narrow definition of ‘catalytic’ used here. The MCC of 0.28 (or 0.32 if clustering is used) is too low to realistically use the simple predictions from the neural network in identifying catalytic residues directly. The main problem is the high number of false positives. 56% of catalytic residues are identified correctly, but only one in seven catalytic predictions are correct.

Visual inspection of the results shows that many of the false positives are other functional residues lying in the active site such as substrate binding and metal binding residues. These residues have very similar properties to the catalytic residues: conserved, low solvent accessibility, lying in clefts and they also lie extremely close to the true catalytic residues and do not form a distinct or separate spatial cluster. A system looking to identify any functional residues at the active site may well consider these false positives to be true positives. However, given the definition used in this study they are errors. As well as the problem of these ‘false’ positives there is the inherent difficulty of picking the handful of catalytic residues from hundreds in the protein. The ratio of catalytic to non-catalytic is around one in one hundred across the entire data set. Given these difficulties, the low success rate is understandable and not as disappointing as first appears.

The network weights and the performance of the sequence-only neural network shows that evolutionary information, encoded in conservation scores is very important in making a prediction. This network reflects the performance that one could expect to achieve when predicting catalytic residues purely from sequence data. We see from the Q_{Observed} and $Q_{\text{Predicted}}$ values in Table 2.2 that 50% of catalytic residues are found by this network, but only one in eight of the predictions is correct.

Structural genomics projects aim to provide some level of structural information for the majority of protein sequences. Some of these proteins will not have any known sequence homologues and the structure will be the only information available. The neural network trained without conservation scores reflects the performance one could expect to achieve when analysing these proteins. Although the network alone performs poorly, the structural information can also be used to cluster the predictions in these proteins. When this form of structural information is incorporated the overall performance rises almost to the level of the sequence network, and 57% of the catalytic residues are correctly predicted, though the true positives are still only one in ten of the catalytic predictions.

For the majority of structural genomics targets there are some sequence homologues and in these cases both types of information can be incorporated. The network trained using sequence and structure outperforms both the other networks with an MCC of 0.28 rising to 0.32 when clustering is used (Table 2.2). 68% of catalytic residues are correctly predicted and one in six of the catalytic predictions is correct.

Although predicting the catalytic residues is difficult, predicting the location of the active site can be done with significant levels of success (Table 2.3). When only structural information is used the clustering algorithm is still able to correctly identify the catalytic cluster in 62% of proteins and a partially correctly in a further 31%. This suggests that even for structural genomics targets where no conservation data is available, it will still be possible to make significant predictions about the location of the active site.

The neural network trained using sequence data identifies 63.5% of sites when all predictions are considered. This level of performance is similar to the technique described by Aloy *et al* [82] which also uses conservation, residue propensities and clustering. It should be noted that Aloy *et al* compared their predictions to the SITE records of PDB files, which are less rigorously defined than the catalytic clusters used in this study, and generally comprise larger number of residues. The performance of the neural networks used in this study are, therefore, likely to be underestimated compared to the method of Aloy *et al*.

As with finding catalytic residues using the neural network output, when structure and sequence are combined, the our performance in finding catalytic sites exceeds that of sequence or structure alone. In this case 69% of sites are correct considering all predictions and 62% considering only the top prediction. A further 25% of sites are partially correctly predicted when all predictions are considered and 22% when only the top prediction is considered. The method fails to make a useful prediction in only 6% of cases when all the predictions are examined

One of the justifications for the large investment made in structural genomics is that it will allow identification of functional sites and residues in cases where it is not possible from sequence. The results we have shown here indicate that structure alone can be used to identify catalytic residues and active sites in enzymes. However, evolutionary history encoded in the form of conservation scores is an extremely rich source of information for making these types of predictions and should be incorporated at every opportunity. The improvement in performance when structure *and* sequence are used, shows that structural information, other than that used for clustering, should be incorporated into *de novo* prediction techniques such as evolutionary trace.

2.4.1 Why Did The Failures Fail?

When considering the top scoring sites in each protein we find that 16% of the proteins failed to find any overlap between the predicted spheres and the known catalytic cluster. It is important to understand why these failures occurred in order to improve the algorithm and assess whether there are specific types of enzyme on which the algorithm performs consistently badly.

Poor Alignments

The alignments automatically generated by PSI-BLAST are the most likely point of failure. The optimal E-value cut-off for each family varies depending on its size and diversity. The single E-value cut-off used represents the best compromise, but still generates poor alignments for some families. To test whether poor alignments are the major source of error the difference between the conservation of the catalytic residues and the conservation of all residues was calculated and averaged for each group of results (correct, partial and incorrect), the results are shown in Figure 2.14. The different groups clearly show a variation in the distinction between conservation of catalytic and non-catalytic residues. In the correctly predicted group the difference is more than 0.3, this falls to 0.25 for the partially correct group, and the incorrect group has an average difference of only 0.15. Clearly, given the importance of conservation scores in making predictions, a lack of differentiation between the conservation of catalytic and non-catalytic residues will reduce the overall accuracy.

This trend implies that unusual conservation scores are responsible for a large part of the failure rate. The low difference in conservation scores in the failure group could be explained if these proteins all had low DOPS. The DOPS for each protein chain were averaged for each category and are also shown in Figure 2.14. There is a correlation between DOPS and the success of a prediction, but when we look

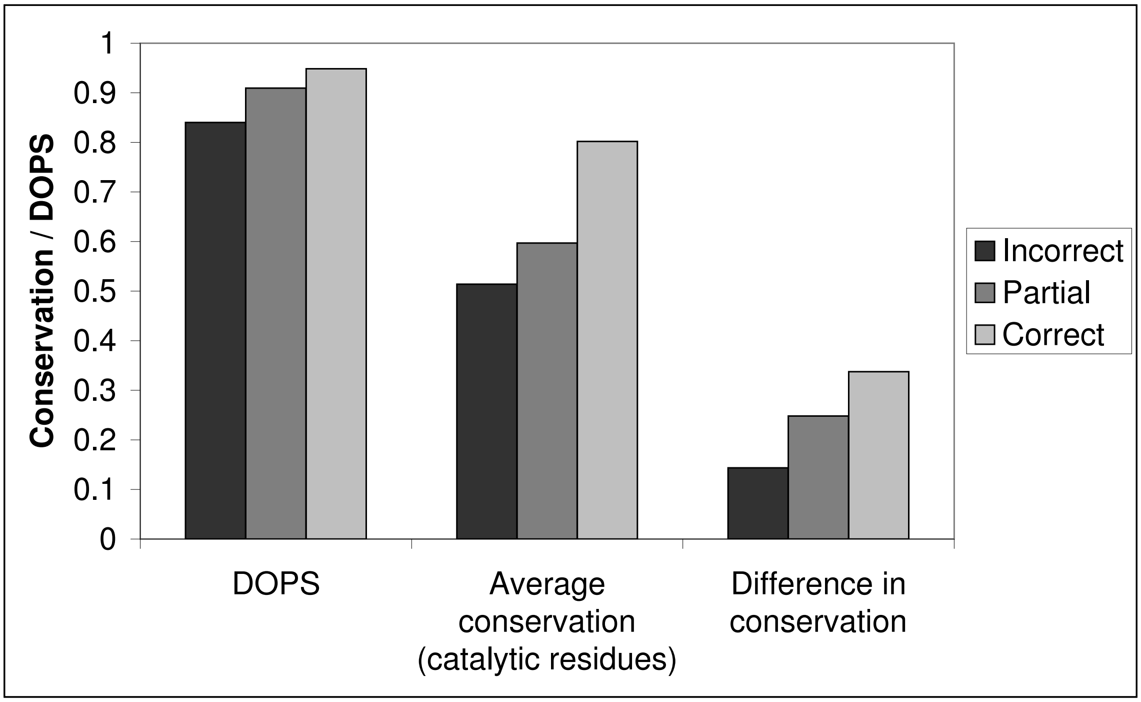


Figure 2.14: The difference in DOPS and conservation between catalytic and non-catalytic residues in the three groups of results.

at the scores themselves we see that, although some chains have very low DOPS, most are just as high as the average correctly predicted protein. If low DOPS were responsible for all of the failures then one would not expect the average conservation of the catalytic residues to vary across the three groups. However a clear trend of increasing catalytic conservation in the correct predictions is detected and shown in Figure 2.14.

How then to explain these anomalous conservation scores? The assumption must be that these enzymes are part of a larger family of proteins, which have different catalytic activities. Catalytic residues conserved within a sub-family would therefore, vary between members of the family and not be necessarily conserved. Several examples of this can be seen in the failed structures. Calpain, for instance, contains an EF-hand domain which is even found in non-enzymes. This means the catalytic residues of Calpain are not conserved in many of the homologues a PSI-BLAST search returns, whilst other residues involved in forming the EF-hand are conserved. This pattern of conservation is the inverse of what the network is expecting, and so it fails to correctly predict the catalytic residues

Clustering Errors

Of the 26 structures that failed to find the active site when only the top site was considered, 10 also failed when all sites were considered. In these 10 cases the error occurs prior to clustering, generally with poor alignments from PSI-BLAST. Of the remaining 16, 11 generated a lower scoring correct cluster and 5 generated a lower scoring partially correct cluster. These 16 cases are failures of the clustering algorithm to find the right cluster, presumably because the signal from the true active site was weak compared to other sites in the protein.

If each structure is analysed by hand, the fault is generally obvious. The single-linkage algorithm is prone to forming long aspherical clusters, since two separate clusters can be joined even if only a single residue joins them. In several failures the true active site is a relatively compact cluster with a few high scoring residues, whilst the top scoring prediction is a large cluster which out-scores the others by its size even if no single residue scores highly. Another problem is that the algorithm tends to select clusters buried in the protein, since these contain more residues than surface clusters, a human can easily spot that these are not suitable active sites. Searching for surface patches rather than spherical clusters may prevent this happening. However, as mentioned above, many catalytic residues do not lie fully on the surface of the protein.

Table 2.1: Observed and expected frequencies of correct results for the three analysis.

Per Site (n=1158)			**Per Protein (n=159)**			**Top site (n=159)**		
Expected (P_{R})	Expected (P_{R}) ($\frac{1}{3}$ Vol)	Observed (P_{O})	Expected (P_{R})	Expected (P_{R}) ($\frac{1}{3}$ Vol)	Observed (P_{O})	Expected (P_{R})	Expected (P_{R}) ($\frac{1}{3}$ Vol)	Observed (P_{O})
1.3%	4.4%	24.7%	9.6%	32.2%	69.2%	2.9%	9.8%	62.3%

Table 2.2: Comparison of the performance of the three different neural networks in predicting catalytic residues.

Data Used	**Before Clustering**			**After Clustering**		
	MCC	$Q_{\mathrm{Predicted}}$	Q_{Observed}	MCC	$Q_{\mathrm{Predicted}}$	Q_{Observed}
Structure	0.19	0.10	0.41	0.23	0.10	0.57
Sequence	0.24	0.13	0.50	0.26	0.13	0.58
Sequence + Structure	0.28	0.14	0.56	0.32	0.16	0.68

Table 2.3: Comparison of the performance of the three different neural networks in locating active sites.

Data Used	**Top Sites Only**			**All Sites**		
	Correct	Partial	Incorrect	Correct	Partial	Incorrect
Structure	**52.8%**	25.8%	21.4%	**62.3%**	31.4%	6.3%
Sequence	**57.2%**	27.7%	15.1%	**63.5%**	28.3%	8.2%
Sequence + Structure	**62.3%**	21.4%	16.4%	**69.2%**	24.5%	6.3%
Aloy *et al*	**-**			**64.2%**	12.3%	23.5%

Chapter 3

Compiling A Dataset of Catalytic Residues

3.1 Introduction

Large scale analyses of protein structures have been used in the past to gain insights into many aspects of protein structure and function. However, using such an approach to study enzyme catalysis has been stymied by the lack of a large, well annotated dataset. For instance, an analysis in 1988 by Zvelebil and Sternberg[110] looked at just 17 different enzymes. The dataset published in the first release of the CSA[75] was an important step forward. It features 178 non-homologous enzymes annotated using a clear definition of what comprises a catalytic residue. However, even this dataset is relatively small compared to the thousands of different enzymes in nature and an order of magnitude smaller than the datasets used in recent general studies on protein structure[111] that use thousands of non-redundant structures.

```
FT   METAL       270    270       MANGANESE 2 (BY SIMILARITY).
FT   METAL       275    275       MANGANESE 1 AND 2 (BY SIMILARITY).
FT   METAL       293    293       MANGANESE 2 (BY SIMILARITY).
FT   METAL       352    352       MANGANESE 1 (BY SIMILARITY).
FT   METAL       354    354       MANGANESE 1 AND 2 (BY SIMILARITY).
FT   ACT_SITE    282    282       POTENTIAL.
FT   ACT_SITE    356    356       POTENTIAL.
FT   MUTAGEN     354    354       E->A: LOSS OF ACTIVITY.
```

Figure 3.1: Extract of the sequence features from the SWISSPROT entry from AMPA_ECOLI, the catalytic residues are shown in the ACT_SITE fields.

3.1.1 Current Data Sources

Both the SWISSPROT/UNIPROT[112] and PDB databases, as well as containing raw sequence and structural data, have some facility for annotating the proteins they contain. However, in neither case are annotations mandatory, and overall annotation coverage is low, particularly in the PDB.

SWISSPROT is a curated protein sequence database which aims to provide a high-level of annotation for the sequences it contains, including active site information. A sample entry for cytosol aminopeptidase from *Escherichia coli* (AMPA_ECOLI) is shown in Figure 3.1. The ACT_SITE fields contain the sequence numbers for potential active site residues. The ACT_SITE definition in the SWISSPROT manual is simply ‘Amino acid(s) involved in the activity of an enzyme’, a broad definition which could potentially include catalytic, binding or regulatory residues. Also included are the METAL fields which record those residues involved in metal ion binding (manganese in this case).

PDB files contain a SITE record similar in purpose to the ACT_SITE record in SWISSPROT. An example from the structure of the AMPA_ECOLI sequence (PDB code: 1GYT) is shown in Figure 3.2. Three separate site records are made: ZA1, ZA2 and COA, corresponding to the two metal binding sites (binding zinc in this

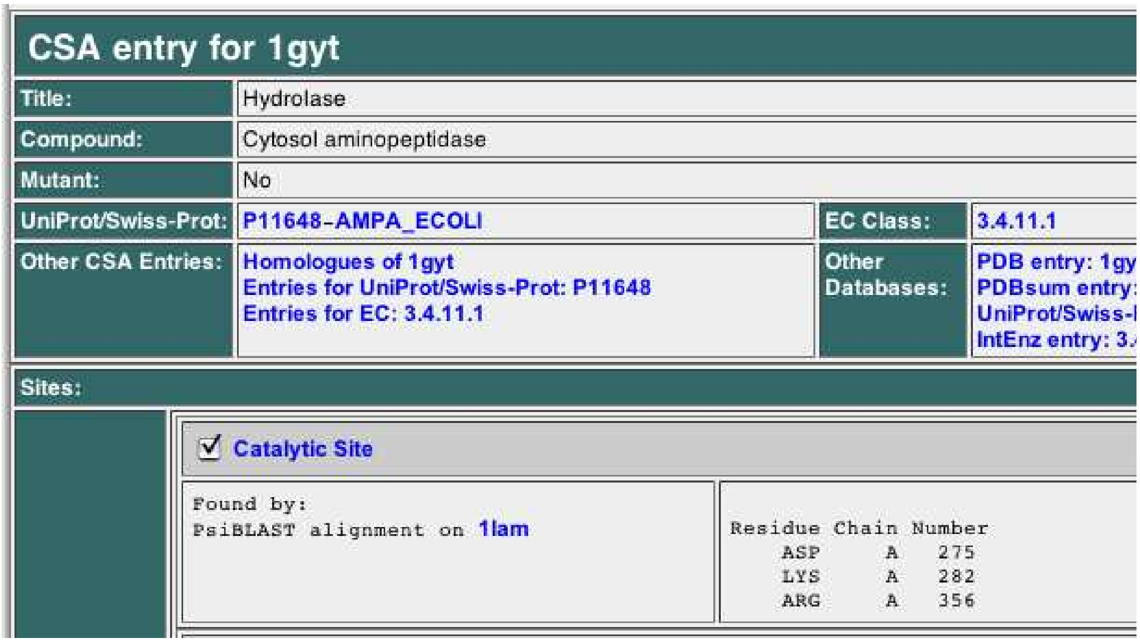
```
SITE     1 ZA1  5 LYS A 270  ASP A 275  ASP A 293  GLU A 354
SITE     2 ZA1  5 HOH A1258
SITE     1 ZA2  4 ASP A 275  ASP A 352  GLU A 354  HOH A1258
SITE     1 COA  7 LYS A 270  ASP A 352  ALA A 353  GLU A 354
SITE     2 COA  7 GLY A 355  ARG A 356  LEU A 380
```

Figure 3.2: Extract of the SITE records from the PDB entry for 1GYT.

case rather than manganese) and a carbonate ion binding site that represents the active site. It would seem that the two databases agree on the residues involved in metal binding, but determining the catalytic residues from this data is impossible. One of the residues mentioned in SWISSPROT (Arg 356) is listed in the COA site, but its function is unclear from this entry.

An entry from the CSA web site for 1GYT is shown in Figure 3.3. This shows that there are in fact three catalytic residues in cytosol aminopeptidase. Two (Lys 282 and Arg 356) are mentioned in the SWISSPROT entry. One of these is also mentioned in the PDB SITE record (Arg 356), along with the third residue (Asp 275). This example is typical of the three data sets, in that SWISSPROT is usually conservative with annotations (missing Asp 275 in this case), PDB SITE records are more general and vary greatly in the type of residues recorded, whilst only the CSA provides specifically catalytic annotations.

One aim for this annotation effort is to cover as broad a range of the functional and structural space, as measured by the EC and CATH classifications, as possible. With this is in mind, we note that the potential size of the dataset is growing all the time, as new structures are solved. Currently ∼100 EC classes that did not previously have a representative structure, have one solved each year, and there are ∼1000 EC classes with representative structures in total. This steady increase in the amount of structural enzyme data available is shown in Figure 3.4. Clearly, a continual annotation effort is required to keep pace.



CSA entry for 1gyt

Title:	Hydrolase		
Compound:	Cytosol aminopeptidase		
Mutant:	No		
UniProt/Swiss-Prot:	P11648-AMPA_ECOLI	EC Class:	3.4.11.1
Other CSA Entries:	Homologues of 1gyt Entries for UniProt/Swiss-Prot: P11648 Entries for EC: 3.4.11.1	Other Databases:	PDB entry: 1gy PDBsum entry: UniProt/Swiss-I IntEnz entry: 3./

Sites:

☑ Catalytic Site

```
Found by:
PsiBLAST alignment on 1lam
```

```
Residue Chain Number
    ASP      A    275
    LYS      A    282
    ARG      A    356
```

Figure 3.3: Extract of the CSA record for 1GYT.

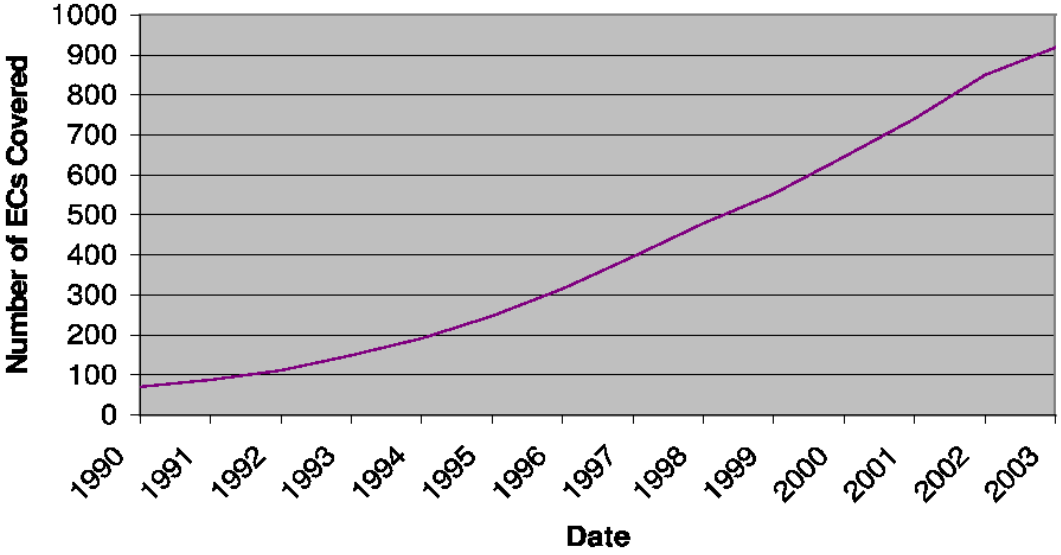


Figure 3.4: The increase in EC class coverage by the PDB over the last 13 years.

3.2 Compiling the Dataset

One of the main reasons for the difficulty in producing a large, well annotated dataset of catalytic (or even functional) residues in proteins is that definitions of ‘catalytic’ vary. Clearly, residues involved in substrate binding or even keeping the active site architecture in the correct orientation are vital for the correct functioning of an enzyme, but they are not catalytic in the usually used sense. Therefore, one of the most important initial tasks in the annotation effort is to precisely define which residues are included as ‘catalytic’ residues. The definition used (given earlier, but repeated here for clarity) comprises four tests, passing any one of which labels a residue as catalytic:

1. Direct involvement in the catalytic mechanism - e.g. as a nucleophile.
2. Exerting an effect on another residue or water molecule, which is directly involved in the catalytic mechanism, which aids catalysis (e.g. by electrostatic or acid-base action).
3. Stabilisation of a proposed transition-state intermediate.
4. Exerting an effect on a substrate or cofactor which aids catalysis, e.g. by polarising a bond which is to be broken. Includes steric and electrostatic effects.

Using the serine proteases as an example: the serine nucleophile and the histidine acid/base are catalytic because they are directly involved in the mechanism. The histidine residue also qualifies by the second rule since it exerts an effect on the serine that is crucial for catalysis, again by acid/base chemistry. The aspartate is also considered catalytic, because it exerts an electrostatic effect on the histidine that primes it for action. The two amide groups that form the oxyanion hole are catalytic, because they stabilise the tetrahedral transition state. They can also be thought of as effecting an effect on the substrate by helping to polarise the carbonyl bond.

In most cases a large quantity of research is required to identify the catalytic residues of an enzyme. The information is generally embedded within the natural language text of many different papers and so cannot be automatically extracted with any reliability. Expert annotators are required to read each paper and weigh the evidence for each residue in any proposed mechanism. This evidence comes from a number of sources: mutagenesis, pH dependence and chemical labelling experiments are all important, but often the annotation comes from simply examining the structure of the ligand bound enzyme to find those groups appropriately positioned relative to the substrate.

3.2.1 Description of the Database

The initial annotation effort to expand the data beyond the core CSA dataset used by Bartlett *et al*[76] is recorded in a simple relational database implemented using the schema shown in Figure 3.5. The 'PDB' table provides a record of each PDB file annotated along with a 'Yes/No' indication of whether the annotation is complete (in some cases only a few of the catalytic residues can be identified), a record of the user who made the annotation and the date the entry was made. This table is linked via the pdb_id filed to the 'Literature' table which simply provides a record of the Pubmed identifiers of the articles which provided the information used to make the annotation. The bulk of the data is stored in the 'Residue' table which records the catalytic residues for each PDB file. The 'res_num', 'chain' and 'res_type' fields record the number, chain and type of the residue in the PDB file. 'res_mod' records

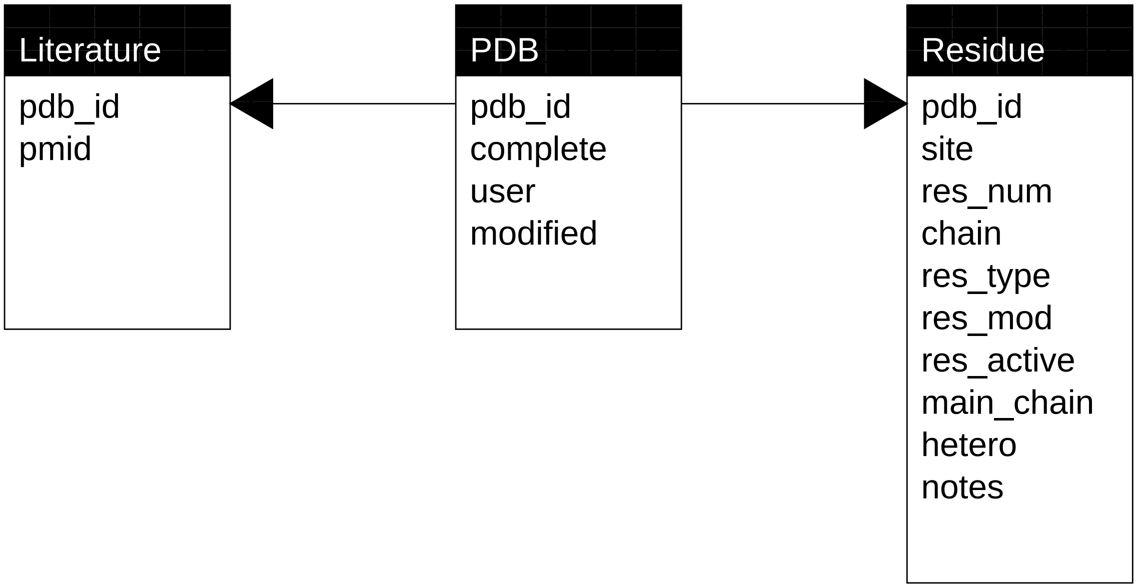


Figure 3.5: The schema used in the catalytic residue database.

the original amino acid residue type in case the PDB file is a mutant. 'res_active' records the three letter code of the residue in the active form of the enzyme, some enzymes, for instance, require a phosphorylated tyrosine for activity, which may not be present in the PDB file. Also recorded is whether the side chain or main chain groups are responsible for catalysis, and whether the entry is recorded in the HETATOM or ATOM field of the PDB file. A notes field is also provided.

Subsequent, ongoing, annotation efforts have seeked to expand the quantity of information recorded to include mechanistic details as to the role of each residue and the type of experimental evidence used to make the annotation. This information should help with more detailed analyses of the data (for instance, allowing a user to look at which types of residue fulfil a certain mechanistic role), and also help to integrate the CSA with the mechanism centred MACiE database.

The extra functional information recorded includes the group upon which the residue acts upon (substrate, transition state, water, another residue or cofactor), and the chemical function that is performed: acid/base, electrostatic interactions, nucleophilic attack and steric effects.

Once a single PDB has been annotated, the annotation can be transferred to homologous proteins which are likely to use identical or similar mechanisms. The procedure for the transfer of annotation is to run PSI-BLAST on a non-redundant sequence database using the sequence of the original annotated protein as the query. Each homologue, returned by PSI-BLAST, is tested to see whether the catalytic residues are conserved, and the annotation is transferred if they are. In some of these cases, the homologue will have diverged sufficiently from the original protein to have a different EC class. In most cases however, the homologue will either perform exactly the same catalysis (being an example of the same enzyme from a different organism or crystallised in a different state) or very similar catalysis, but with different substrate specificity.

3.3 Results

3.3.1 PDB and EC Coverage

The results in this section refer to the state of the CSA database in June 2003, at the end of the effort to expand the CSA annotation beyond the initial core dataset. The ongoing annotation effort has expanded the coverage significantly in recent months (July-August 2005).

There are ∼10,000 PDB files in the ENZYME database, each of which has an associated EC class. The coverage of these entries by the CSA is shown in Figure 3.6: 493 PDB files are directly annotated (373 having complete annotations and 120 incomplete), this annotation can be transfered to a further ∼4500 PDB entries by homology with a directly annotated entry. The final annotation coverage represents around half of the ∼10,000 PDB entries in ENZYME.

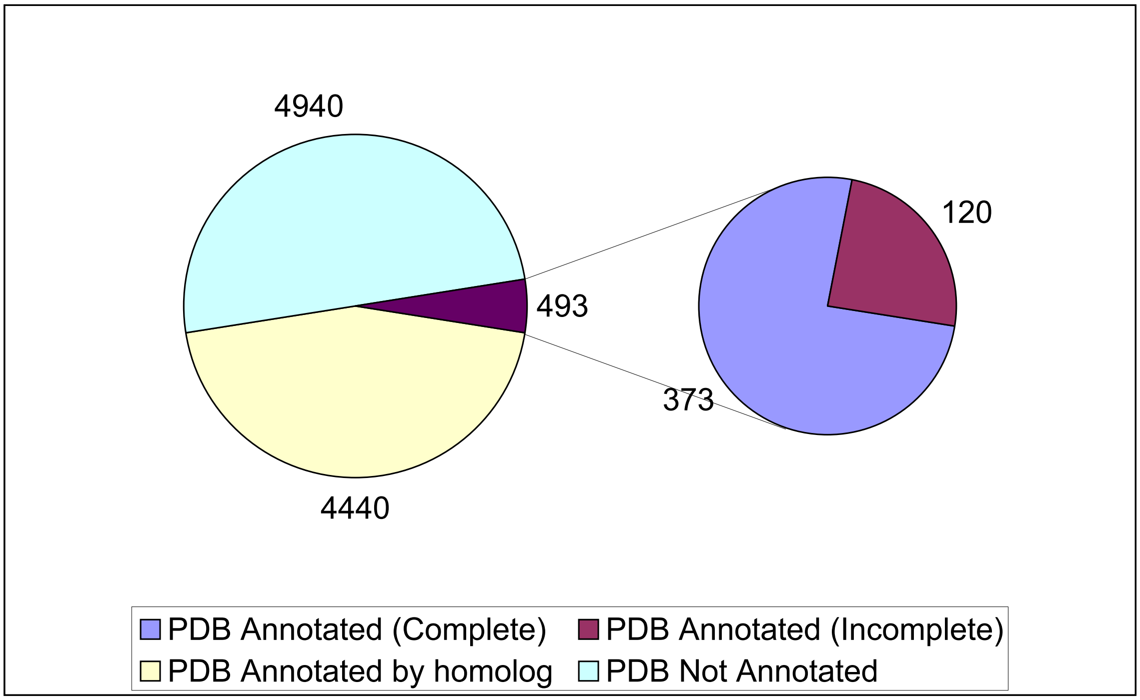


Figure 3.6: Coverage of the PDB by the CSA as of June 2003.

The EC classification contains ∼4000 different enzyme reactions, of which 971 have at least one member with a known structure. It should be noted again at this point, that a single EC could be catalysed by several different families of enzymes with different mechanisms, which means that for complete coverage, more than one annotation for each EC may have to be made. 482 different EC classes were annotated and the annotation was transferred by homology to a further 48 EC classes (demonstrating how rarely the annotation transfer crosses EC classes). Around half the structurally represented EC classes are annotated therefore, though only a tenth of the ∼4000 defined EC classes are covered. The coverage of the EC is shown in Figure 3.7.

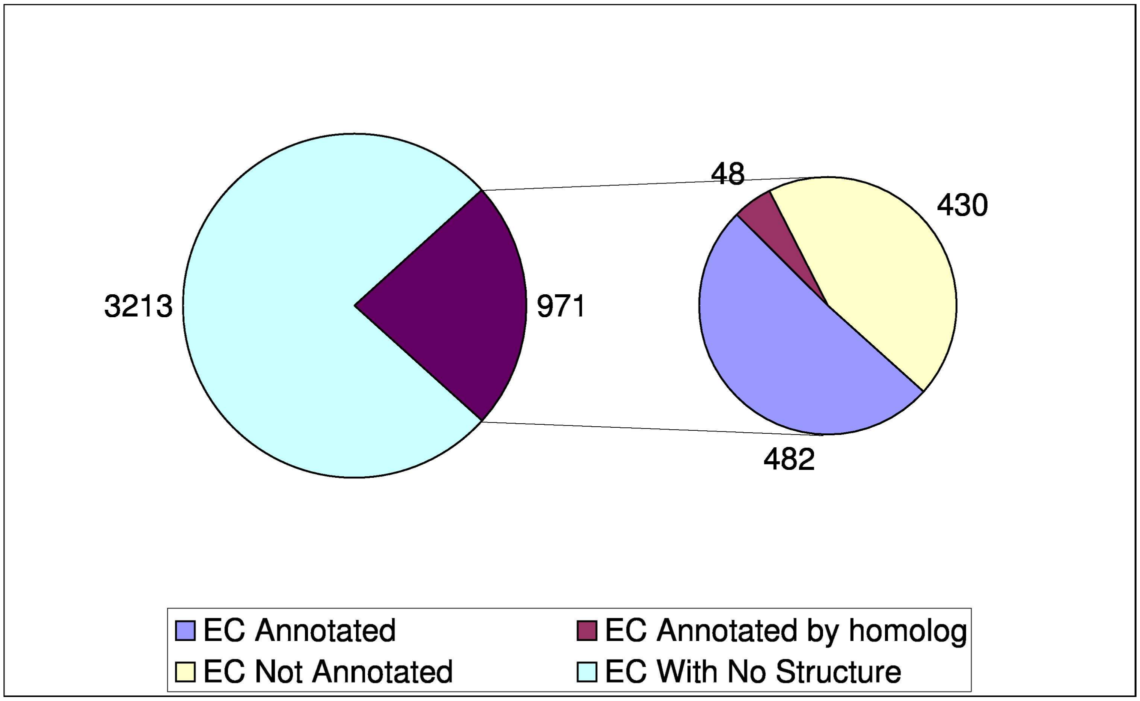


Figure 3.7: Coverage of the EC by the CSA as of June 2003.

3.3.2 Bias and Redundancy

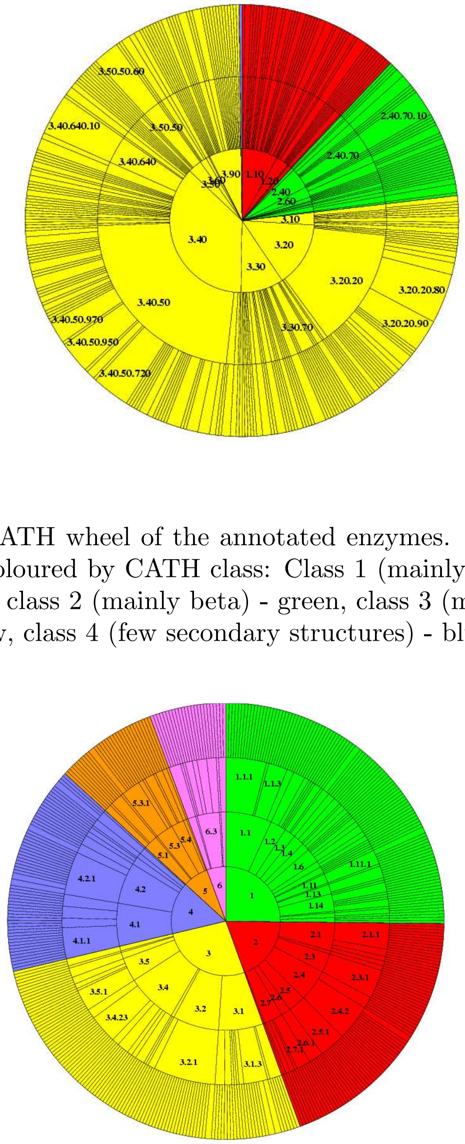
It is well known that there is not a simple one-to-one relationship between EC classes and evolutionary families of enzymes. Some families catalyse many different EC classes[33], while others only catalyse one. Similarly, some EC classes are catalysed by more than one different family of enzymes while others are not.

The enzymes chosen in this annotation effort were largely done so on the basis of achieving a high EC class coverage, so this means that the dataset is not non-redundant, and several of the annotated enzymes are related to each other. These related entries are redundant as far as subsequent analyses are concerned, because they will bias any results towards populated families and so have to be excluded.

We use the CATH classification to examine the evolutionary redundancy of the dataset. A CATH wheel, shown in Figure 3.8(a) shows the distribution of the annotation across the CATH database allowing us to visually see the level of redundancy in the CSA. The lowest level of the CATH wheel (closest to the centre) is split into sectors, with each one representing a different architecture (the second level of the CATH hierarchy, the top level is shown by the colouring). The next level out shows the topologies, with the final outermost layer divided according to homologous super families. We can see, therefore, that while many families are represented just once, many are also over-represented. The most over-represented families are the Type I PLP-dependent aspartate aminotransferase like (CATH 3.40.640.10), TIM barrels (3.20.20.80 and 3.20.20.90) and Rossmann-like domains (CATH 3.40.50.720, 3.40.50.950 and 3.40.50.970). The 493 annotations made can be divided into 437 separate CATH superfamilies. We can also see from Figure 3.8(a) that the mixed alpha/beta class of structures (coloured yellow) is by far the most commonly annotated. This largely reflects the distribution of enzyme structures in the PDB, rather than any bias in annotation.

It is also possible for there to be some redundancy in the EC classes annotated since a single EC class can include multiple families, each of which may catalyse the same reaction by quite different means. An EC wheel is shown in Figure 3.8(b). Each division in the outer most layer represents a single EC class. We can see that of the 493 annotations, there are 11 cases which share an EC number. Each of the six top level EC divisions are represented, though the first three classes (oxidoreductases, transferases and hydrolases) have more examples annotated than the other classes. Again, this reflects the distribution of enzymes with known structure in the EC classification rather than any bias in annotation.

This expanded CSA dataset, made non-redundant by filtering out homologues using CATH, is used in subsequent chapters. However, in many cases the directly annotated enzyme structure is not used, since structures onto which the annotation has been transferred may be of a higher resolution than the original, or have otherwise interesting properties.



(a) CATH wheel of the annotated enzymes. Sectors are coloured by CATH class: Class 1 (mainly alpha) - red, class 2 (mainly beta) - green, class 3 (mixed) - yellow, class 4 (few secondary structures) - blue.

(b) EC wheel of the annotated enzymes. Sectors are coloured by the top level of the EC classification: Class 1 (oxidoreductases) - green, class 2 (transferases) - red, class 3 (hydrolases) - yellow, class 4 (lyases) - blue, class 5 (isomerases) - orange, class 6 (ligases) - pink.

Figure 3.8: CATH and EC wheels of the annotated enzymes.

Chapter 4

Conformational Change at the Active Site

4.1 Introduction

In this chapter we aim to understand more about the nature of conformational change in the catalytic cycle, which as well as being of interest in its own right, has significance for subsequent analyses, which may be made more complex if large changes are routinely observed in the structure of different forms of the same enzyme.

All proteins, and hence enzymes, are inherently flexible molecules because of the many, relatively weak non-covalent bonds that define their folded 3D structure. So, while crystal structures make it convenient to think of a protein as existing in a single state, in reality a protein exists in a range of conformations often with relatively small energy differences between them. In enzymes the differences between these conformations can have important functional consequences. For instance, the conformation that can bind the substrate, may be different to that required for catalysis to occur.

There are a number of different types of motions undergone by protein structures, which take place on different time and length scales.

- Breathing: Experiments measuring the rate of exchange of buried backbone NH hydrogen atoms show that even buried groups are exposed to the solvent by nanosecond scale vibrations in the protein structure.

- Segmental flexibility and hinges: Some parts of the structure can be flexibly attached to the rest of the protein by hinge regions. The regions that move can themselves be flexible loops (a few residues in length) or whole domains that move as a rigid block. These motions occur on timescales between nanoseconds and microsecond for loop motions, up to milliseconds for whole domain motions.

- Side chain rotation: Many side chains rotate at rates between 1 and 10^5 Hz. Surface amino acids may have no unique conformation.

In this work, we only consider the motions of flexible segments that we can observe in different crystal structures, rather than the more dynamic breathing and side chain motions (though we do briefly consider side chain motions).

The importance of conformational change in the study of enzyme catalysis became clear when Daniel Koshland's theory of induced fit [113] replaced the Lock and Key theory as the main model of enzyme action. The Lock and Key theory, which describes both substrate and enzyme as rigid bodies, could not explain certain aspects of enzyme kinetics, particularly allosteric interactions and co-operativity. These effects can only be understood if enzymes are capable of changing their substrate affinities and hence changing their active site structure.

The induced fit theory proposes a general mechanism whereby an 'open' form of the enzyme binds the substrate, and in doing so closes around the substrate into a ‘closed’ form. Catalysis takes place in the closed form, before the enzyme opens again to release the product. This motion of open to closed is proposed to fulfil a number of requirements for the enzyme [114]. Firstly, if the active site is not preformed, conformational change can arrange the catalytic residues of the enzyme into the correct orientation for catalysis to take place. The substrate may also be held in a more restricted conformation when enclosed by the enzyme, thus reducing the entropy of the substrate and any entropic cost of the catalytic reaction. Closing the active site also prevents intermediates escaping before the reaction has completed, and stops solvent from entering and reacting with fragile intermediate species.

Although induced fit provides many potential benefits to an enzyme, the process is in fact catalytically unfavourable because it reduces k_{cat}/K_M. This is because energy is required to distort the enzyme on substrate binding, thus raising K_M without raising k_{cat}. Since proteins are inherently flexible, these distortions are likely to have a low energy cost, which may make change worthwhile if they provide the benefits mentioned above.

Previous studies have categorised the types of motion seen in proteins according to the size of the fragment involved, and the type of motion observed. Fragment sizes are divided into regions smaller than domains (loops, small secondary structures), domains, and finally, whole subunits. These three groups form the highest level of the classification used by Gerstein *et al*’s MolMovDB database of macromolecular motions [115].

The type of motion observed can also be divided into a few simple classes [116]. Hinge motions are characterised by the motion of two fragments towards each other by pivoting around a flexible hinge region. The effect of this is to create a new interface between the two fragments, and the motion of the two fragments is perpendicular to this interface. The classic example of small hinged motion in an enzyme is triosephosphate isomerase (TIM), in which a small loop closes the active site [117]. Hinged domain motion is demonstrated by many kinases, where two large lobes move towards each other when the substrate binds [118].

In contrast to hinge motion, where a new interface is formed, shear motion is characterised by a sliding motion of two fragments parallel to an already existing interface. An example of shear motion in enzyme catalysis is aspartate aminotransferase[119]. It is also possible to have motion based around the unfolding and refolding of secondary structure. Dihydrofolate reductase for instance, has an active site loop which is found in three distinct secondary structures: a 3_{10} helix, β-sheet and loop [120]. Some enzymes show motions which have components of all these different types.

Knowing the extent of conformational change undergone by enzymes is important in other applications, besides our understanding of catalysis. Firstly, ligand docking is a problem with great scientific and commercial interest because of its importance in drug discovery [121, 122]. One of the difficulties in making accurate docking calculations is factoring in the induced fit motion of the enzyme. Calculating large scale motions of the peptide backbone is computationally intractable with current methods [123]. Knowing the extent of conformational change in enzymes in general will help assess the importance of this problem for studying ligand docking in enzymes, and help formulate strategies for solving it.

Secondly, structural genomics projects are now starting to provide significant numbers of structures for proteins for which there is little or no functional annotation [48, 49]. One of the strategies used for annotating these structures is the use of small structural templates [124, 57]. In enzymes the most obvious strategy is to build templates from the catalytic residues of already annotated structures. However, a potential problem with using templates in this way, is that if most enzymes undergo significant rearrangement of their catalytic residues on ligand binding, a single static template will be insufficient to describe all the potential crystal forms that might be observed. An example of this problem is illustrated in Figure 4.1, which shows the catalytic residues of UDP-N-Acetylglucosamine enolpyruvyl transferase (EPT). EPT has four catalytic residues: Arg 397, Asn 23 and Asp 305 shown at the bottom of Figure 4.1, and Cys 115 which lies on a flexible loop shown in two different conformations. In the apo form the loop is open, allowing the substrate to bind in the space between the loop and the other three catalytic residues. Once the substrate has bound, the cysteine residue moves ∼10Å to seal the active site. Also note that Arg 397 rotates its side chain to interact with the cysteine. A template built from the catalytic residues in the apo state would not match a substrate bound structure since their relative positions change so dramatically.

Knowing the extent of conformational change in most enzymes will allow template builders to judge how much induced fit needs to be taken into account, either by removing flexible residues from their templates, or by constructing more complex templates capable of describing flexible residues in a multitude of different conformations.

4.2 Methods

4.2.1 Compiling the Dataset

To study the conformational changes due to ligand binding requires a set of enzyme structures with accurate annotation on the type of ligand bound in each. Unfortunately, the PDB does not provide such annotation directly, though it may describe bound ligands in some detail, it does not label them as to whether they represent

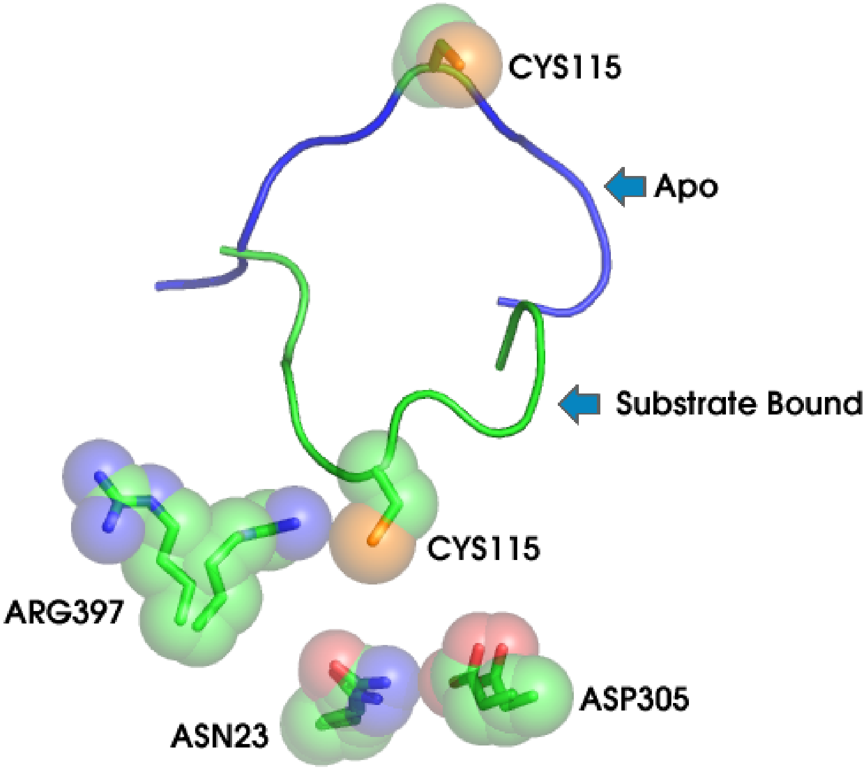


Figure 4.1: Superposition of the apo(PDB: 1EJD [125]) and substrate bound (PDB: 1UAE [126]) forms of EPT. The four catalytic residues are shown in sticks and transparent spacefill, the loop carrying the mobile Cys 115 is shown in both the apo and substrate bound conformations.

Total Enzymes	297
Enzymes with Apo structures	225
Enzymes with structures with some or all substrates	214
Enzymes with structures with transition states	48
Enzymes with structures with some or all products	89
Enzymes with structures with unclassified ligands	81
Enzymes with Apo and All Substrates structures	109
Enzymes with Apo and All Substrates and All Products	31
Enzymes with Apo and All Substrates and All Products and Transition State	6

Table 4.1: Numbers of enzyme structures in the conformational change dataset

substrates, transition states, products or some other ligand. It was necessary, therefore, to annotate as many of the enzymes in the CSA with this information. Each PDB file directly annotated in the CSA was taken as a starting point and all the PDB files with identical sequences were retrieved. Each of these files as well as the original was examined and labelled according to the type of the ligands bound. The following classes were used: apo (representing those structures with no bound ligand), some substrate, all substrates, transition state, all products, some products and other (where the ligand bore no chemical resemblance to any of the natural ligands). Where the ligand was a substrate/product analog it was examined to see whether it was close enough to be used as a substitute for the real substrate or product.

4.2.2 Size of the Dataset

A break down of the numbers of structures annotated is shown in Table 4.1. Two analyses were performed on this dataset: The first on the 31 enzymes where apo, substrate bound and product bound structures were available (in order to examine complete reaction cycles) and the second on the 109 where apo and substrate bound structures were available (to examine conformational change on substrate binding in more detail).

Before the analysis, the dataset was further refined by filtering using a resolution cutoff of 2.5Å and removing any homologous enzymes by filtering by CATH superfamilies. This resulted in 11 enzymes with apo, substrate and product structures; and 58 with apo and substrate structures.

Catalytic residues were extracted from the CSA and binding residues were defined as any residue with any atom placed within 4Åof a ligand atom in any of the ligand bound states.

4.2.3 Measuring Conformational Changes

To measure the extent of conformational change between two states of an enzyme we use root mean square deviation (RMSD). RMSD is measured using a superposition of Cα atoms performed using the ProFit program[127]. RMSD provides a convenient, simple measure of the changes between two structures. However, it is dependent on the number of atoms used to make the superposition and the physical size of the enzyme, so comparisons of RMSDs between enzymes have to be made with care.

To compare the flexibility of the catalytic and binding regions to the rest of the protein, we compare the changes (measured in RMSD) undergone by these functional residues with a large set of randomly chosen residues from the non-functional parts of the structure. For each set of functional residues, a sphere is placed over their geometric centre such that all their Cα atoms fit within the sphere. Each group of random residues is chosen so that it has the same number of residues as the functional region and the residues lie within a sphere with a radius within 10% of the functional residues sphere. 1000 different randomly selected groups are chosen from each structure and the RMSD is calculated for each. The percentile-rank (P) of the RMSD of the functional residues within the 1000 random samples is then calculated. A P-value of 1 means that the functional residues undergo the most conformational change in the structure, whilst a P-value of 0 means they undergo the least amount of change.

4.3 Results

4.3.1 Catalytic Cycles

The 11 enzymes and the PDB files that have apo, substrate bound and product bound structures are summarised in Table 4.2. For each enzyme the structure of the apo form (labelled E), substrate bound form (labelled 'ES') and product bound form (labelled 'EP') are known. In protein farnesyltransferase (FTase) and dihydrofolate reductase (DHFR) the apo form is not part of the usual catalytic cycle, instead the binding of fresh substrate causes the release of product from the previous cycle. In both cases the structure with both substrate and product is known (labelled ESP). In thymidylate synthase (TS) the folate substrate binds in an unreactive conformation, and then opens to form a reactive complex. This reactive structure is labelled as ES*.

Conformational Changes Seen

The observed conformational changes between states are categorised into four different types of motion:

1. Loop motions: Movements of small (2-10 residues) segments of structure.
2. Domain motions. Movements of protein domains.
3. Side Chain rotation. Rotation of side chains which alters the position of the functional atoms of the side chain.

4. Secondary structure change.

In addition, it should be noted that only those motions that effect the conformation of the active site are reported. All of the enzymes show side chain rotation and loop motions in some parts of the structure, but these are not considered significant if they are not part of the active site. It may be that these motions have other roles such as allostery or are simply background noise.

Methylmalonyl-CoA mutase (MUT) is unique in this set of enzyme in the extent of the conformational change it undergoes, with some residues moving over 10Å when the substrate binds. The conformational changes include loop and domain movements, which form the substrate binding $(\alpha\beta)_8$ barrel and close the active site to exclude solvent.

DHFR also undergoes important conformational change. It uses the 'MET20' loop to control access to the NADPH binding pocket of the active site. As well as moving to block and cover the NADPH binding site, the MET20 loop undergoes changes in secondary structure, changing from disordered in the apo form, to β-sheet in the closed form, a 3_{10} helix in the occluded form and an ordered loop in the open form. There is also domain motion in DHFR: the two sub-domains move towards each other by $\sim$1Å, closing the active site cleft around the bound substrates.

As well as MUT and DHFR, significant movement of active site loop regions are seen, on substrate binding, in TS, dethiobiotin Synthase (DTBS), FTase and cytochrome P450cam (P450cam). These movements are characterised by a general closing of the active site, with the surface loop regions moving in towards the rigid core of the protein, closing over the bound substrate.

TS uses the C-terminal tail, rather than a loop, to close the active site. The tail moves to cover the active site, serving to exclude the bulk solvent and trap intermediates. DTBS uses three small loops to bind the substrate and ATP, with motions of ∼2Å . FTase undergoes very small loop motions in the active site (∼1Å), though these small motions are still important, since one of these loops contains the catalytic lysine which is moved into the correct alignment with the substrate. P450cam undergoes a small rearrangement of the peptide mainchain to allow binding of a catalytic water molecule. This motion is classified as a loop motion here, though it is much smaller than the other motions.

There are significant rotations of functional side chains in DTBS, MUT, DHFR, peptide deformylase (PDF) and chalcone synthase (CHS). In DTBS and DHFR, catalytic residues, a threonine and methionine respectively, rotate so that their side chain is correctly placed near the substrate. In CHS and PDF, hydrophobic residues, a phenylalanine and leucine respectively, rotate so as to form close Van Der Waals contacts with the substrate and intermediates. In MUT, TYR89 rotates so as to push the adenosyl cofactor off the cobalt atom to which it is attached, which generates the adenosyl radical required for the reaction.

In general the changes seen between the substrate and product bound forms are smaller than that seen between the apo and substrate bound forms. There are only obvious functional changes in P450cam, TS and PDF. In P450cam the bound water molecule accommodated in the substrate bound form, is not present in the product bound form, and the peptide mainchain moves back into its previous conformation. In TS and PDF, there are suggestions that steric clashes between the product and the enzyme may encourage product release. In TS the extra methyl group of dTMP clashes with a bound water molecule, and in PDF an active site leucine rotates such that the side chain replaces the leaving formyl group.

Motion of The Functional Residues

For each enzyme, the binding residues are defined as any residue with any atom within 4Å of a bound substrate atom and the catalytic residues are taken from the CSA.

The catalytic residues of four of the enzymes are shown in Figure 4.2. We see in Figure 4.2A the multiple conformations of the catalytic phenylalanine side chain in PDF. This is an example of side chain rotation without loop motion. In Figure 4.2B we see a change in the peptide bond joining the catalytic aspartate and threonine residues. This moves the mainchain carbonyl oxygen ∼2Å and allows binding of an ordered water molecule. In Figure 4.2C we see an example of side chain rotation and loop motion. The catalytic threonine of DTBS moves ∼1Å and rotates to correctly position the hydroxyl group. Finally, in Figure 4.2D we see the active site of MUT, which undergoes the largest conformational change in this sample. There are large domain and loop motions as well as side chain rotation of the catalytic tyrosine which acts to generate the adenosyl radical required to start the reaction.

Reaction Cycles

For each enzyme, we construct a simple reaction cycle comprising the apo, substrate bound and product bound forms. In FTase and DHFR, product release is only achieved by binding of fresh substrate and so the apo form is not part of the normal in vivo reaction cycle; in these cases the enzyme-substrate-product complex is used instead of the apo form.

We measure the RMSD using all the Cα atoms between each structure in the cycle. Figure 4.3 shows these RMSDs by drawing a triangle for each enzyme such that each vertex represents a structure in the cycle and the length of the edge connecting two vertices is proportional to the RMSD between those two structures.

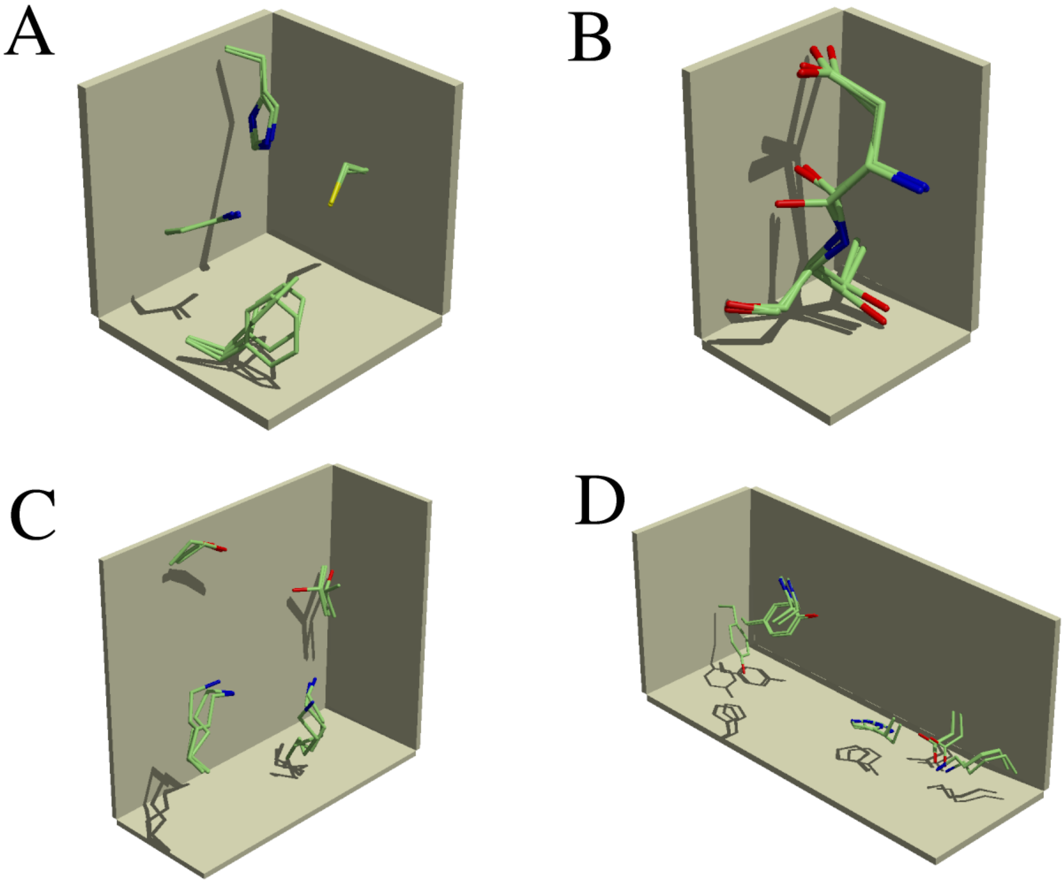


Figure 4.2: Four examples of superpositions of catalytic residues from different stages of the reaction cycle. A: The catalytic residues of PDF, the multiple conformations of the phenylalanine can be seen at the bottom. B: P450cam, the central carbonyl oxygen can be seen in two different conformations which allows binding of an ordered water molecule in the substrate bound form. C: DTBS, the threonine in the top right moves ∼1Å on substrate binding and rotates to correctly position the hydroxyl. D: MUT, all the residues move, but the tyrosine on the left hand side is the most important, rotating as well as moving to generate the adenosyl radical.

Enzyme	PDB Files					Motion Observed				Notes
	E	ES	ES*	EP	ESP	Loop	Domain	Side Chain	2^o Structure	
Napthalene Dioxygenase	1O7H	1O7N	-	1O7P	-	-	-	-	-	No significant motion observed [128].
Hal2p	1K9Z	1KA1	-	1KA0	-	-	-	-	-	No significant motion observed [129, 130].
Aconitase	1AMJ	1C96	-	1C97	-	-	-	-	-	No significant motion observed [131, 132].
Chalcone Synthase	1BI5	1CML	-	1CGK	-	-	-	●	-	PHE215 side chain rotates through a number of different conformations [133].
Peptide Deformylase	1BS5	1LRU	-	1BS8	-	-	-	●	-	LEU91 rotates to occupy the space left by the leaving formate group [134, 135].
P450cam	1PHC	1DZ8	-	1NOO	-	●	-	-	-	Mainchain carbonyl oxygen of ASP251 moves to accommodate a bound water molecule on substrate binding [136].
Protein Farnesyltransferase	1FT1	1D8D	-	1KZP	1KZO	●	-	-	-	LYS164 moves ∼1Å on substrate binding, small motions in other loop residues. [137].
Thymidylate Synthase	3TMS	1BJG	2TSC	1TYS	-	●	-	-	-	C-terminal tail moves to cover active site [138, 139, 140, 141, 142].
Dethiobiotin Synthase	1BYI	1DAH	-	1DAF	-	●	-	●	-	THR11 moves ∼1Å and rotates on substrate binding, also other loop movements [143, 144].
Methylmalonyl-CoA Mutase	2REQ	4REQ	-	6REQ	-	●	●	●	-	Large loop and domain motion on substrate binding, TYR89 rotates also [145, 146].
Dihydrofolate Reductase	5DFR	1RA2	-	1RX4	1RX6	●	●	●	●	MET20 loop adopts four conformations: unordered, closed, open and occluded. Subdomain motion acts to open and close the active site [120].

Table 4.2: Enzymes and PDB files used in the analysis. E is the resting state of the enzyme with no ligands bound. ‘ES’ is the structure with substrate bound. ‘ES*’ is the structure with activated substrate bound. ‘EP’ is the structure with product bound. ‘ESP’ is the structure with product and fresh substrate bound.

In most cases, the RMSD is small (less than 1Å), reflecting the localised nature of the motions observed. Usually conformational change is restricted to small loop regions, only a few residues in length, moving to close around a rigid core. Only MUT, with RMSD ~2.5Å ,and to a lesser extent DHFR, TS and DTBS, with RMSD ~1Å undergo large scale motions. In these enzymes the motion between the apo and substrate bound forms is much greater than the motion between the substrate bound and product bound forms. In contrast to those enzymes undergoing smaller motions, where the three sides of the triangle are usually similar in length.

To show the motion of the functional residues, in Figure 4.4 the RMSD is calculated using a superposition of the Cα atoms of the binding residues only, and in Figure 4.5 the RMSD is calculated using only the Cα atoms of the catalytic residues.

The pattern seen in Figures 4.4 and 4.5 is similar to that seen in Figure 4.3. A few enzymes (MUT, DTBS and TS) undergo relatively large motions of the binding and catalytic residues, whilst the remainder are relatively static. As with the whole protein RMSDs, the difference between substrate bound and product bound forms is small in most cases, though in both P450cam and DHFR the product bound state has moved closer to the apo form.

RMSDs and P Values

The distribution of the RMSDs and the P-values (calculated as described earlier) for the catalytic residues are shown in Figure 4.6 and for the binding residues in Figure 4.7. In each case the distribution of RMSDs for the apo-substrate bound and product bound-apo pairs peak at low RMSD, with an extended tail representing the examples of large conformational change. The substrate bound-product bound pair does not have a tail of high RMSD changes and tends to small values only, reflecting

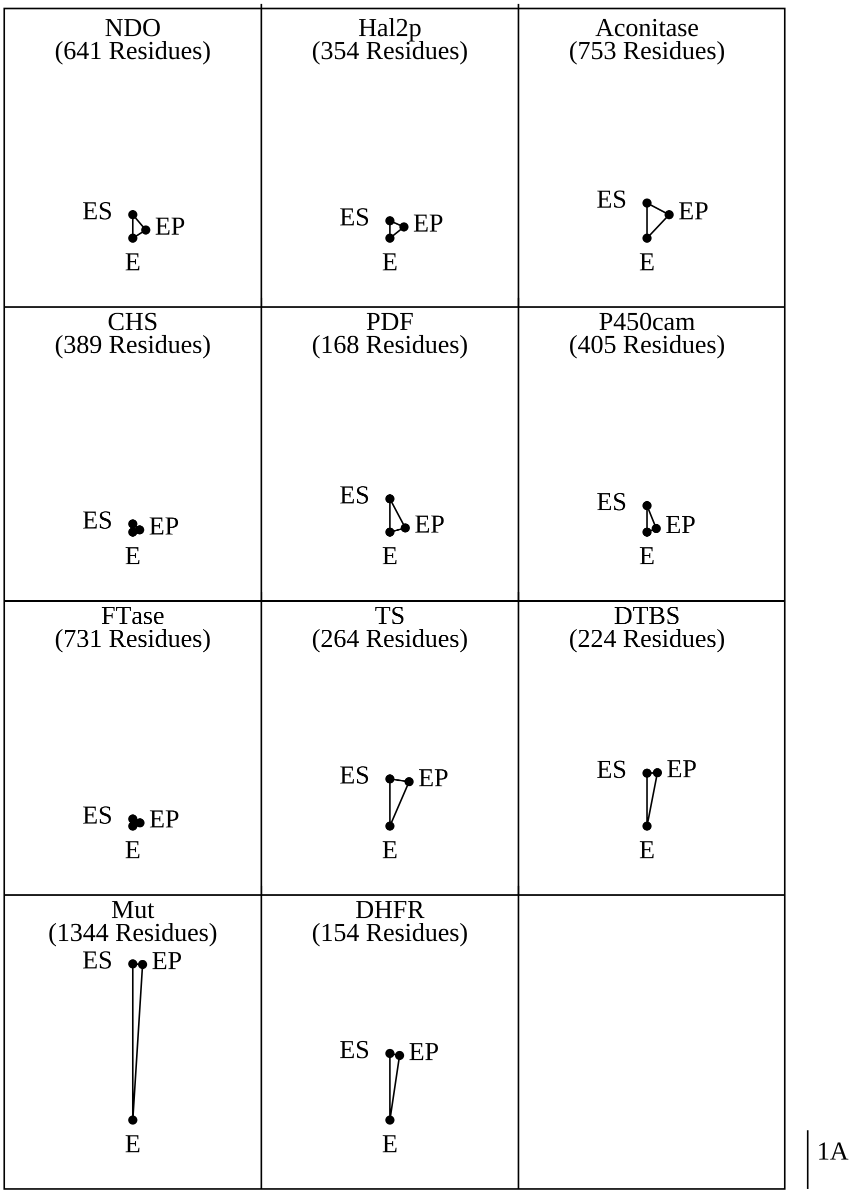


Figure 4.3: Triangles representing the conformational change undergone by each enzyme. Each side of a triangle represents the RMSD between the structures represented by each vertex. RMSD is calculated by performing a superposition of Cα atoms. For DHFR and FTase, the structure with product and fresh substrate bound is used instead of the apo form, as this better represents the *in vivo* enzyme cycle. For TS the structure with activated substrate is used as the structure bound form. A line representing a 1Å RMSD is shown as scale.

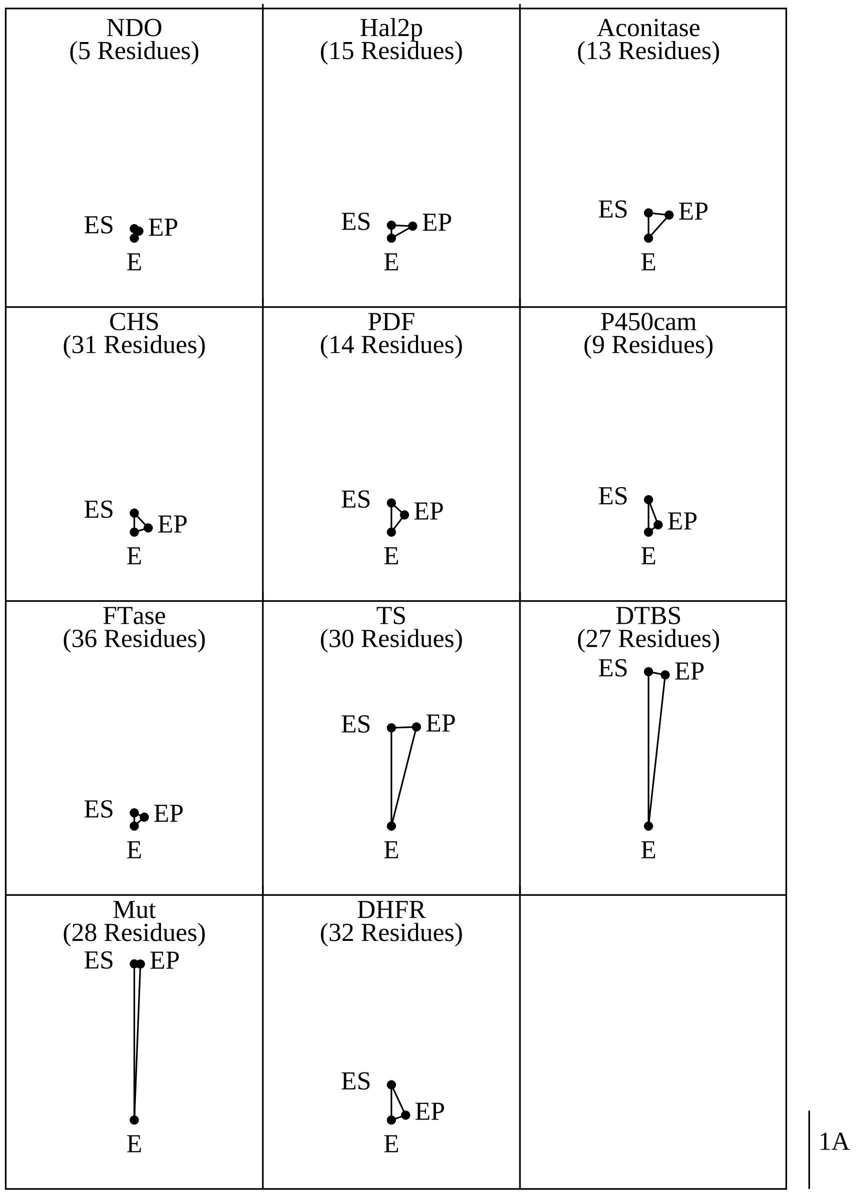


Figure 4.4: Triangles representing the conformational change undergone by the binding residues of each enzyme. Each side of a triangle represents the RMSD between the position of the binding residues in the structures represented by each vertex. RMSD is calculated by performing a superposition of Cα atoms. For FTase and DHFR, the structure with product and fresh substrate bound is used instead of the E form, as this better represents the *in vivo* enzyme cycle. For TS the structure with activated substrate is used as the structure bound form. A line representing a 1Å RMSD is shown as scale.

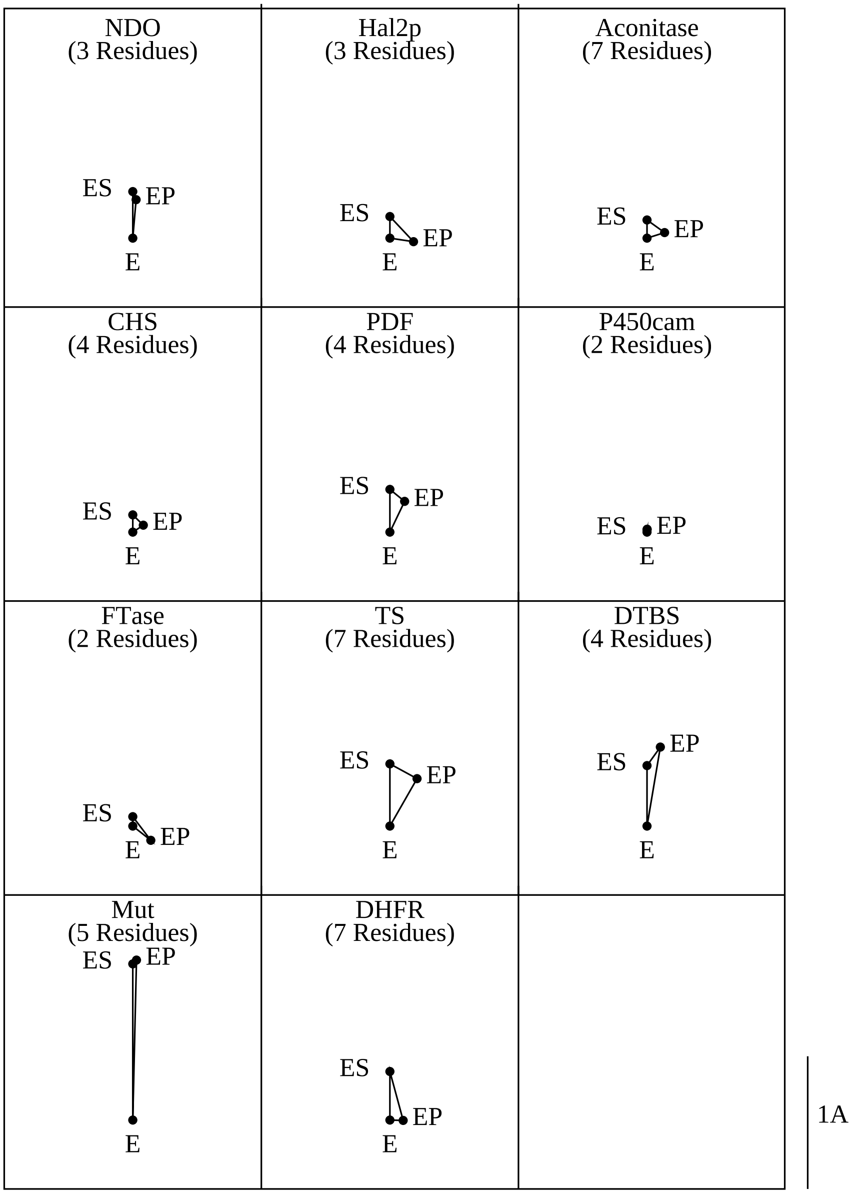
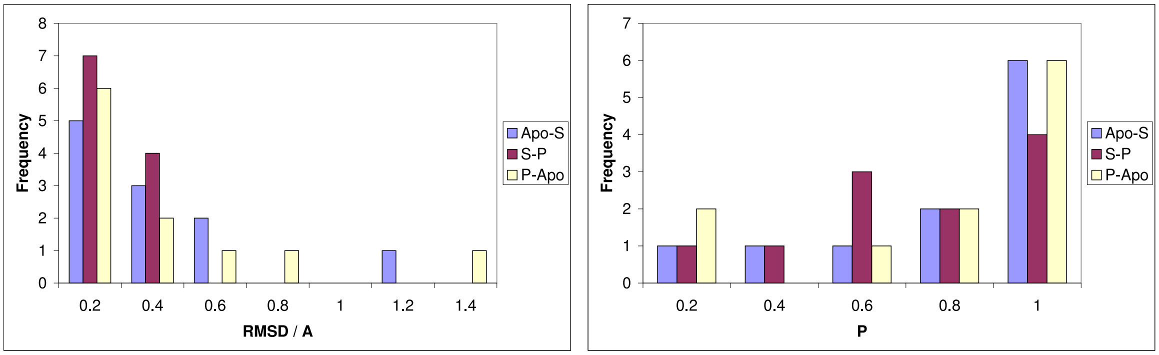


Figure 4.5: Triangles representing the conformational change undergone by the catalytic residues of each enzyme. Each side of a triangle represents the RMSD between the position of the catalytic residues in the structures represented by each vertex. RMSD is calculated by performing a superposition of Cα atoms. For FTase and DHFR, the structure with product and fresh substrate bound is used instead of the E form, as this better represents the *in vivo* enzyme cycle. For TS the structure with activated substrate is used as the structure bound form. A line representing a 1Å RMSD is shown as scale.

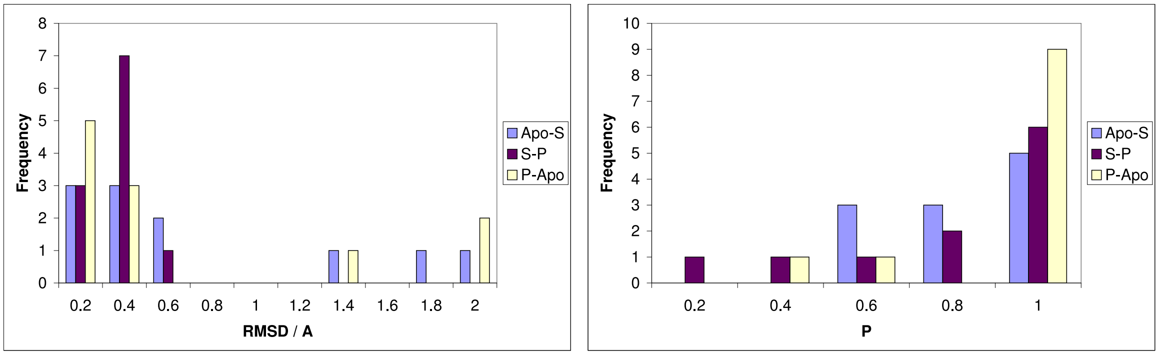
the consistently small changes we see between the substrate bound and product bound forms. The P values show that both the catalytic and binding residues are often amongst the most flexible parts of the structure, despite the small values of the RMSDs. However, we do see a few examples where, even in the apo-substrate bound transition, the catalytic residues are amongst the most rigid parts of the structure.



(a) RMSDs for catalytic residues.

(b) P values for catalytic residues.

Figure 4.6: Graphs of RMSD and P for the catalytic residues.



(a) RMSDs for binding residues.

(b) P values for binding residues.

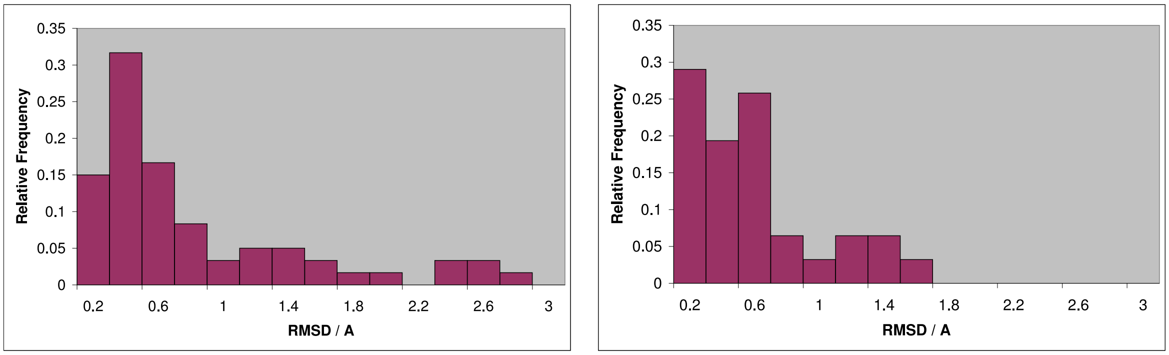
Figure 4.7: Graphs of RMSD and P for the binding residues.

4.3.2 Substrate Binding

RMSDs Observed

To improve our picture of the events associated with substrate binding we now analyse the larger dataset of enzymes where the apo and substrate bound forms are known. Figure 4.8(a) shows a histogram of the RMSDs observed calculated from superpositions over all the Cα in each enzyme. 75% of the enzymes have an RMSD less than 1Å, and 91% have an RMSD less than 2Å, between the apo and substrate bound forms. The average RMSD observed is 0.7Å. From inspecting the data by eye, it seems that a reasonable rule of thumb is that RMSDs >2Å represent those enzymes undergoing either significant domain motions, or very large loop motions, it would appear therefore, that motions such as this are relatively rare, and seen in around 10% of cases.

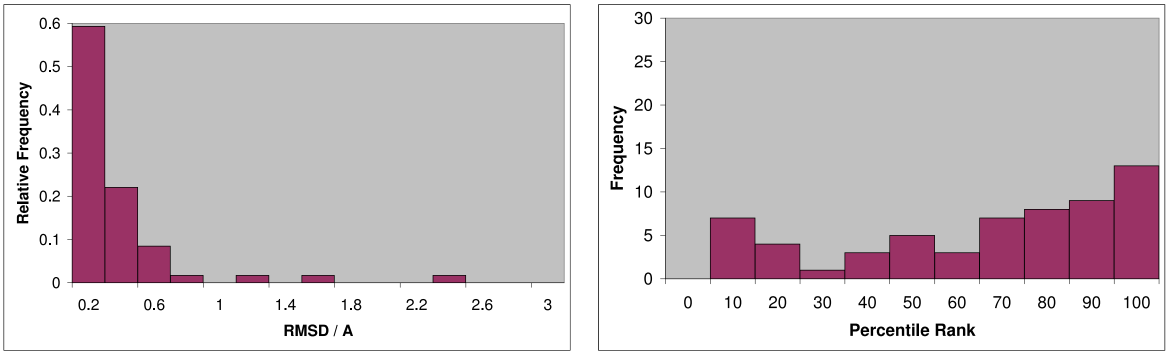
To put this in context, we compare the RMSDs obtained from comparing apo with substrate bound structures with a second data set made of pairs of apo structures. This dataset of 31 enzymes is smaller because not all the enzymes in the original dataset have two separate apo structures solved. All apo crystal structures were considered, so some pairs may have been solved by the same investigators, whilst others will have been solved independently. Where more than two apo structures of the same enzyme were solved, two were chosen at random from those available. The RMSDs obtained from comparing two apo structures for a single enzyme are shown in Figure 4.8(b). The average RMSD is 0.5Å, only a little smaller than the 0.7Å found between apo and substrate bound structures. Comparing Figure 4.8(a) and Figure 4.8(b), we see that many of the motions seen when comparing the apo and substrate bound forms are no larger than those seen when comparing two apo structures of the same enzyme.



(a) Apo to substrate bound

(b) Apo to apo

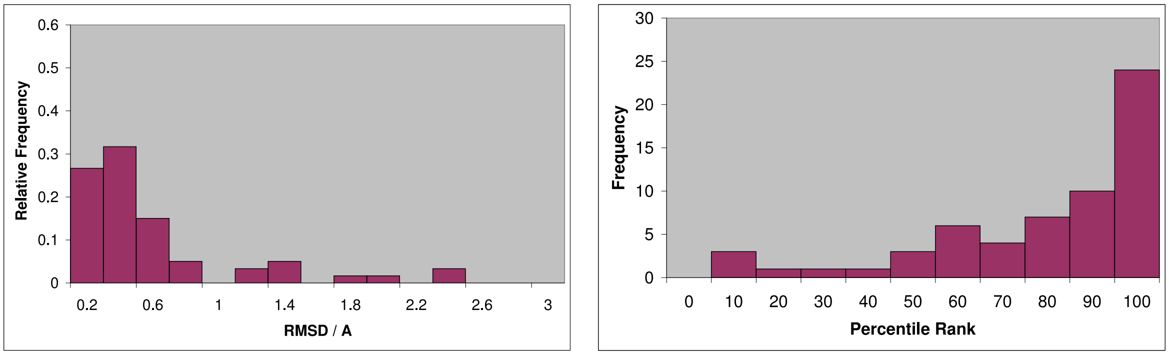
Figure 4.8: (a) Histogram of the RMSDs seen between the apo and substrate bound structures over all residues. (b) Histogram of the RMSDs seen between two apo structures of the same enzyme over all residues.



(a) Catalytic residue RMSDs

(b) Catalytic residue P-values

Figure 4.9: (a) Histogram of the RMSDs seen in the catalytic residues between the apo and substrate bound structures. (b) Histogram of the P-values for the catalytic residues compared to the rest of the structure between the apo and substrate bound structures.



(a) Binding residue RMSDs

(b) Binding residue P-values

Figure 4.10: (a) Histogram of the RMSDs seen in the binding residues between the apo and substrate bound structures. (b) Histogram of the P-values for the binding residues compared to the rest of the structure between the apo and substrate bound structures.

The distribution of RMSDs for the binding and catalytic residues are shown in Figure 4.9(a) and Figure 4.10(a). For both the binding and catalytic residues the RMSDs tend to be less than 1Å, though there is a tail in the distribution going up to 8Å in both cases. 13 out of the 60 enzymes (22%) in this dataset have RMSDs larger than 1Å for the binding residues, and 5 (8%) have RMSDs larger than 1Å for the catalytic residues. Inspecting these examples by hand shows that in all of these cases the high RMSD is due to a single residue moving on a flexible loop towards a stationary catalytic core, as demonstrated in EPT shown in Figure 4.1.

Figures 4.10(b) and 4.9(b) show the P-values for the catalytic and binding residues respectively. Both regions were found to undergo more conformational change than the rest of the protein, with average P-values of 0.74 for the binding residues and 0.61 for the catalytic residues. The larger P-value for the binding residues suggests that the binding residues undergo more motion than the catalytic residues. The difference between the distribution of P-values for the binding residues and catalytic residues is found to be significant using a Wilcoxon signed rank test;

with a probability of it occurring by chance of 2×10^{-2}.

4.3.3 Absolute Motion

To further examine the difference between the binding and catalytic residues we also measure the absolute motion of residues upon substrate binding, rather than the relative motions given by the RMSDs. Absolute motion is determined by superposing the structure as before, and then, for each residue, measuring the distance between the position of the Cα atoms in the apo and substrate bound structures. Figure 4.11(a) is a relative cumulative frequency plot, showing the proportion of residues that move less than a given distance. The non-functional and catalytic residues have an almost identical distribution, but the binding residues show a significant difference over the range 0.5Å-3Å. For instance, 20% of binding residues move more than 1Å, compared to only 10% of catalytic and non-functional residues.

4.3.4 Side Chain Motion

For most catalytic and binding interactions it is the terminal side chain atoms which are the key functional groups. All residues apart from glycine, proline and alanine are capable of moving the terminal side chain groups through rotation of side chain bonds. This means that the previous analysis, based on Cα atoms, may be missing the complete story.

To measure the motion of the side chain atoms a superposition was made on the four main chain atoms of each residue (C, O, N, Cα) in the apo and substrate bound forms. The distance between the geometric centre of the functional atoms (for instance, the carboxyl group of aspartate) in the apo and substrate bound forms was then measured. Figure 4.11(b) shows the cumulative relative frequency of the motions of all the catalytic, binding and non-functional residue side chains. In this plot we observe that the side chains of the binding residues and catalytic residues move a similar amount, unlike the backbone motions. In both cases 20% of residues move their side chains more than 1Å, and 15% more than 2Å. Both sets of residues move their side chains slightly more than the non-functional residues.

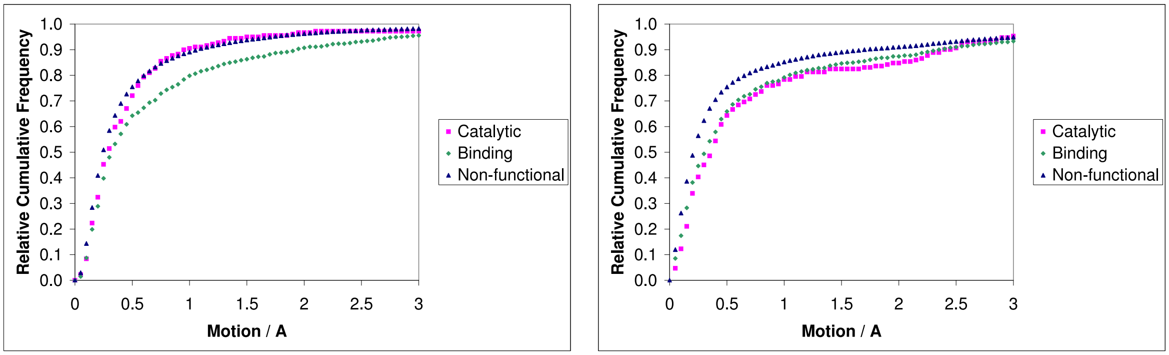


Figure 4.11: (a) Relative cumulative frequency plot of the motion undergone by the Cα atoms of catalytic, binding and non-functional residues. (b) Relative cumulative frequency plot of the motion of the functional atoms for the catalytic, binding and non-functional residues.

4.4 Conclusions

4.4.1 Overall Results

One of the surprising results of this survey is the small size of the motions undergone by most of the enzymes. Of the 11 enzymes in the first dataset, only DHFR and MUT undergo large conformational change. DTBS and TS have a few loop regions which undergo localised changes, but the remaining seven enzymes undergo only very subtle conformational changes. This pattern is repeated when we look at the larger dataset which shows that many changes observed on substrate binding are of a similar magnitude to what we observe when comparing two different apo structures of the same enzyme. We also note that the catalytic residues show a different pattern of motion from the binding residues, with significantly less backbone motion, though a similar amount of side chain motion. This suggests that the catalytic site is often pre-formed, though the binding site may not necessarily be.

Although the motions observed are often small, they are not insignificant when considered in the context of catalysis and catalytic mechanisms. Mesecar *et al* [147, 148] demonstrate that movements in the active site of fractions of angstroms can alter the rate of catalysis by an enzyme by many orders of magnitude. They present evidence that perturbations in the active site of isocitrate dehydrogenase (IDH) lead to a change in the distance between the cofactor NADP and the isocitrate substrate of 0.56Å. This small change in distance leads to a reduction in the reaction rate of almost 3 orders of magnitude. We expect therefore, that even the small motions observed in most of the active sites here will have important consequences in catalysis. It is clear that one of the difficulties of studying enzyme catalysis using structural biology, is that very accurate structures (both in terms of resolution and similarity of ligand analogs to the natural ligand) are required for a full understanding of the catalytic mechanism.

Bias in the Dataset?

One possible bias in this dataset, is that the structures we have chosen, are self-selected to be those which undergo minor conformational changes. There are two reasons for this: firstly rigid enzymes are likely to be easier to crystallise in multiple states, and secondly, in enzymes which do undergo large change, it may be that soaking the ligand into the apo form destroys the crystal, and so crystal structures cannot be obtained from these enzymes. In the first dataset, six of the eleven enzymes had structures generated by soaking the ligand into crystals of the apo form, and five by direct crystallisation of the ligand bound form.

Another potential problem with the analysis is that those enzymes which undergo a transition from a disordered to ordered state upon substrate binding will not be represented in this dataset. Disordered regions, such as substrate binding loops, do not generally appear in the crystal structure and so will be ignored by this analysis.

It is also possible, that the available structures are only revealing part of the story. Certainly in two cases: P450cam and aconitase, the crystal structures cannot be giving the full picture, as the active site is entirely closed even in the apo form. In these enzymes, there must be hidden flexibility in the structure that the static snapshots given by X-ray crystallography do not detect. However, we believe that in the majority of cases the crystal structure does represent a true picture of the enzyme at that point in the reaction cycle.

4.4.2 Motion and Energy

Do the triangle diagrams of the catalytic cycle tell us anything about the Gibbs free energy profiles of these enzymes? There are a number of difficulties with interpreting the diagrams this way: firstly, the diagrams only deal with the conformational change of the enzyme, and so can only describe the reaction from the proteins point of view, which forms just one component of the total free energy. Also the RMSD of a conformational change is not proportional to the energy involved in making that change. Unravelling an alpha helix, for example, may take more energy than moving a surface loop, though the RMSD may well be greater for the loop motion.

Given these caveats, what would we expect to see in terms of conformational change from an energy perspective? The free energy profile found by Knowles and Albery[31] for triose phosphate isomerase (TIM), shows that the apo form has the lowest free energy, followed by the substrate bound species, followed by the product bound species. Given this pattern, and making the large assumption that conformational change is related to free energy change, we would expect to see a triangle with EP-E as the longest side, representing the large energy change between the product bound and resting states. Looking at Figure 4.3 we don't see any examples of triangles with this shape. In fact, what we see is that in those enzymes which do undergo significant induced fit (TS, DTBS, MUT and DHFR), the ES-EP side is short, leading to E-ES and EP-E sides of similar length. For enzymes undergoing smaller conformational change, some have a similar length for all three sides, whilst others have short EP-E sides. The RMSDs of the binding residues shown in Figure 4.4 show a similar pattern.

These observations suggest that the increase in free energy from the ES to the EP form is not spent in further conformational change of the enzyme, but in some form of strain on the substrate. In this study, we have only analysed the motion of the enzyme, not the changes undergone by the substrates, so we do not observe this aspect of the reaction. It is clear that in some cases, the conformational change of the substrate is extremely important. In FTase for example, the two substrates are bound with the acceptor carbon of the farnesyl and the attacking cysteine sulphur of the peptide 7Å apart. The reaction must, therefore, involve substantial motion of the substrates during catalysis. The enzyme itself, appears not to have moved between the substrate bound and product bound states, but we cannot rule out the possibility that there is motion between the static snapshots given by the crystal structures.

4.4.3 Implications for Other Applications

Our results indicate that modelling side chain flexibility in ligand docking applications is an important step forward, as many enzymes show significant side chain motion in both the binding and catalytic residues, though we note that most residues move the functional atoms less than 1Å. Looking at the backbone motion we observe that the majority of enzymes do not undergo large domain or loop motions. However, even the relatively small motions that we do see (<1Å), can have a significant effect on docking scores and predicted binding conformations. Progress still needs to be made in accurately modelling these types of conformational changes.

The other application of relevance for this study, is the use of structural templates as annotation tools. Our results show that building stable templates from the side chain atoms of catalytic residues is likely to be harder than building them from Cα atoms. However, templates built purely from Cα positions may not discover functional similarities between unrelated (or distantly related) enzymes such as the serine proteases subtilisin and trypsin. Template builders will have to take into account these two considerations: the relative immobility of Cα atoms, and the functional importance of the side chain atoms; when constructing templates, and choose appropriately.

4.4.4 Stability vs induced fit

The final question we consider is whether there is a conflict between the requirement for enzymes to precisely position their functional groups and the conformational change required by theories of induced fit. Intuitively, it would seem harder to precisely position residues that are in a flexible part of the enzyme than residues in the rigid core. An obvious solution to this apparent contradiction is to restrict induced fit motions to surface loops which can close over the catalytic machinery located at the base of the active site. Looking at the P-values for the catalytic and binding residues, which measure the conformational change seen in these residues compared to the rest of the protein, we find that the average P-value for the catalytic residues is 0.61 compared to 0.74 for the binding residues. This suggests that the binding residues are more flexible than the catalytic residues. There is also a suggestion of this in the results obtained by Bartlett *et al*[76] which showed that catalytic residues generally have small solvent exposure and so do not lie on surface exposed loops.

Chapter 5

Structural Features of Catalytic Side Chain Interactions

5.1 Introduction

In this chapter we will investigate the roles of interactions between amino acid side chains in enzyme catalysis. The interactions of side chains are extremely important when considering any aspect of protein structure, because of their role in determining both structure and function. Side chains can interact in a number of different ways: firstly the larger aromatic and aliphatic side chains form hydrophobic contacts with each other[149], secondly, oppositely charged groups can form salt bridges[73], thirdly, polar residues can form hydrogen bonds[150], and lastly, cysteine residues can form covalent links between one another[151]. These different interactions all constrain the geometry of the interacting residues in some way that is usually structurally significant and can, on occasion, be functionally significant as well.

Although most side chain interactions serve a structural or stabilising purpose, in catalytic systems they can also perform functional roles by modifying the chemical

Group	pK$_a$ Model compounds	 Usual range in proteins
Asp (CO_2H)	3.9	2-5.5
Glu (CO_2H)	4.3	2-5.5
His (imidazole)	6.4	5-8
Cys (SH)	8.3	8-11
Lys (NH2)	10.8	∼10
Tyr (OH)	11	9-12
Arg (guanidine)	12.5	-

Table 5.1: Sample pK$_a$ values for ionizable side chains. Data taken from 'Enzyme Structure and Mechanism' (Freeman)[13]

properties of the interacting residues. The most obvious example of a property effected in this way is the pK$_a$ of a side chain. pK$_a$ is defined in terms of the ionization constant K$_a$, defined in Equation 5.1.

$$K_a = \frac{[B][H^+]}{[BH^+]} \tag{5.1}$$

Where $[B]$ is the concentration of a basic group, $[H^+]$ is the proton concentration and $[BH^+]$ is the concentration of the protonated base. The pK$_a$ is then defined in Equation 5.2.

$$pK_a = -logK_a \tag{5.2}$$

The outcome of this is that pK$_a$ is the pH at which the residue is half-protonated. Sample pK$_a$ values for ionizable residues are shown in Table 5.1. However, the pK$_a$ values observed in model compounds can be significantly altered by the environment in which a residue finds itself in. An example of this effect is the interaction of two carboxylate groups [152, 153]. In these interactions the two carboxylates overcome their mutual repulsion to form a hydrogen bond. The presence of a negative charge on one of the residues raises the pK$_a$ of the other, allowing it to be protonated at higher pH and hence, a bond can form between the two. Several enzymes, such as HIV-1 protease, use such pairs to perform catalysis.

Another early example of this phenomenon was found in acetoacetate decarboxylase[154], where two adjacent lysines mutually destabilise their protonated forms because of their proximity, allowing one of them to act as a nucleophile. Similarly, the well known serine proteases use a triad of interacting residues to perform their chemistry[27]. However, not all combinations of residues will be useful: some combinations may have no effect or even reduce the power of their component residues. Using an analogy with a toolkit: a hammer and chisel is a useful combination, whilst a hammer and a saw is a poor one.

5.2 Method

5.2.1 Generating Data

We have compiled from the literature a set of 191 enzymes to study. The set is non-redundant, with no two enzymes being evolutionarily related as defined from sequence and structure comparisons in the CATH database. The catalytic mechanism for each enzyme is extracted from the CSA and the crystal structure from the PDB. All the structures are high quality (<2Å resolution, <0.3 R-factor) and the catalytic residues are found in a single conformation (all atoms have occupancy 1). To find interacting residue, we define an interaction as taking place between two residues if any of their side chain atoms are within 4Å of each other. We only consider interactions between polar residues and ignore interactions involving the hydrophobic residues.

Hydrophobic residues such as phenylalanine, tryptophan and the smaller aliphatic residues do have important effects on catalytic residues and mechanisms. By providing a non-polar environment for instance, they tend to raise the pK_a of acidic residues and lower the pK_a of basic ones. However, these effects (also known as medium or solvent effects)[155] are often spread out over the whole, or parts of, the active site and are, therefore, hard to localise to a specific residue. For this reason we only consider interactions between polar or charged residues. We also restrict our attention to side chain groups, though main chain amides and carbonyls are also important polar groups, used most famously in the oxyanion hole of serine proteases.

We find that, on average, each polar catalytic residue interacts with 0.4 other polar catalytic residues and 1.9 other polar residues (giving a total of 2.3 interactions per catalytic residue). In contrast, non-catalytic buried polar residues have interactions with 1.2 other polar residues on average, significantly less than the catalytic residues. Only 88 of the 191 enzymes contain one or more interactions between two of the defined catalytic residues, suggesting that most catalytic residues do not require direct interactions with other catalytic residues to be active. However, the fact that the catalytic residues have a larger number of interactions than non-catalytic residues, suggests that at least some of these interactions between catalytic and non-catalytic residues are functional. The annotation in the CSA is derived from literature searching using strict criteria for defining a catalytic residue. Despite this, it seems likely that some of these secondary interactions, between residues annotated as catalytic and residues annotated as non-catalytic, do have a role to play in catalysis. These residues, annotated as non-catalytic, have not been previously identified, because their effect is likely to be subtler than other residues that are directly involved in the mechanism.

5.2.2 The Functions of Secondary Residues

The simplest of the functions performed by these secondary residues is orientation. Making bonds between residues restricts their motion and so ensures that they are positioned correctly relative to the substrate. Since restricting motion reduces entropy, there is an energetic cost to this orientation. By pre-arranging the active site, this entropic cost is paid for during the folding of the enzyme rather than during catalysis, and so is beneficial to the enzyme. The individual contribution of a single residue whose only role is to orientate another residue is likely to be small, but the effect of taking all such residues in an enzyme together will be significant.

Where charged groups are required to interact with a substrate, there may well be secondary groups that stabilise the charge required by providing oppositely charged groups nearby. In many cases these residues have already been annotated as catalytic. However, in some enzymes the importance of these residues may be less obvious, and so they have may have been overlooked.

In the case of histidine, secondary residues which control the tautomerisation of the imidazole ring may be important. Histidine exists in two neutral tautomeric states, protonated on either the Nδ or Nϵ atoms. Free in solution, these two forms exist in roughly equal proportions, but in an enzyme it is essential that the correct tautomer is preferred.

5.3 Results

5.3.1 Common Catalytic Dyads

It is expected that certain residue interactions will be seen much more than others. There are two effects which contribute to this: Firstly certain residue interactions occur more often in protein structures than others since, for instance, two oppositely charged residues are more likely to interact than two residues of the same charge. Secondly, certain residues have a much higher propensity to be catalytic than others and so are over-represented in the catalytic set. Histidine, for instance, is common in active sites and so interactions involving histidine are especially numerous.

Given this, are any interaction pairs seen significantly more or less than we would expect? To answer this, we need to define what we would expect to observe. First we need to take into account the fact that certain residue pairs are seen more often than others throughout protein structures (aspartate-arginine interactions for instance). We measure the relative propensity for two residues to interact using Equation 5.3.

$$I_{i,j} = \frac{\sum Non_{i,j}}{\frac{\sum Non_i}{\sum Non} \times \frac{\sum Non_j}{\sum Non} \times 2 \times \sum Non} \tag{5.3}$$

Where i and j are two residue types, $\sum Non_{i,j}$ is the total number of observed non-catalytic interactions between residue i and residue j, $\sum Non_i$ is the total number of i residues observed and $\sum Non$ is the total number of non-catalytic residues (of all types) in the dataset. We multiply the proportions of each residue i and j by two because an i-j interaction is equivalent to a j-i interaction. A propensity greater than 1 means that a pair of residues is observed interacting more often than we would expect if the residues interact at random.

To take into account the fact that some residues have a high catalytic propensity we use Equation 5.4 to calculate the expected number of interactions observed if the catalytic residues combined randomly. By combining this with the observed non-catalytic interaction propensity we get Equation 5.5 which we use to calculate the expected number of interactions if the residues in the catalytic dataset interacted in the same way as the non-catalytic residues.

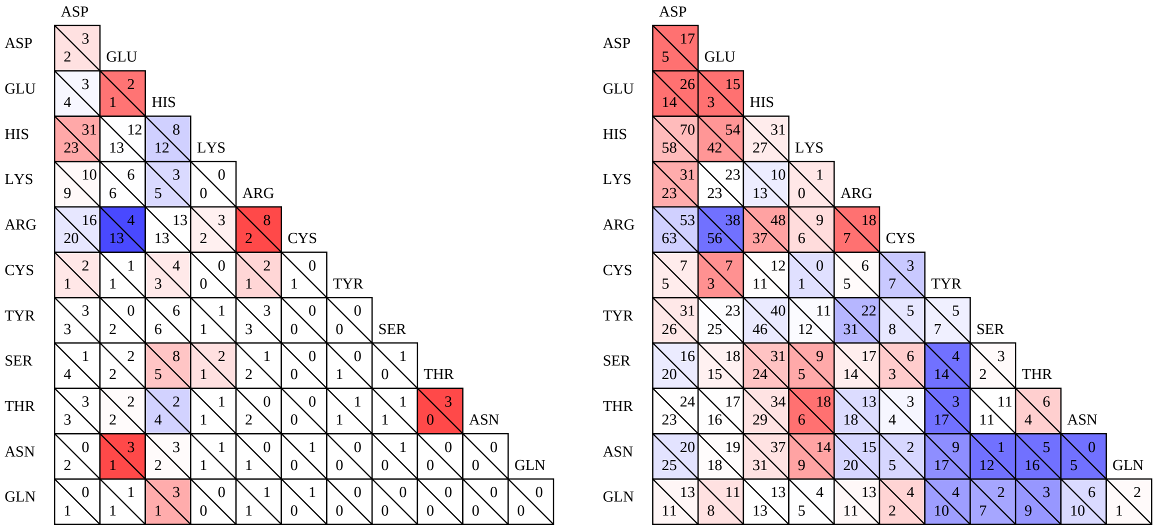
$$E_{i,j} = \frac{\sum Cat_i}{\sum Cat} \times \frac{\sum Cat_j}{\sum Cat} \times 2 \times \sum Cat \tag{5.4}$$

$$E_{i,j} = \frac{\sum Cat_i}{\sum Cat} \times \frac{\sum Cat_j}{\sum Cat} \times 2 \times \sum Cat \times I_{i,j} \tag{5.5}$$

Figure 5.1(a) shows the numbers of interactions observed between each pair of residues, where both residues are annotated as catalytic. It is clear, that many potential interactions, particularly between the polar residues (clustered in the bottom right of the diagram) are rarely observed in active sites. In contrast, interactions between charged residues, and to a lesser extent between polar and charged residues, are more common. It is also noticeable that certain combinations such as histidine-aspartate are observed much more than expected, whilst others such as arginine-carboxylate interactions are observed much less.

Figure 5.1(b) shows the number of interactions where only one of the residues is catalytic, there are many more of these interactions, and, as explained above, we expect many to be functionally important even though they have not been annotated as such. The larger numbers allow us to see more general trends: not only are the polar-polar interactions rare, they are also observed much less often than we would expect. In contrast, we again see a large number of interactions between charged groups. Some, like carboxylate-carboxylate and arginine-arginine interactions are observed more often than we would expect, whilst others, like arginine-carboxylate interactions are observed less than we would expect.

We call these commonly observed interactions ‘catalytic units’. Each catalytic unit is a small (two or three residues) group of interacting residues which fulfil a particular catalytic function such as acid/base chemistry or nucleophilic attack.



	ASP	GLU	HIS	LYS	ARG	CYS	TYR	SER	THR	ASN	GLN
ASP	3 2										
GLU	3 4	2 1									
HIS	31 23	12 13	8 12								
LYS	10 9	6 6	3 5	0 0							
ARG	16 20	4 13	13 13	3 2	8 2						
CYS	2 1	1 1	4 3	0 0	2 1	0 1					
TYR	3 3	0 2	6 6	1 1	3 3	0 0	0 0				
SER	1 4	2 2	8 5	2 1	1 2	0 0	0 1	1 0			
THR	3 3	2 2	2 4	1 1	0 2	0 0	1 1	1 1	3 0		
ASN	0 2	3 1	3 2	1 1	0 1	1 0	0 0	1 0	0 0	0 0	
GLN	0 1	1 1	3 1	0 0	1 1	1 0	0 0	0 0	0 0	0 0	0 0

(a)

	ASP	GLU	HIS	LYS	ARG	CYS	TYR	SER	THR	ASN	GLN
ASP	17 5										
GLU	26 14	15 3									
HIS	70 58	54 42	31 27								
LYS	31 23	23 23	10 13	1 0							
ARG	53 63	38 56	48 37	9 6	18 7						
CYS	7 5	7 3	12 11	0 1	6 5	3 7					
TYR	31 26	23 25	40 46	11 12	22 31	5 8	5 7				
SER	16 20	18 15	31 24	9 5	17 14	6 3	4 14	3 2			
THR	24 23	17 16	34 29	18 6	13 18	3 4	3 17	11 11	6 4		
ASN	20 25	19 18	37 31	14 9	15 20	2 5	9 17	1 12	5 16	0 5	
GLN	13 11	11 8	13 13	4 5	13 11	4 2	4 10	2 7	3 9	6 10	2 1

(b)

Figure 5.1: (a) Counts of residue interactions in the dataset where both residues are annotated as catalytic in the CSA. (b) Counts of residue interactions in the dataset where either residue is annotated as catalytic.

Numbers above the diagonal lines in each box are the observed number of interactions, numbers below the diagonal lines are the expected number of interactions. The expected number of interactions takes into account the catalytic propensities of each residue and the propensity for a given pair of residue types to interact in non-catalytic regions of protein structure.

Boxes are coloured such that red boxes represent interactions that are over-represented and blue boxes represent those that are under-represented. The colouring is done by calculating $\frac{(O-E)^2}{E}$ (χ^2) for each box (where O equals the observed number of interactions and E equals the expected number. The deepest blue and red boxes have $\chi^2 \geq 5$ with $O < E$ and $O > E$ respectively. Boxes where χ^2 is 0 are left clear. Other boxes are scaled between these two extremes according to the χ^2 value. In Figure (a) boxes where $O < 2$ are also left clear to improve clarity.

5.3.2 Common Catalytic Units and their functions

5.3.3 Arginine-Arginine

The 8 catalytic arginine-arginine interactions we find come from 5 different enzymes: arginine kinase, flavocytochrome c, phytase, undecaprenyl pyrophosphate synthase and adenylate kinase. Four of these five enzymes catalyse reactions involving phosphate chemistry. In each case, the arginines form bonds to the phosphate oxygens and so polarise the phosphate group making it a better leaving group. Figure 5.2(a) shows the arrangement of arginines around an substrate analog in adenylate kinase [156]. The arginines are close enough to destabilise each other until the negatively charged phosphate groups bind. The nearby aspartates, Asp 162 and Asp 163, also provide stabilising negative charges though this may be more important for the global folding and stability of the protein and active site rather than for the mechanism of the enzyme.

5.3.4 Carboxylate/Carboxylate

The effect of placing two carboxylates together is that their pK_a values are raised and so they tend to be protonated at a higher pH than is normal. This prevents the unfavourable interaction of two negative charges and a hydrogen bond can form between the two carboxylates. Carboxylate dyads are used in two particularly important classes of enzymes: aspartic proteases and glycosidases.

Aspartic Proteases

In aspartic proteases both the carboxylates engage in acid/base chemistry. One of the aspartates is protonated at the start of the reaction, because of its raised pK_a, allowing it to donate a proton to the substrate. The second aspartate is unprotonated and so can accept a proton from water forming a nucleophilic OH^- which then attacks the substrate. In the second stage the roles of the two aspartates are reversed, with the first aspartate accepting a proton from the protonated intermediate and the second aspartate donating a proton to the leaving substrate. The active site of cardosin [157] is shown in Figure 5.2(b).

Glycosidases

Glycosidases, such as cellobiohydrolase Cel6A[158] shown in Figure 5.2(c), also use two interacting carboxylates. This interaction raises the pK_a of Asp 226, allowing it to operate as an acid/base, but in contrast to the aspartic proteases Asp 180 operates as a nucleophile. This difference in mechanism is due to the interaction between Asp 180 and Arg 179. This lowers the pK_a of Asp 180 and prevents it from becoming protonated. In aspartic proteases, both aspartates are hydrogen bonded to hydroxyl groups (as shown in Figure 5.2(b)) which do not alter the pK_a of the carboxylates in the same way as the arginine.

5.3.5 Carboxylate/Arginine

Placing these residues together stabilises the charged form of each residue so that neither residue can easily gain or loose protons.

Carboxylate-arginine interactions are often found where either a positive charge (from the arginine) or a negative charge (from the carboxylate) is required to polarise a substrate. This is demonstrated by the use of carboxylates to stabilise the arginines used in adenylate kinase, as described above.

Carboxylate oxygens that act as nucleophiles are also found to interaction with arginines. An example of this is the active site of sucrose phosphorylase[159] shown in Figure 5.2(d). Arg 190 reduces the pK_a of Asp 192 ensuring it is unprotonated and so able to perform a nucleophilic attack on the substrate (an analog of which is shown).

The carboxylate-arginine interaction is also seen as part of a larger unit comprising two carboxylates and an arginine, as demonstrated by the glycosidase described above.

5.3.6 Carboxylate/Lysine

Since the amino group of lysine is usually protonated, it can play a similar role to arginine in its interactions with carboxylate. However, lysine has a lower pK_a than arginine (10 compared to 12) and is found in a neutral state given the correct conditions.

Lysine Containing Triads

A threonine containing triad (analogous to the serine protease triad) is found in L-asparaginase, shown in Figure 5.3(a)[160]. The side chain of Thr 95 is used as the nucleophile and Lys 168 as the acid/base instead of serine and histidine respectively. Asp 96 retains its role in orientating and altering the pK_a of the lysine.

Another lysine-carboxylate containing triad is seen in aldo-keto reductases[161], shown in Figure 5.3(b). However, Tyr 58 is not used as a nucleophile, instead lysine lowers its pK_a so it can act as an acid.

A third triad containing two lysine-carboxylate interactions is found in indole-3-glycerol-phosphate synthase shown in Figure 5.3(c)[162]. Two lysines form salt bridges to a single glutamate. The charged forms of the glutamate and one of the lysines are required to stabilise charges in the transition state. The second lysine engages in general acid catalysis and it is speculated that its reprotonation is mediated by its involvement in the triad.

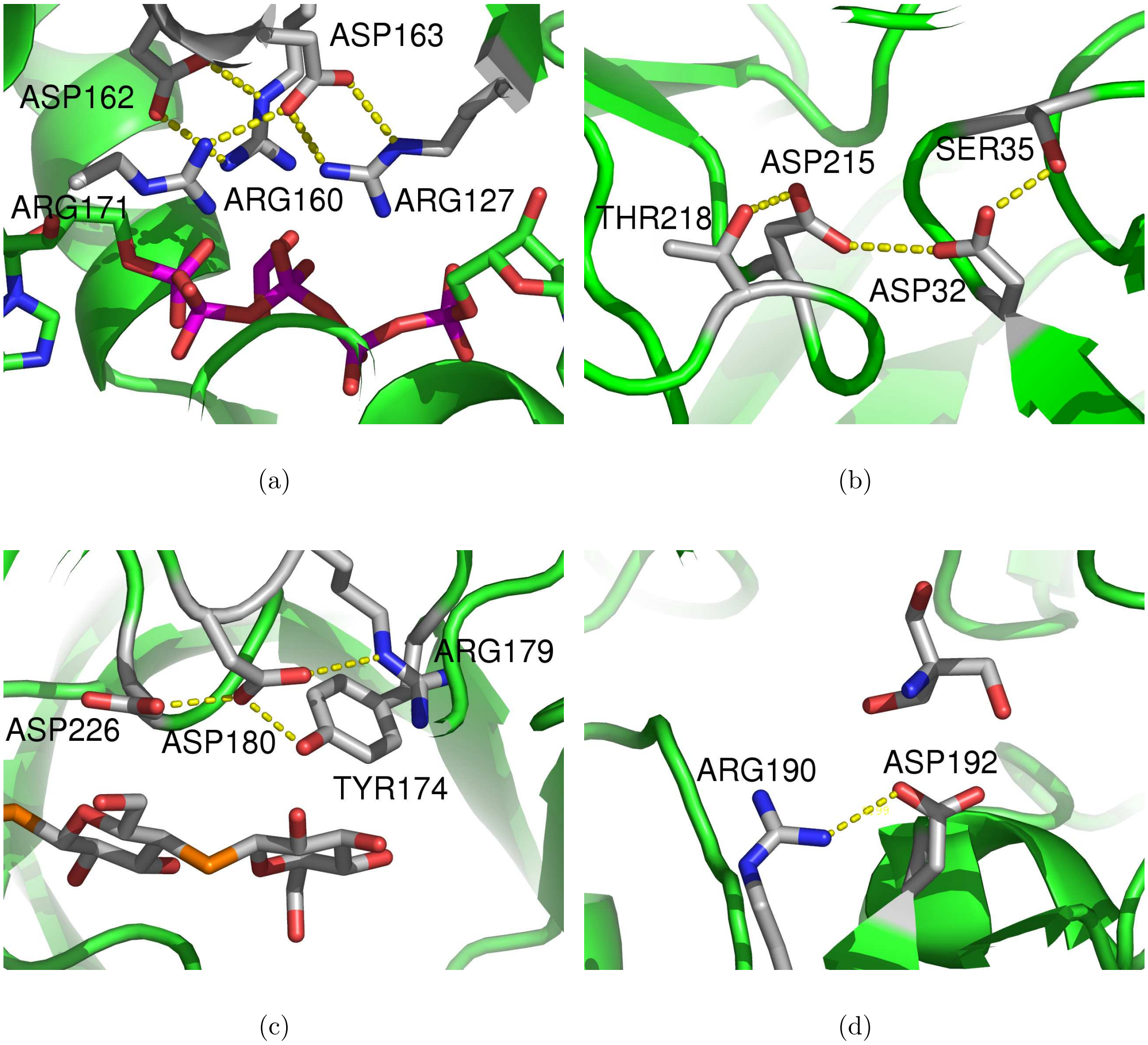


Figure 5.2: Interactions involving arginine and carboxylate. (a) The arginines in adenylate kinase (1ZIN) polarise the substrate phosphates shown in sticks below the arginines. Two aspartates stabilise the concentration of positive charge required. (b) Asp 215 and Asp 32 in cardosin (1B5F) form an interaction which allows them both to act as acid/bases. The hydroxyl groups of Thr 218 and Ser 35 orientate the carboxyls without affecting their pK_a (c) The Asp 180-Arg 179 ion pair in cellobiohydrolase (1OC7) raises the pK_a of Asp 226, which can then engage in acid/base chemistry. (d) Asp 192 in sucrose phosphorylase (1R7A) acts as a nucleophile forming a covalent bond to the substrate, the nearby Arg 190 reduces the pK_a of the aspartate.

Lysine-Carboxylate in Acid/Base Chemistry

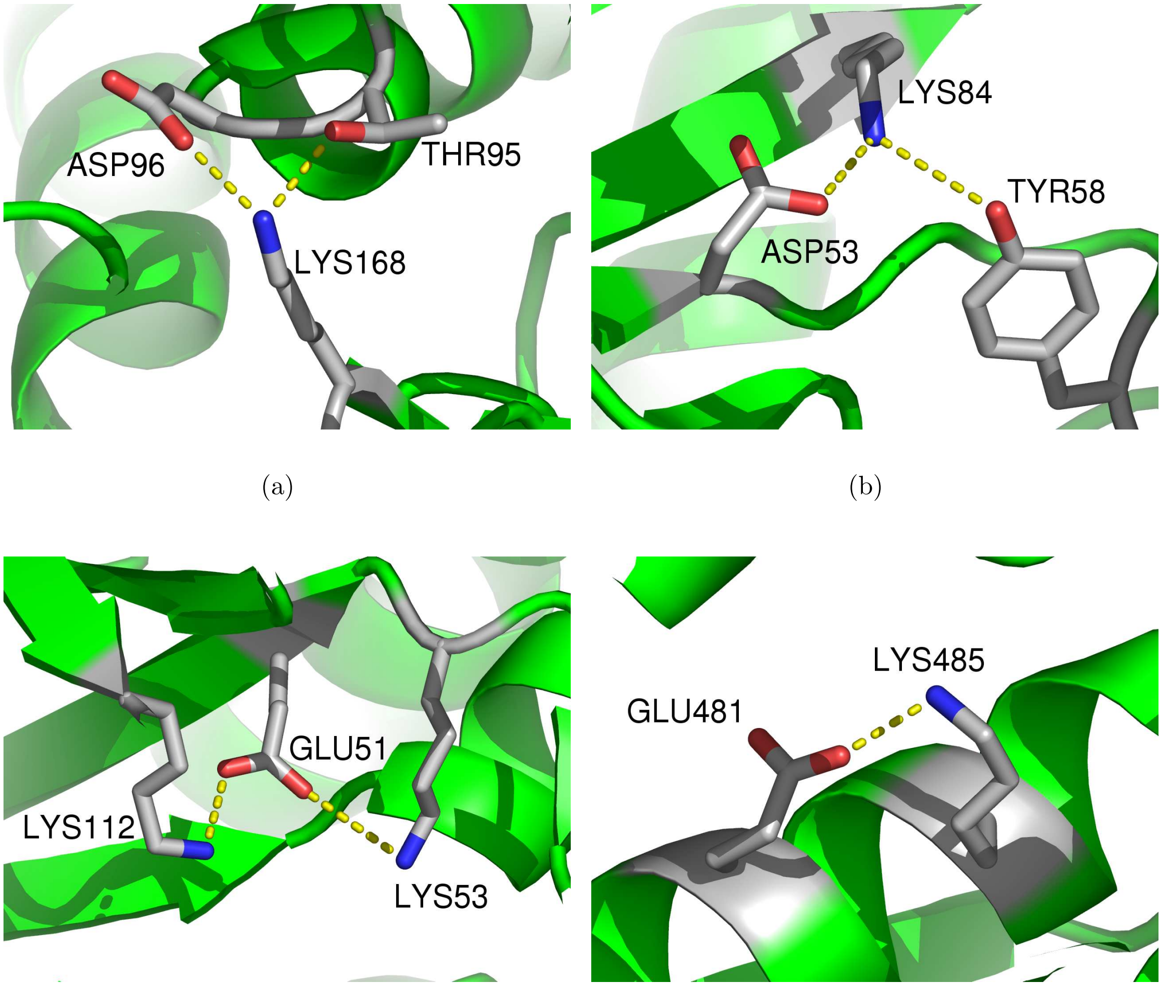
In contrast to arginine which seems to remain protonated at all times, lysine can gain and loose protons. Interactions with carboxylate groups will tend to raise the pK_a of lysine making it less capable of loosing protons. However, we observe several cases where the lysine of a lysine-carboxylate dyad is involved in acid base chemistry. The question is, given its high pK_a, how is the lysine ever deprotonated? Proton relay chains have been proposed for this role, but this is still speculative. Figure 5.3(d) shows the lysine-aspartate dyad from glucosamine-6-phosphate synthase[163], which is proposed to deprotonate a substrate hydroxyl group[164] using the lysine.

5.3.7 Carboxylate/Histidine

The effect of the interaction between histidine and carboxylate is to raise the pK_a of the histidine which helps the histidine act as an acid/base. The classic example of this is the serine protease triad from trypsin shown in Figure 5.4(a)[165] where the histidine-aspartate dyad is used to deprotonate Ser 195. However, most of the examples we see use the histidine-carboxylate dyad acting directly on the substrate. Figure 5.4(b) shows just such a dyad and a substrate analog in the active site of aconitase[132].

5.3.8 Histidine/Hydroxyl

The most well known example of a catalytic histidine-hydroxyl interaction is found in the catalytic triad, that catalyses many different reactions. In the triad, histidine extracts a proton from serine to prime the serine as a nucleophile as described above. However, we also see hydroxyls priming histidines. For instance, in phosphotransferase, shown in Figure 5.4(c)[166], a threonine hydroxyl hydrogen bonds to His 68.



(a) (b)

(c) (d)

Figure 5.3: Interactions involving lysine. (a) The 'asparaginase triad' in L-asparaginase (1O7J) features an aspartate-lysine pair. They are used to activate a threonine as a nucleophile by extracting a proton from the threonine hydroxyl. (b) Another aspartate-lysine containing triad found in aldo-keto reductase AKR11A (1PYF) uses tyrosine not as a nucleophile, but as an acid/base. The lysine-aspartate dyad controls the pK_a of the tyrosine. (c) A triad of two lysines and a glutamate are used in indole-3-glycerol phosphate synthase (1VC4). Lys 112 acts as an acid/base and is believed to have its pK_a modulated by its involvement in the triad. Glu 51 and Lys 53 have additional roles in providing electrostatic stabilisation during the reaction. (d) The lysine in a simple glutamate-lysine dyad is used to perform acid/base chemistry in glucosamine-6-phosphate Synthase (1MOQ).

This hydrogen bond forces the Nδ of His 68 to be unprotonated and the Nϵ to be protonated. This tautomeric state is essential for His 68 in priming His 83 for its role as a nucleophile.

5.3.9 Histidine/Histidine

As the most commonly used catalytic residue it is surprising to see only 8 catalytic histidine-histidine interactions. Furthermore, we find that in all but one of these cases, the two histidines happen to be close enough to each other but the interaction between them does not seem to be functionally important.

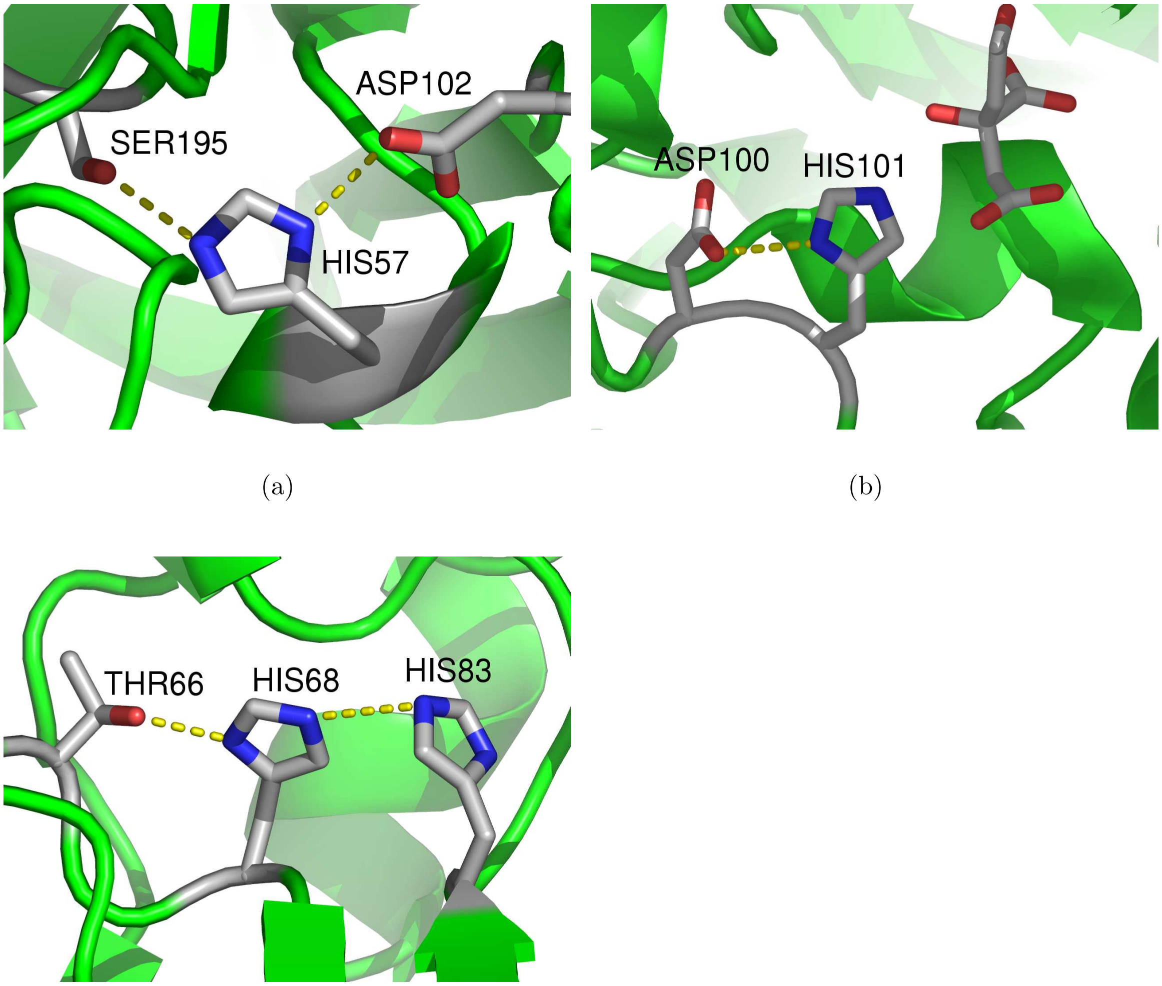
The one exception to this is phosphotransferase mentioned above. His 83 acts as a nucleophile which attacks a phosphate group it is kept in its correct tautomeric state by its bond to His 68.

5.4 Discussion

5.5 The Roles of Catalytic Interactions

Previous studies have shown that ∼40% of catalytic residues are involved in either transition state stabilisation or substrate activation[76]; processes which generally involve simply providing the appropriate charged groups or hydrogen bond partners around the substrate. Given this, it is not surprising that most catalytic residues interact directly with the substrate, rather than each other. Interactions with other groups are not required for most residues to fulfil these types of roles.

However, it also seems that many important catalytic functions are best performed by particular combinations of residues. In some cases, like carboxylate-carboxylate interactions, the effect is to change the chemical character of a group



(a)

(b)

(c)

Figure 5.4: Interactions involving histidine. (a) The classic Ser-His-Asp triad found in trypsin (1AVW), the aspartate-histidine dyad extracts a proton from the serine hydroxyl. (b) The aspartate-histidine dyad from aconitase (1C96) acts as an acid/base directly on the substrate which is also shown. (c) A rare functionally important histidine-histidine interaction is found in the phosphotransferase domain of glucose permease (1GPR). The Nϵ atom of His 83 acts as a nucleophile in attacking phosphate, His 68 ensures that His 83 exists in the correct tautomer and stabilises the transition state. A threonine ensures that His 68 is in the correct tautomeric state.

(from negatively charged to neutral in this case), whilst in others, like arginine-carboxylate interactions, the effect is to enhance already existing properties (by stabilising charges). The range of functions we find performed by interacting residues is summarised below with examples of the groups involved.

- Interactions between like charges
  - Provide charge concentration E.g. Arg-Arg, His-Arg
  - Provide acid/base (by depolarisation) E.g. Asp-Asp.
  - Provide nucleophile E.g. Lys-Lys
- Interactions between opposite charges
  - Stabilisation of charge concentration e.g. Asp-Arg
  - Provide ion pair to depolarise third residue E.g. Arg-Asp-Asp
  - Provide charges for transition state stabilisation e.g. Glu-Lys
  - Provide nucleophile E.g. Arg-Asp
  - Provide acid/base E.g. Glu-Lys
- Interactions between charged and polar residues
  - Provide nucleophile E.g. Lys-Thr
  - Provide acid/base E.g. Lys-Tyr, Glu-Thr, Asp-His
- Interactions between polar residues
  - Provide nucleophile E.g. His-Ser, His-His
  - Provide acid/base E.g. Asn-His
  - Tautomerisation E.g. Thr-His

The conclusion we draw from this, is that the range of roles played by interactions involving charged residues is greater than that played by interactions involving polar residues. The charged residues have important roles in transition state stabilisation and substrate polarisation, but they also have the ability to modify the pK_a of other residues, allowing those residues to perform functions they wouldn't otherwise be able to do.

In contrast, apart from histidine (which is often found charged in protein structures), the polar residues are only catalytically active when combined with a charged residue. These interactions generally involve a charged residue priming the polar residue for action, either as a nucleophile as shown in the threonine-lysine interaction or as an acid/base as shown by aspartate-histidine. A polar residue rarely primes a charged residue, presumably because it has little effect on its properties.

An interesting question in enzyme catalysis is how enzymes catalyse such a diverse range of reactions using its limited toolkit of the polar and charged amino acids, metals, cofactors and water. One explanation could be that combinations of residues are used to provide new chemical tools. This analysis, and the data shown in Figure 5.1 shows that only a few of the different dyads that could be formed in enzyme active sites are actually used in catalysis. The seven combinations we describe above (arginine-arginine, carboxylate-carboxylate, carboxylate-arginine, carboxylate-lysine, carboxylate-histidine, histidine-hydroxyl and histidine-histidine) account for ~65% of the interactions between catalytic residues, and probably an even higher proportion of the really key functional interactions. So, although combinations of residues can produce new or enhanced chemical activity in the residue side chains, the catalytic toolkit used by enzymes seems as small as ever.

The answer to how such a small set of tools can catalyse such a diversity of reactions probably lies partly in the nature of the catalysed reactions themselves.

Results from an analysis of the MACiE database suggest that most reactions can be broken down into individual steps, each of which is chemically simple. For instance, 75% of reaction steps involve a simple proton transfer [Holliday *et al*, Personal communication]. Since there are a restricted number of these simple steps, it follows that the number of chemical groups required to catalyse these steps is also small.

The other part of the answer is that the power of enzymes to catalyse reactions is due to much more than simply providing certain residue or residue combinations close to the substrate. The repeated evolution of some units, such as the catalytic triad, implies that these units are genuinely useful to enzymes. However it would clearly be wrong to ascribe all the catalytic power of any enzyme to the formation of these units. Rather, it is the combination of these very 'local' structural features, along with the more 'global' features of the enzyme, such as the dynamics of the structure and the overall microenvironment of the active site, where the real power of an enzyme lies.

One possible use for these results is in the development of templates to annotate enzyme structures. The use of templates built from the structure of active sites to find similarities is well developed[60]. However, these templates generally involve three or more residues and may incorporate residues providing quite different functions within a single template. These large templates then make finding similarities between enzymes difficult, because the complete active site has to be conserved, rather than the smaller functional units we have described here.

Shaw *et al*[167] describe an interesting example where two of these smaller units: a classic Glu-Glu cellulase dyad (similar to that shown in Figure 5.2(c)) and a Ser-His-Glu triad; combine in a novel way. The triad is used to couple one of the catalytic glutamates with another deeply buried glutamate, which has a high pK_a, due to its burial in the core of the protein. This serves to raise the pK_a of the catalytic glutamate and ensures it is protonated and so ready to act as a proton donor in the reaction. Creating small two or three residue templates and searching for these types of combinations of catalytic units could help predict catalytic functions from structure. The difficulty of using such small templates is that the number of false positive hits rises dramatically. This could be avoided by stipulating that combinations of units have to be found or by enforcing certain other constraints on the regions of structures that are searched (restricting the search space to cleft regions for instance).

Chapter 6

Priming Catalytic Histidine

6.1 Introduction

The importance of the imidazole side chain of histidine in enzyme catalysis is well known. It is the most commonly observed catalytic residue and has by far the highest catalytic propensity of all the amino acids[76]. This is mainly because of its unique property of having a pK_a close to neutral (∼6-7), allowing the imidazole to interconvert between charged and uncharged forms and making it an ideal group for engaging in acid/base catalysis at neutral pH.

In protein structures, the imidazole ring generally exists in one of the three states shown in Figure 6.1: there are two neutral tautomers where a single proton exists on either the Nδ or Nϵ atom, and a charged form where both Nδ and Nϵ are protonated. The doubly unprotonated negative form is not generally observed in protein structures. However, it is seen in some enzyme mechanisms such as triose phosphate isomerase, (TIM) where a neutral histidine engages in acid/base catalysis with the substrate[168]. The two neutral tautomers are not identical, and in solution the Nϵ protonated tautomer is predominant because of its slightly higher pK_a[169].

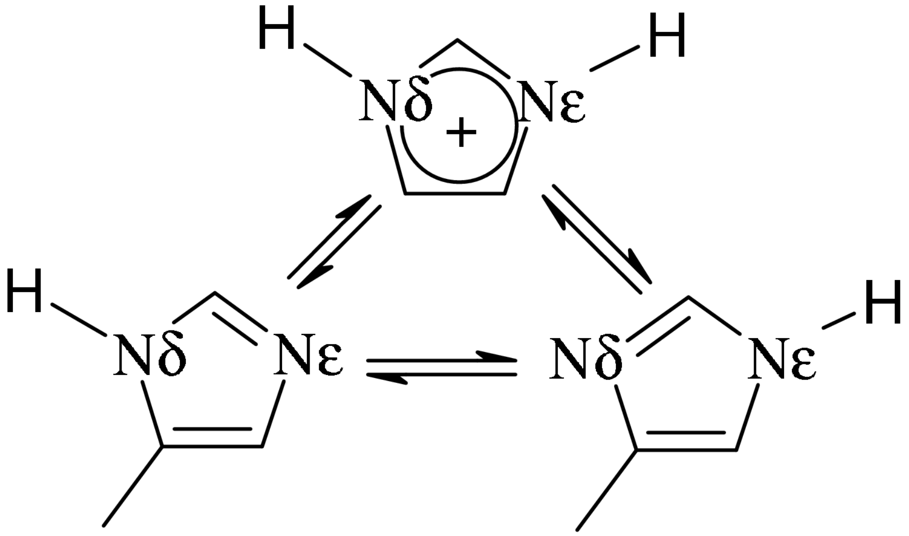


Figure 6.1: The three tautomers of histidine.

In several enzyme mechanisms a histidine residue operates as part of a dyad or triad of interacting residues[170, 171, 166]. These residues can often be thought of as priming the histidine for action, modulating its properties so that it can fulfil its catalytic function more effectively. Histidine can be primed because a residue interacting with one of the two nitrogens effects the properties of the other. The classic example of this priming effect is in the serine protease catalytic triad[45], where an aspartate residue interacts with one of the imidazole nitrogens (usually Nδ). This has the effect of raising the pK_a of the imidazole, and makes it a more effective base, so the Nϵ can abstract a proton from a serine hydroxyl which in turn acts as a nucleophile.

In this chapter we investigate the types of residues used to prime histidine, and the geometries of some of the important interactions made between histidine and these residues. There have been many investigations into side chain interactions in proteins[149, 172, 173, 174, 72, 175, 74, 153, 152, 71]. The geometry between two residues is usually defined by three parameters: the two spherical polar angles, the azimuthal angle (θ) and the equatorial angle (ϕ), (shown in Figure 6.2 for a carboxylate relative to an arginine), and the interplanar angle P (shown in Figure 6.3 for a histidine imidazole interacting with an phenylalanine ring).

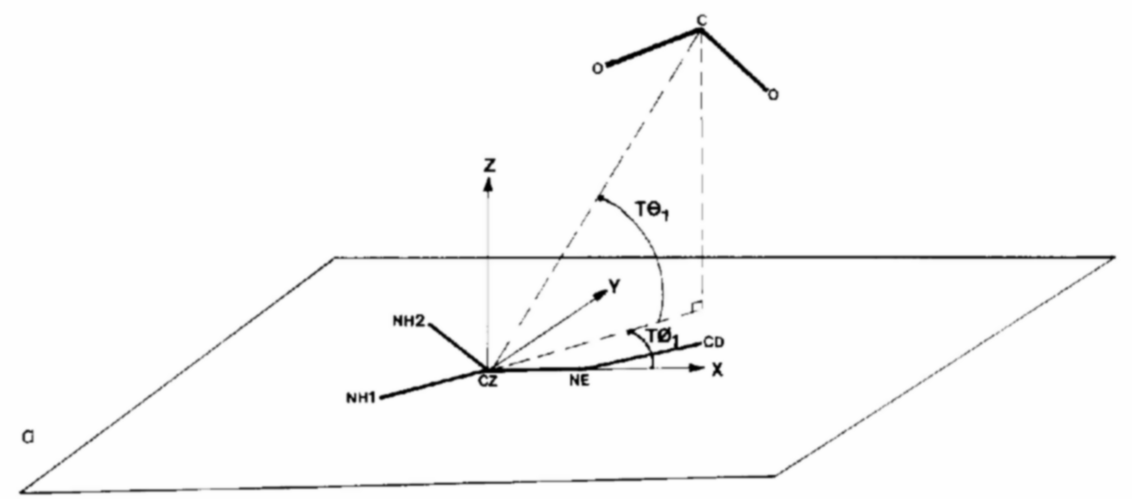


Figure 6.2: Angles defining the interaction of a carboxylate relative to an arginine. θ is the azimuthal angle, ϕ is the equatorial angle. The interplanar angle is not shown. Taken from Singh *et al* [72].

Ippolito *et al*[71] used these parameters to analyse the geometry of all the different hydrogen bonds formed between polar residues in protein structures. They found that each residue has a particular set of preferred hydrogen bond geometries. This is due to the electronic configuration of the acceptor atom, the steric accessibilities of the donor atom and the particular conformation of the residue side chain. Since many important catalytic interactions involving histidine are hydrogen bonds, this work provides a useful base line for comparing the geometry of catalytic interactions with non-catalytic interactions.

Bhattacharyya *et al*[175] looked at the interactions made by histidine side chains and again found that the interaction geometries are non-random, though they restrict themselves to examining the azimuthal angle and the interplanar angle. They also briefly compare the geometry of the histidine-aspartate interaction in the cat-

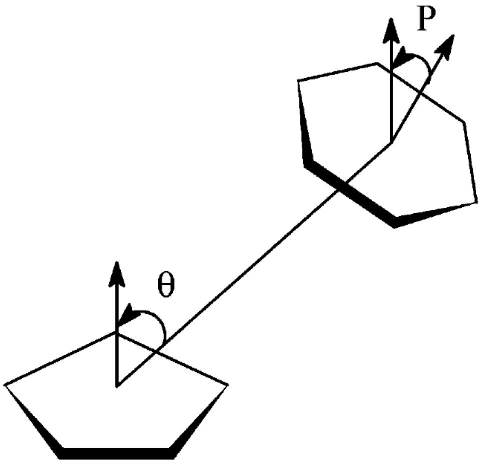


Figure 6.3: P is the interplanar angle between two residues, θ is the azimuthal angle. Taken from Bhattacharya *et al* [175]

alytic triad to the observed background and find no significant difference between the two.

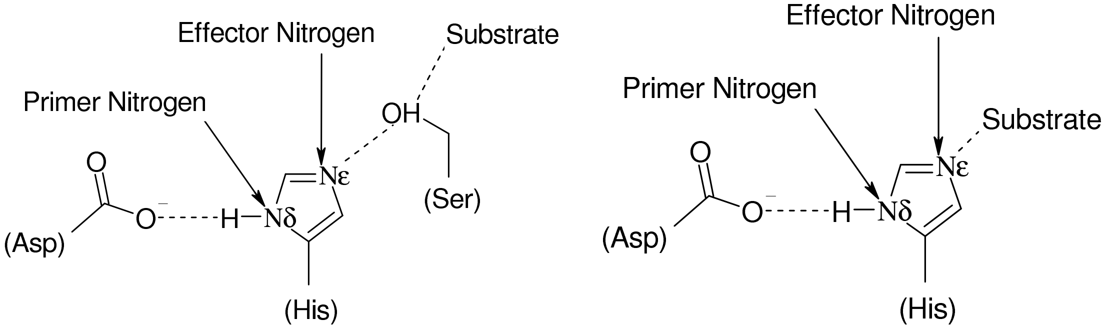
We look specifically at the interactions of histidine in catalytic systems, and look in detail at the interactions with aspartate and serine, which form the components of the catalytic triad.

6.2 Methods

6.2.1 The Dataset

We have compiled a set of 191 enzymes, as described in Chapter 5, with known catalytic residues and crystal structures taken from the CSA and the PDB respectively. The enzymes are non-redundant, with no two enzymes belonging to the same homologous superfamily as defined by CATH. All the structures are high quality (<2Å resolution, <0.3 R-factor) and the catalytic residues are found in a single conformation (all atoms have occupancy 1).

For each catalytic histidine in the dataset we identify those groups that the histidine is interacting with, and how those interactions relate to the substrate and other residues known to take part in the reaction mechanism. The imidazole nitrogen that interacts with the substrate is labelled the ‘effector’ nitrogen and the other is labelled the ‘primer’ nitrogen. If the histidine doesn’t interact directly with the substrate, but with another residue which in turn interacts with the substrate then the nitrogen that interacts with that residue is the ‘effector’. This is shown in Figures 6.4(a) and 6.4(b) for the example of a catalytic triad and a simple carbonyl-histidine dyad.



(a) Primer and effector nitrogens in the catalytic triad. Nϵ interacts with a serine which interacts with the substrate so it is the effector.

(b) Primer and effector nitrogens in a histidine-carbonyl dyad. Nϵ interacts with the substrate so it is the effector.

Figure 6.4: Identifying the primer and effector nitrogens for catalytic histidines. The nitrogen interacting with the substrate or a substrate interacting residue is the effector.

This system allows us to identify those residues or groups that are used to prime histidine (by interacting with the primer nitrogen) for action and also which of the two nitrogens (Nδ and Nϵ), if any, is the preferred nitrogen for priming.

In the dataset of 191 enzymes, 77 contain at least one catalytic histidine. To further investigate the priming of histidine, we have surveyed 49 of the 77 histidine utilising enzymes (only enzymes with well defined catalytic mechanisms were used) and assessed the role of priming in each one. Since some of these enzymes have more than one catalytic histidine there are 61 histidine residues in our final dataset for this section. All the data collected is shown in Table 6.1.

PDB	Histidine	Function	Primer	Primer Function	Effector	Primer N	Effector N
1AVW	57:A	Acid/base	Aspartate	Hbond acceptor	Serine	Nδ	Nϵ
1AYL	232:	TS stab.	Main chain O	Hbond acceptor	Substrate	Nδ	Nϵ
1B93	19:A	TS stab.	Main chain NH	Hbond donor	Substrate	Nδ	Nϵ
1B93	98:A	TS stab.	-	-	Substrate	Nδ	Nϵ
1C96	147:A	Charge stab.	Water	Hbond acceptor	Aspartate	Nδ	Nϵ
1C96	167:A	TS stab.	Glutamate	Hbond acceptor	Substrate	Nϵ	Nδ
1C96	101:A	Acid/base	Aspartate	Hbond acceptor	Substrate	Nδ	Nϵ
1CCW	16:A	Charge stab.	Aspartate	Hbond acceptor	Cofactor	Nδ	Nϵ
1CHD	190:	Acid/base	Aspartate	Hbond acceptor	Serine	Nϵ	Nδ
1CHM	232:A	Acid/base	Water	Hbond acceptor	Substrate	Nδ	Nϵ
1CPO	105:	Charge stab.	Main chain O	Hbond acceptor	Glutamate	Nδ	Nϵ
1CRU	144:A	Acid/base	Aspartate	Hbond acceptor	Substrate	Nδ	Nϵ
1D4A	161:A	Acid/base	Tyrosine	Hbond acceptor	Substrate	Nδ	Nϵ
1DUV	133:G	TS stab.	Water	Unclear	Substrate	(Nδ)	Nϵ
1EJX	1219:C	Sub. Pol.	Main chain NH	Hbond donor	Substrate	Nδ	Nϵ
1EYB	200:A	Acid/base	Water	Hbond acceptor	Substrate	Nδ	Nϵ
1EYB	248:A	Acid/base	Main chain O	Hbond acceptor	Tyrosine	Nδ	Nϵ
1FS5	143:A	Acid/base	Aspartate	Hbond acceptor	Substrate	Nδ	Nϵ
1G6S	385:A	Acid/base	Threonine	Hbond acceptor	Glutamate	Nδ	Nϵ

continued on next page

continued from previous page

PDB	Histidine	Function	Primer	Primer Function	Effector	Primer N	Effector N
1GWE	61:A	Acid/base	Serine	Hbond acceptor	Substrate	Nδ	Nϵ
1HN0	501:A	Acid/base	Glutamate	Hbond acceptor	Substrate	Nδ	Nϵ
1I0V	40:A	Acid/base	Water	Hbond acceptor	Substrate	Nδ	Nϵ
1I8D	97:A	Acid/base	Water	Hbond acceptor	Serine	Nϵ	Nδ
1IDP	85:A	Acid/base	Aspartate	Hbond acceptor	Substrate	Nδ	Nϵ
1IOM	219:A	Acid/base	Serine	Hbond acceptor	Substrate	Nϵ	Nδ
1IOM	258:A	Sub. Pol.	Main chain NH	Hbond donor	Substrate	Nδ	Nϵ
1JDW	303:	Acid/base	Aspartate	Hbond acceptor	Cysteine	Nϵ	Nδ
1MC2	1048:A	Acid/base	Aspartate	Hbond acceptor	Water	Nϵ	Nδ
1MDW	262:A	Acid/base	Asparagine	Hbond acceptor	Cysteine	Nϵ	Nδ
1MGT	142:A	Acid/base	Glutamate	Hbond acceptor	Water	Nϵ	Nδ
1MKA	70:A	Acid/base	Main chain O	Hbond acceptor	Substrate	Nδ	Nϵ
1MLA	201:	Acid/base	Main chain O	Hbond acceptor	Serine	Nδ	Nϵ
1N4W	447:A	Sub. Pol.	Asparagine	Hbond donor	Substrate	Nδ	Nϵ
1NPK	122:	Nucleophile	Glutamate	Hbond acceptor	Substrate	Nϵ	Nδ
1OBF	181:O	Acid/base	Main chain O	Hbond acceptor	Substrate	Nδ	Nϵ
1OJ8	10:A	Acid/base	Main chain O	Hbond acceptor	Substrate	Nδ	Nϵ
1OJ8	97:A	Acid/base	-	-	Substrate	-	Nϵ
1OX0	337:A	TS stab.	Main chain NH	Hbond donor	Substrate	Nδ	Nϵ
1OX0	303:A	TS stab.	Lysine	Polarises ring	Substrate	-	Nϵ
1QD1	82:A	Acid/base	Main chain O	Hbond acceptor	Substrate	Nδ	Nϵ
1QJ4	235:A	Acid/base	Aspartate	Hbond acceptor	Serine	Nϵ	Nδ

continued on next page

continued from previous page

PDB	Histidine	Function	Primer	Primer Function	Effector	Primer N	Effector N
1Q9I	365:A	TS stab.	Main chain NH	Hbond donor	Substrate	Nδ	Nϵ
1Q9I	504:A	Acid/base	Threonine	Hbond acceptor	Substrate	Nδ	Nϵ
1QL0	89:A	Acid/base	Asparagine	Hbond acceptor	Water	Nϵ	Nδ
1QQQ	207:A	Acid/base	Aspartate	Hbond acceptor	Water	Nδ	Nϵ
1QWO	59:A	Nucleophile	Main chain O	Hbond acceptor	Substrate	Nδ	Nϵ
1R0R	64:E	Acid/base	Aspartate	Hbond acceptor	Serine	Nδ	Nϵ
1S95	304:A	Acid/base	Aspartate	Hbond acceptor	Substrate	Nδ	Nϵ
1T2D	181:A	Acid/base	Aspartate	Hbond acceptor	Substrate	Nδ	Nϵ
1T48	274:A	Acid/base	Glutamine	Hbond acceptor	Cysteine	Nδ	Nϵ
1UCD	88:A	Acid/base	Water	Unclear	Substrate	Nδ	Nϵ
1UCD	34:A	Acid/base	Glutamine	Hbond donor	Substrate	Nδ	Nϵ
1USM	40:A	Acid/base	Aspartate	Hbond acceptor	Substrate	Nϵ	Nδ
1USM	41:A	Acid/base	Glutamate	Hbond acceptor	Substrate	Nϵ	Nδ
1USM	58:A	Acid/base	Glutamate	Hbond acceptor	Substrate	Nϵ	Nδ
1VDK	188:A	Acid/base	Glutamate	Hbond acceptor	Water	Nδ	Nϵ
1VNS	404:A	Acid/base	Main chain O	Hbond acceptor	Substrate	Nϵ	Nδ
2PTD	32:	Acid/base	Aspartate	Hbond acceptor	Substrate	Nδ	Nϵ
2PTD	82:	Acid/base	Aspartate	Hbond acceptor	Substrate	Nδ	Nϵ
2PTH	20:	Acid/base	Aspartate	Hbond acceptor	Water	Nδ	Nϵ
3CLA	195:	Acid/base	Main chain O	Hbond acceptor	Substrate	Nδ	Nϵ

Table 6.1: The roles and priming groups for each histidine in the dataset. ‘TS Stab.’ refers to transition state stabilisation, ’Charge Stab.’ to charge stabilisation and ’Sub. Pol.’ to substrate polarisation functions.

6.3 Analysis and Results

6.3.1 Priming of Histidine

Histidine residues fulfil a number of different roles in catalysis, the relative proportions found each one is shown in Figure 6.5. It is clear from this data that acid/base chemistry is by far the most common role for histidine in catalysis, comprising 75% of the cases in this dataset, though histidine can fulfil a number of other roles. In transition state and charge stabilisation either the positively charged imidazole ring or a protonated nitrogen is used to stabilise a negative charge on the transition state, or some other residue such as an aspartate. Substrate polarisation involves using a protonated nitrogen to donate a hydrogen bond in such a way that other bonds in the substrate are polarised to facilitate their cleavage. Histidine can also use the imidazole nitrogens as nucleophiles.

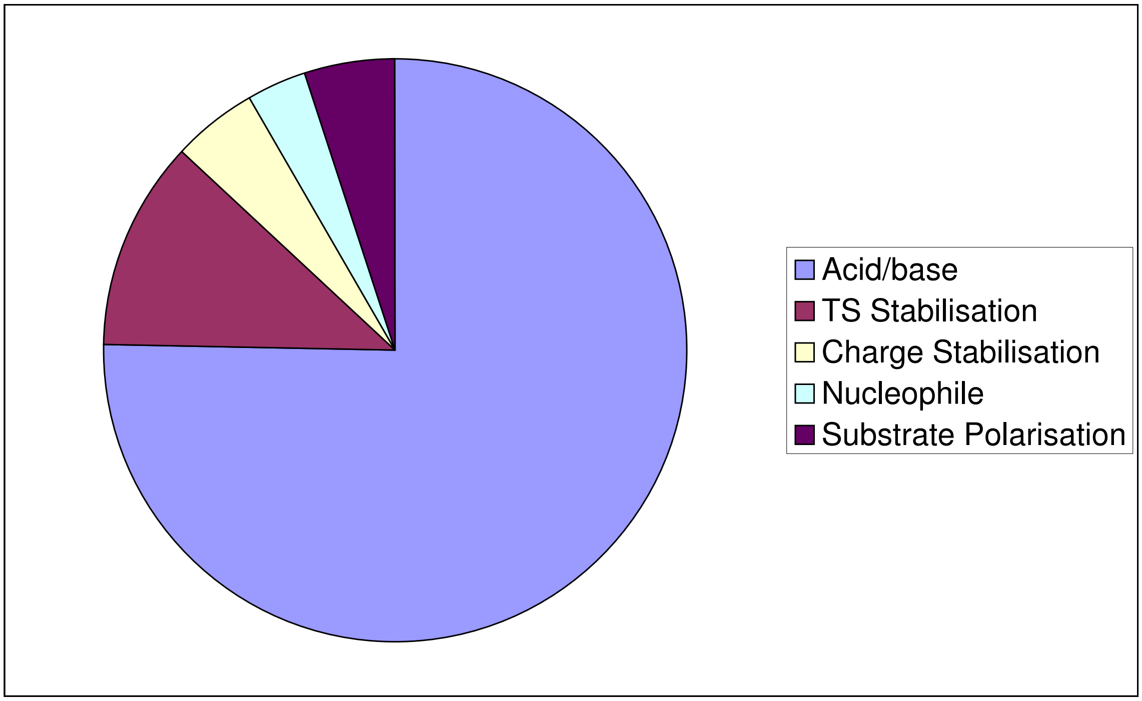


Figure 6.5: Counts of the different roles for histidine.

Figure 6.6 shows the relative frequency of primer groups observed in this dataset compared with the interaction partners of all, partially buried and fully buried non-catalytic residues. Partially buried residues are defined as those with a relative surface accessibility of less than 20% and fully buried residues are defined as those with a surface accessibility of 0. We see that water is the most common interaction partner for non-catalytic histidines, surprisingly this is true even for completely buried histidines. However, water is seen relatively rarely as a priming group, and in those cases where it is used, it is almost always used as a bridging molecule connecting the histidine to another residue such as aspartate.

Figure 6.7 shows the same data as Figure 6.6 but excludes water. From this we see that aspartate is the most common priming residue (priming over 30% of histidines) and is seen as a priming residue much more often than it is seen in non-catalytic interactions (glutamate is used more rarely, but shows the same pattern of being seen more often in catalytic systems). In contrast, the hydroxyl residues: serine, threonine and tyrosine are all observed less often in catalytic systems than non-catalytic. Carbonyl and amide groups (including both main chain groups and asparagine and glutamine side chains are seen at similar levels in both non-catalytic and catalytic systems.

The partners histidine interacts with at the other nitrogen, once primed is shown in Figure 6.8. 41 of the 61 histidines act directly on a substrate, whilst 13 act on another residue which in turn interacts with the substrate, 6 interact with water and one with the cofactor. In the cases of serine, cysteine, tyrosine and water the histidine is almost always used as a base to activate the -OH,-SH or HOH as a nucleophile (-O$^-$,-S$^-$ and OH$^-$ respectively). For interactions with glutamate and aspartate the role of histidine is either as a base (again activating the carboxylate to act as a nucleophile) or simply to stabilise the negatively charged side chain.

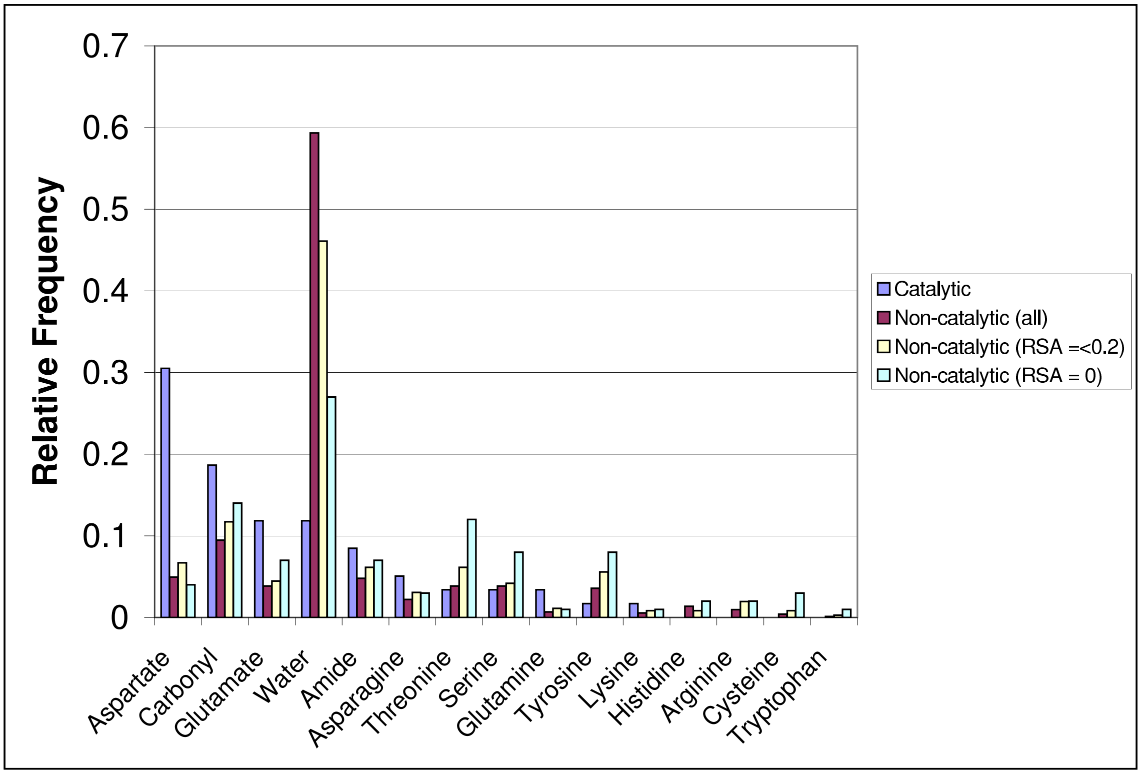


Figure 6.6: Relative frequency of interacting partners for non-catalytic histidine, buried non-catalytic histidine and primers of catalytic histidine.

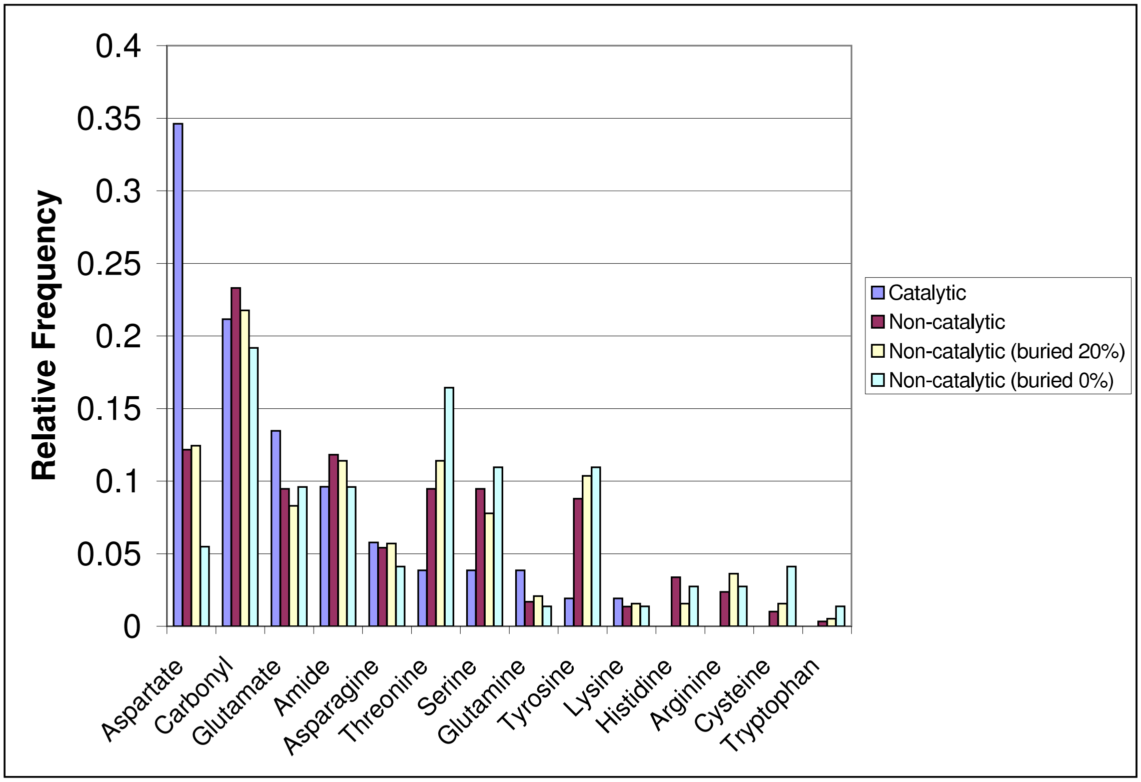


Figure 6.7: Relative frequency of interacting partners for non-catalytic histidine, buried non-catalytic histidine and primers of catalytic histidine excluding water.

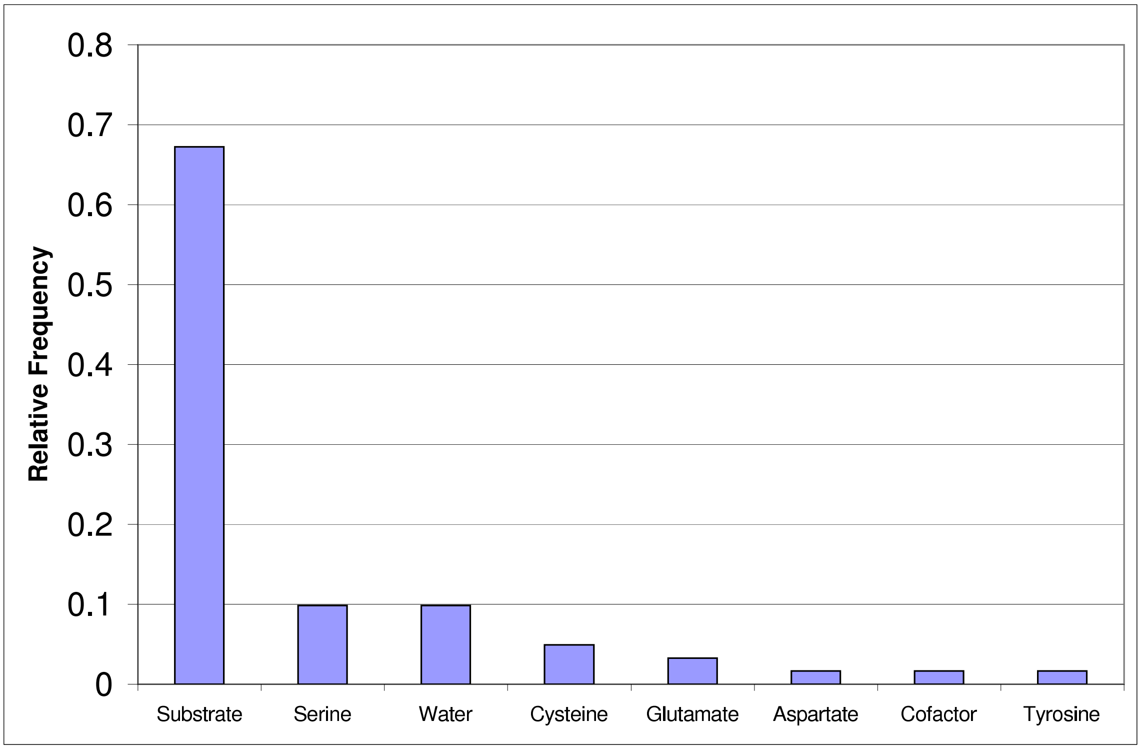


Figure 6.8: Counts of the different effectors for histidine.

Almost all the priming interactions involve the formation of a hydrogen bond from the priming group to one of the imidazole nitrogen atoms. Of the 61 histidines, 56 are involved in a hydrogen bond to the priming group. Imidazole nitrogens can be protonated or unprotonated so can act as both a hydrogen bond donor and an acceptor.

Of the priming residues only the hydroxyl residues, water and the amide residues are also capable of acting as donors and acceptors. Carboxylates have a low pK_a so are almost always unprotonated and act as hydrogen bond acceptors only. Similarly, carbonyls only act as hydrogen bond acceptors and amides only act as hydrogen bond donors.

Of the 56 priming hydrogen bonds only 6 use the priming group as a donor, whilst in 50 the priming residue is the acceptor. We also note that water, hydroxyl and amide almost always act as hydrogen bond acceptors rather than donors.

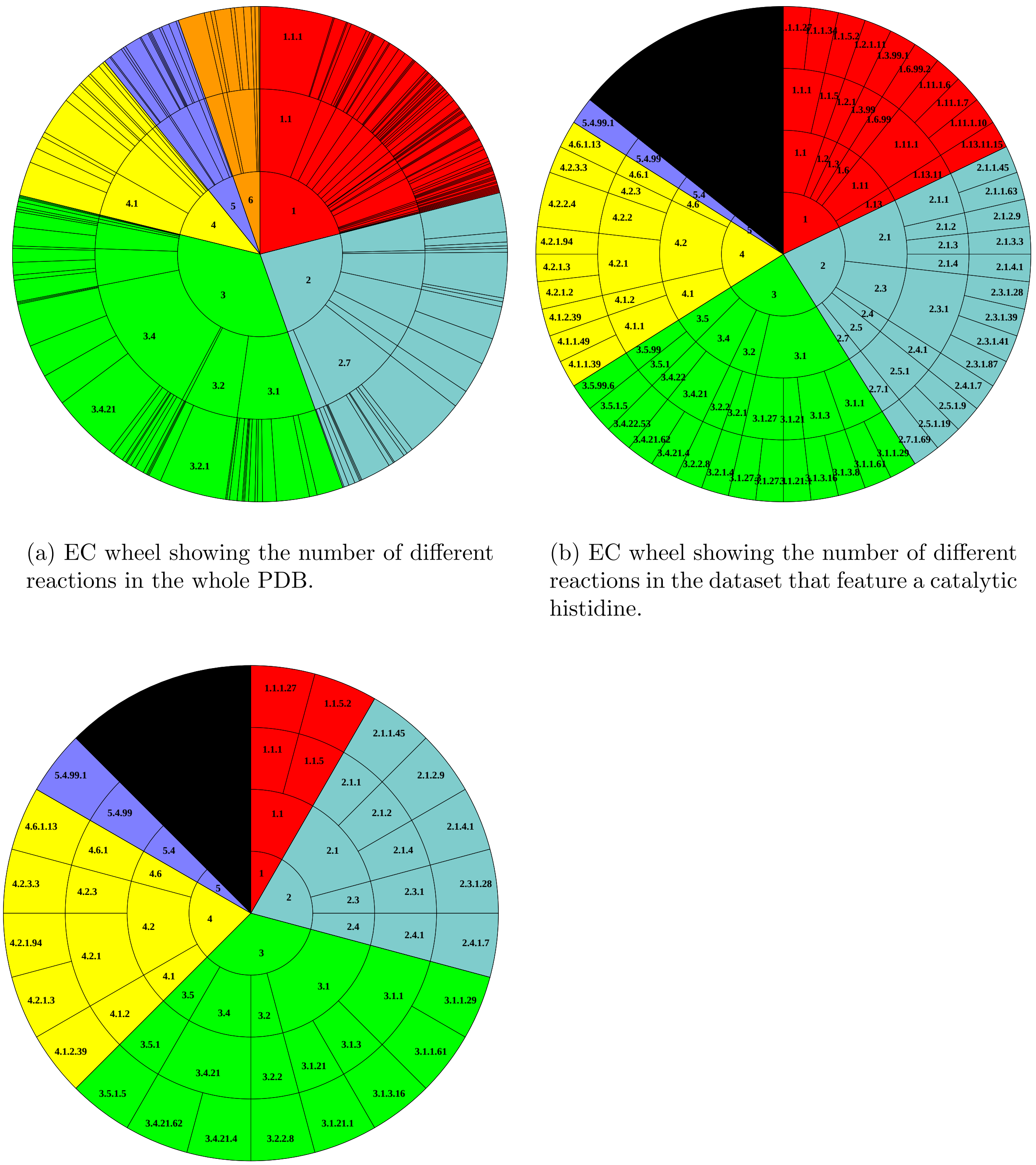
We also see a preference for priming at the Nδ nitrogen. Of the 56 primed histidines, 14 are primed at the Nϵ and 42 are primed at the Nδ. This tendency is particularly strong for histidine-aspartate bonds. Searching through all the catalytic histidine-aspartate hydrogen bonds in the original dataset of 191 enzymes, we find that 74% are formed between the carboxylate and Nδ and only 26% between the carboxylate and Nϵ. In the same 191 enzymes, the non-catalytic histidine-aspartate interactions are formed at Nδ 40% of the time and 60% of the time at Nϵ. This reflects the tendency for Nϵ to be protonated more often than Nδ.

6.3.2 Histidine Interactions

Histidine-Aspartate Interactions

The previous section seems to suggest an important role for aspartate-histidine interactions in catalysis. Figure 6.9(a) shows the range of different reactions (EC classes) that are represented by a structure in the PDB (for clarity, only the first three EC levels are shown). Figures 6.9(b) and 6.9(c) shows those EC classes in this dataset that feature a histidine and a histidine-aspartate dyad in the mechanism respectively. We can see that histidine and histidine-aspartate interactions take part in almost every class of reaction covered by the PDB and the distribution across the top level EC numbers is roughly equal to that seen for the whole PDB.

We extract 31 interactions between histidine and aspartate from the full dataset of 191 enzymes, where both residues are annotated in the literature as catalytic. We also extract 370 interactions between histidine and aspartate, from these same enzymes, where both residues are annotated as non-catalytic. Residues are defined as interacting if any two of their side chain atoms are closer than 4Å. This definition includes interactions which are not necessarily hydrogen bonds.



(a) EC wheel showing the number of different reactions in the whole PDB.

(b) EC wheel showing the number of different reactions in the dataset that feature a catalytic histidine.

(c) EC wheel showing the number of different reactions that feature a catalytic histidine-aspartate dyad.

Figure 6.9: EC wheels of the whole PDB, histidine utilising enzymes and histidine-aspartate utilising enzymes. Different top level EC classes (1.*,2.*,etc) are coloured differently. Enzymes in the dataset without a EC classification are coloured black.

For each interaction we measure the equatorial angle of the aspartate relative to the imidazole sidechain. The equatorial angle is defined such that the vector formed by the Cβ-Cγ bond is 0^o and the imidazole ring is aligned as shown in Figure 6.10. The azimuthal and interplanar angles were also measured, but no significant differences between the catalytic and non-catalytic datasets were observed.

To visualise the difference between the catalytic and non-catalytic interactions more clearly, we use a density map. First we translate all the interactions into a common frame of reference so that the histidines are superposed in each case. Then, we build a density map by placing a Gaussian shaped 'blob' of density at the centroid of each carboxylate, so that regions with many carboxylates are represented by a region with high density in the map.

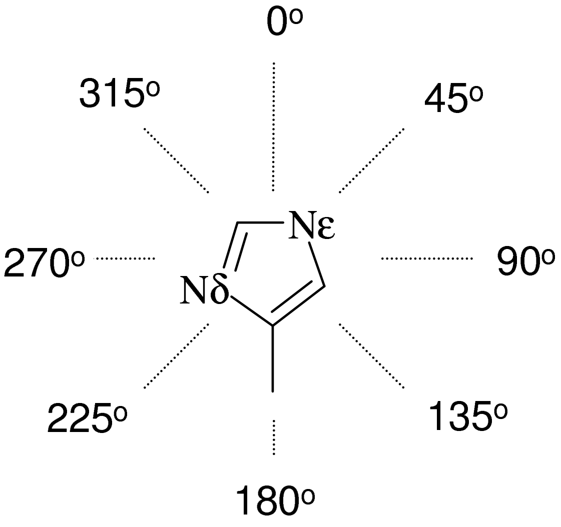


Figure 6.10: The definition of the equatorial angles around the imidazole ring of histidine.

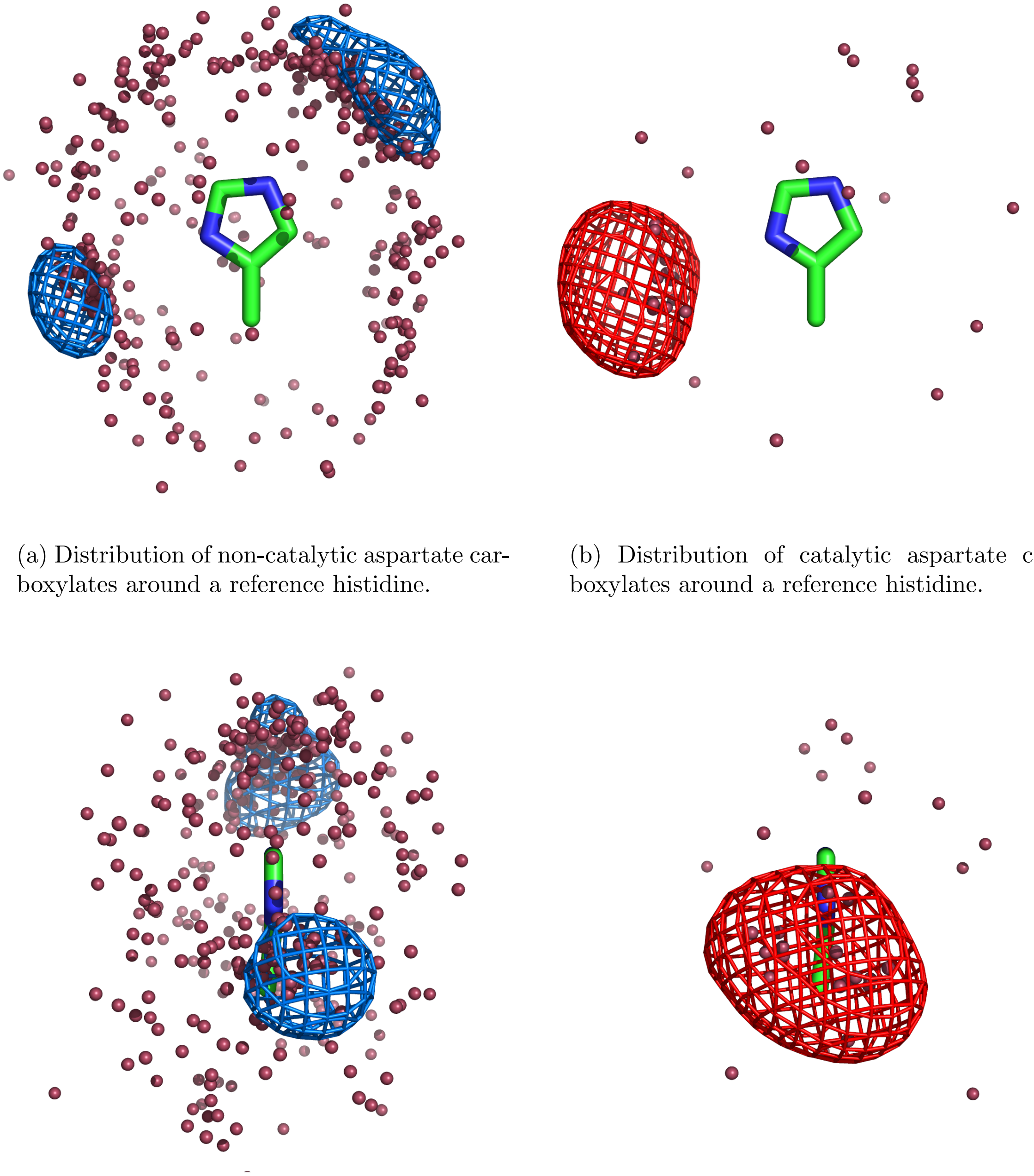
Figures 6.11(a) and 6.11(c) show the distribution of non-catalytic aspartate residues around a reference histidine side chain. Each aspartate sidechain is represented by a sphere placed at the Cγ atom. The mesh represents a contour of the density map to show regions with a high density of aspartates. Clusters of aspartates can be seen around the Nδ and Nϵ atoms as expected. Figures 6.11(b) and 6.11(d) show the same data, but restricted to the catalytic interactions. The density maps for the catalytic and non-catalytic data are normalised so that the average density is equal in each case, and the contour levels are the same in each diagram to allow them to be compared. We can immediately see that, in contrast to the non-catalytic set, only the Nδ cluster is visible in the catalytic set.

By dividing both the catalytic and non-catalytic interactions into bins according to their equatorial angle, we confirm that there is an excess of catalytic interactions from aspartate to the Nδ of histidine using a χ^2 test. Figure 6.12 shows this data by plotting the number of observed and expected catalytic interactions against the equatorial angle around the imidazole. There is a large excess in catalytic interactions where the equatorial angle is in the range 240-270^o, representing interactions between the carboxylate and the Nδ atom.

We also note that the non-catalytic cluster around the Nϵ atom is considerably more spread out than the cluster around the Nδ atom. This confirms an observation made by Ippolito *et al*[71] for all histidine hydrogen bonds (not just histidine-aspartate). However, the authors of that study suggested this was due to the preponderance of hydrogen bond donors at Nδ (which is more likely to be unprotonated) which they suggest are more constrained than hydrogen bond acceptors. In histidine-aspartate interactions, the histidine is always the acceptor so this cannot explain the clustering that we observe.

Histidine - Serine Interactions

Although less important for histidine priming, the histidine-serine interaction is interesting because it forms the other half of the catalytic triad. Although the



(a) Distribution of non-catalytic aspartate carboxylates around a reference histidine.

(b) Distribution of catalytic aspartate carboxylates around a reference histidine.

(c) Distribution of non-catalytic aspartate carboxylates around a reference histidine (rotated by 90^o).

(d) Distribution of catalytic aspartate carboxylates around a reference histidine (rotated by 90^o).

Figure 6.11: Plots of aspartate carboxylate groups around a reference histidine residue.

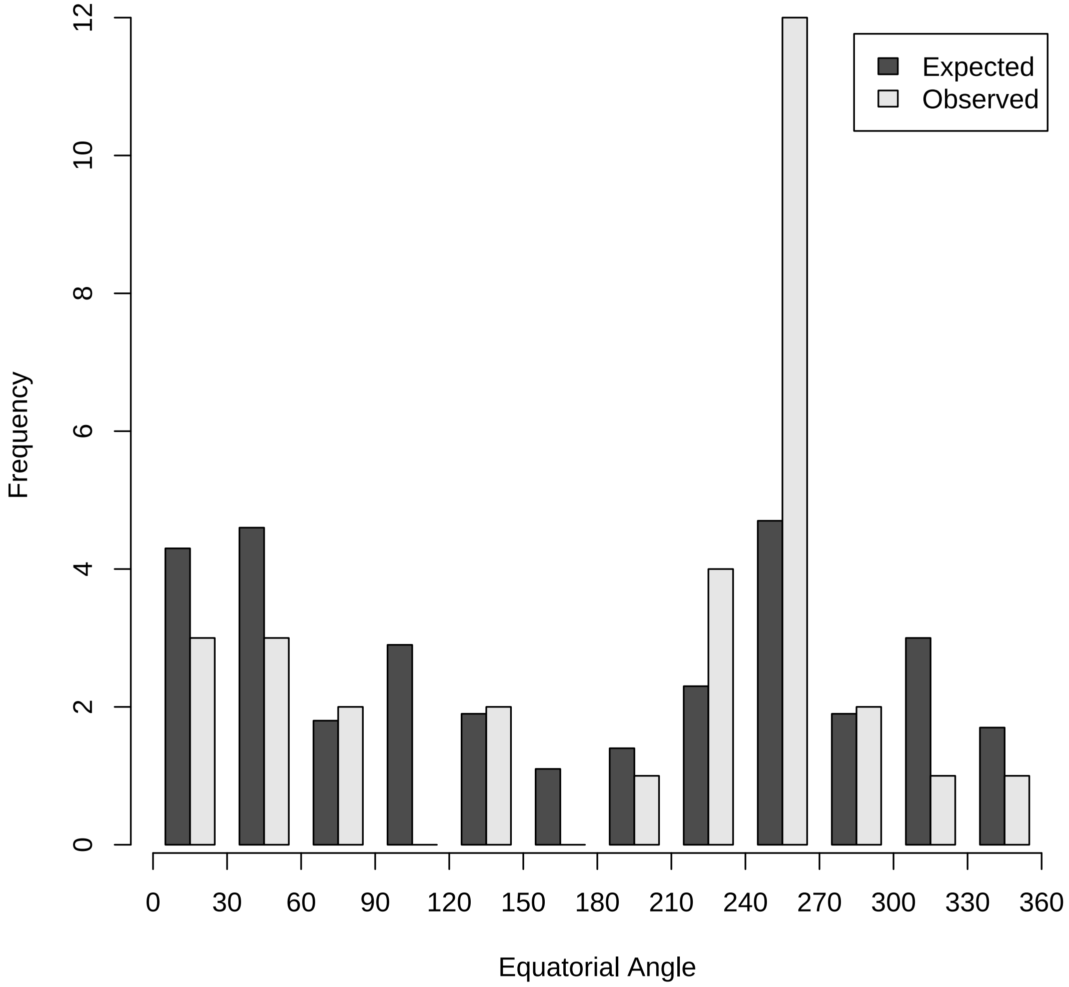


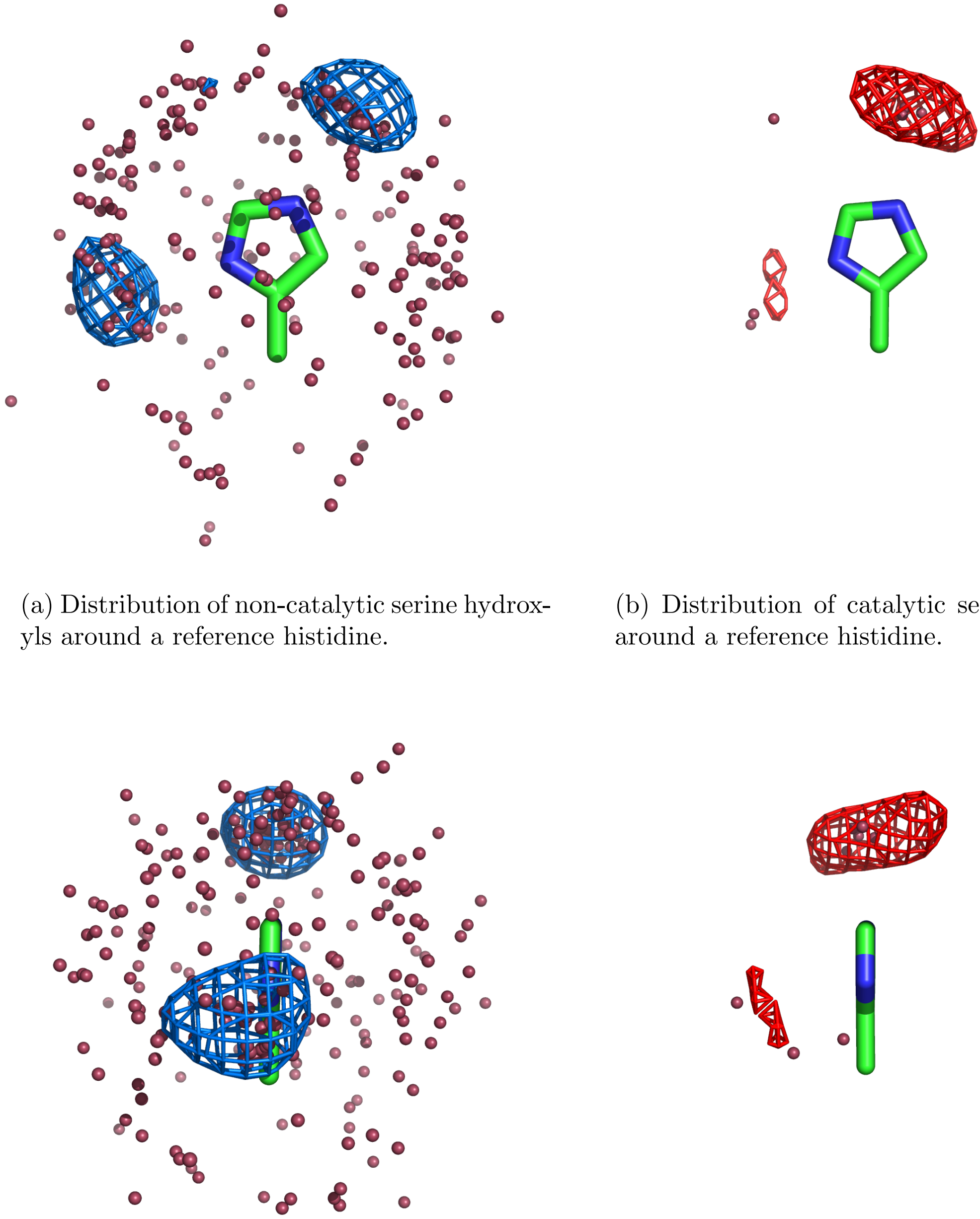
Figure 6.12: Expected and observed counts of equatorial angles (relative to histidine) for histidine interacting with aspartate. There is an excess of observed interactions at 210-270^o. This shows the preference for aspartate to interact with the Nδ atom of histidine in catalytic systems

number of histidine-serine catalytic interactions (8) in our dataset is far smaller than for the histidine-aspartate interaction (31), and so firm conclusions are harder to make. Four of the eight histidine-serine interactions are part of a Ser-His-Asp catalytic triad. Figure 6.13 shows the density of catalytic and non-catalytic serine side chains around a reference histidine. The spheres represent the Oγ atoms of serine in these figures.

Comparing this to the histidine-aspartate data shown in Figure 6.11, at first glance we see an apparent tendency for Nϵ to be preferred by the catalytic serines (the reverse of the aspartate preference). Closer examination of Figure 6.13(d) and Figure 6.14 (which plots the equatorial angle of the serines), shows that the numbers are roughly equal, with four serines at the Nϵ, three at Nδ and one closer to Cϵ. It is possible that this last example is a case where the histidine ring is incorrectly orientated in the crystal structure, and so should be interacting at Nϵ. The serines at Nϵ are also much more tightly clustered than the serines at Nδ, which explains why the density of serines is less defined at the Nδ. Of the four catalytic triads, three have the serine at Nϵ and one has the serine at Nδ.

We can also look at these interactions from the perspective of the serine, using serine as the reference residue and measuring where the histidines are found relative to the hydroxyl side chain. Figure 6.15 shows the a density plot based on this data, the imidazole ring of each histidine is also shown. The non-catalytic density is concentrated in two areas below the Cβ-Cγ bond. Again, this is a pattern seen previously[71] in hydrogen bond interactions with serine. In contrast, the catalytic histidines all lie in the area above the Cβ-Cγ bond.

Freely rotating bonds like the Cβ-Oγ bond are constantly spinning. However, this places the Oγ hydrogen atom into different conformations that are not all equally favourable. The most stable conformation is where the Oγ hydrogen is staggered



(a) Distribution of non-catalytic serine hydroxyls around a reference histidine.

(b) Distribution of catalytic serine hydroxyls around a reference histidine.

(c) Distribution of non-catalytic serine hydroxyls around a reference histidine (rotated by 90^o).

(d) Distribution of catalytic serine hydroxyls around a reference histidine (rotated by 90^o).

Figure 6.13: Plots of catalytic and non-catalytic serine hydroxyls groups around a reference histidine residue.

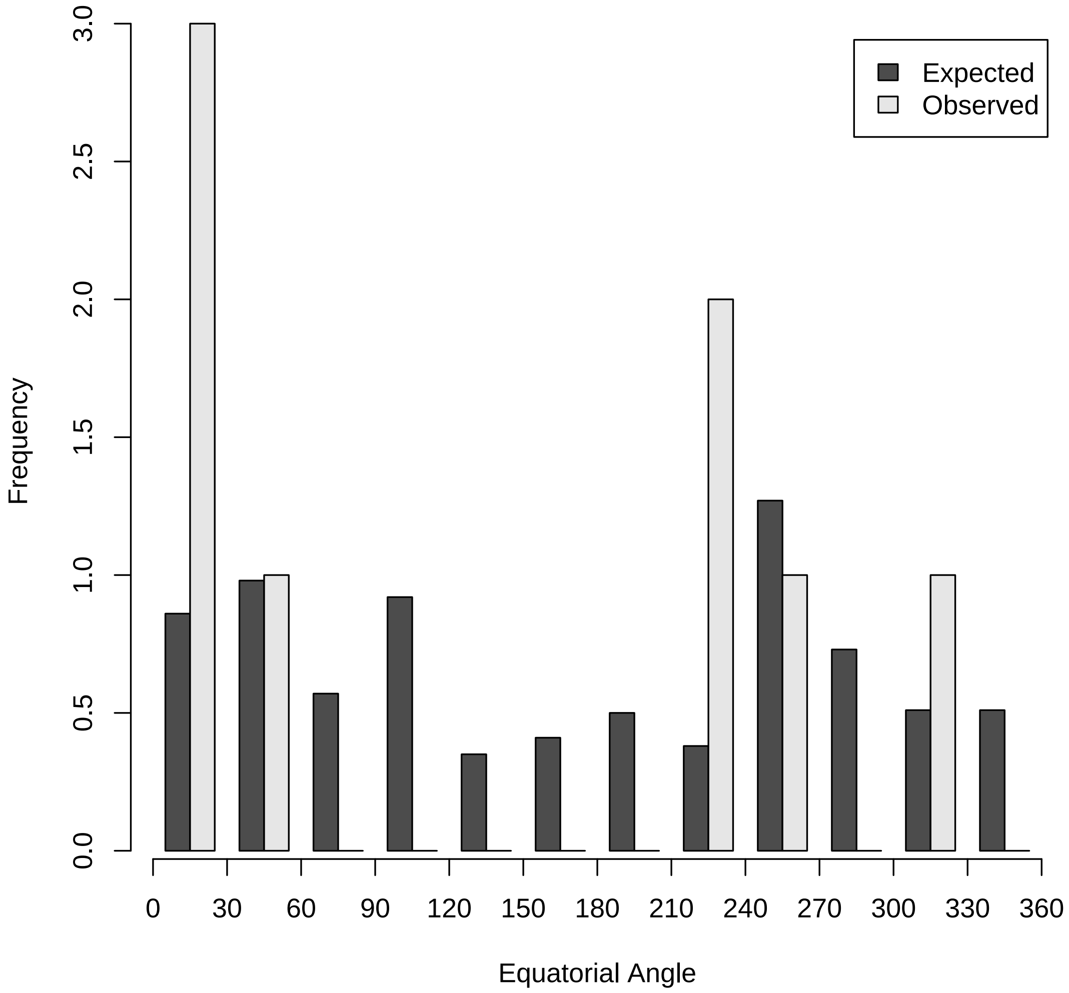
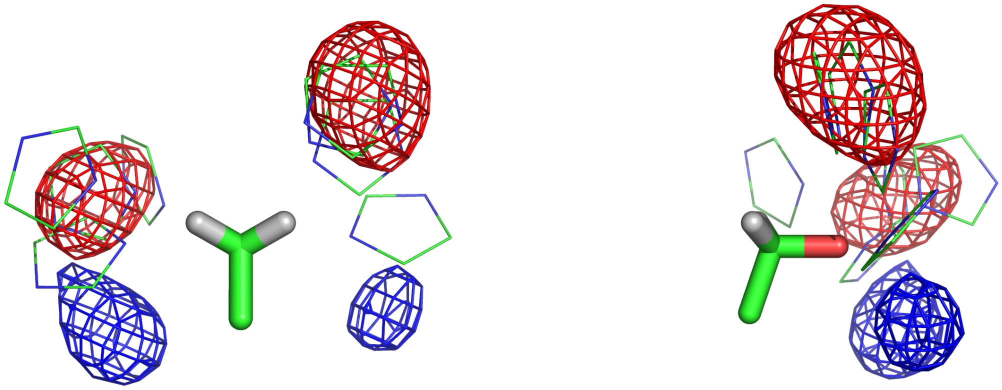


Figure 6.14: Expected and observed counts of equatorial angles (relative to histidine) for serine interacting with histidine.



(a) Distribution of catalytic and non-catalytic histidines around a reference serine residues.

(b) Distribution of catalytic and non-catalytic histidines around a reference serine residues (rotated 90^o.

Figure 6.15: Plots of aspartate carboxylate groups around a reference histidine residue.

with the Cβ hydrogens as this prevents overlap of their orbitals. The most unstable conformation is the eclipsed conformation where the hydrogens directly line up. This is shown as a Newman projection in Figure 6.16.

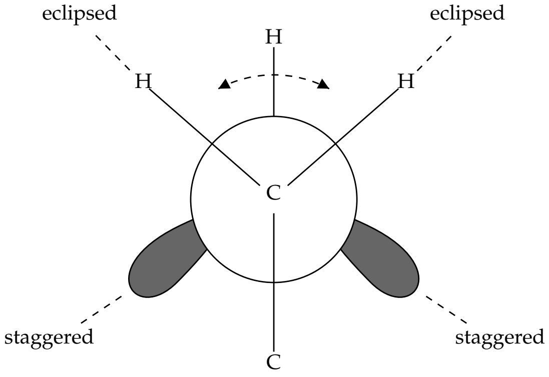


Figure 6.16: Newman projections of serine. We find that non-catalytic hydrogen bonds tend to form in the gauche regions whilst catalytic bonds are staggered.

By placing the histidine in the regions shown in Figure 6.15, the serine hydrogen atom involved in the hydrogen bond is forced into the most unfavourable eclipsed conformation. This could encourage the loss of the proton to the histidine, a key step in the mechanism of enzymes utilising a catalytic triad. A problem with this explanation is that one would expect the highest energy state to be where the hydrogen is opposite the Cβ-Cα bond (the orbital overlap would be greatest in that conformation) and we see no density here. Possibly steric constraints prevent this being a useful conformation for catalytic activity.

6.4 Discussion

6.4.1 Histidine Priming

The survey of histidine priming residues reveals a number of interesting observations. To put them in context we can consider the three possible roles for priming residues interacting with histidine:

- Orientation. The histidine sidechain is freely rotatable around the Cα-Cβ and Cβ-Cγ bonds, allowing the imidazole ring to rotate and move relative to the Cα. Since correct positioning of catalytic groups is essential in most enzymes the catalytic histidine must usually be held in place by interactions with another group. This pre-organisation also means that the entropic costs of restricting the motion of the histidine is paid for during the folding of the protein rather than during catalysis.

  Orientation can in theory be done by almost any group capable of interacting with an imidazole ring. These interactions can be electrostatic (assuming the imidazole is charged), hydrogen bonding to either nitrogen (or carbon) or via hydrophobic stacking interactions. Almost every histidine in the dataset has at least one interaction that maintains orientation.

- Tautomerisation. The histidine sidechain can exist in two neutral tautomers with the proton on either of the two nitrogens. The higher pK$_a$ of Nϵ means that the preferred tautomer is protonated at the Nϵ, but interconversion will occur in free solution. It is essential in almost all reactions that the neutral histidine exists in one particular tautomer as this has obvious consequences for the ability of specific nitrogens to engage in acid/base chemistry and to provide hydrogen bonding.

  Priming a histidine with a hydrogen bond donor ensures that the nitrogen interacting with the primer is unprotonated (otherwise there is no lone pair to engage in bonding). This means that the other nitrogen will usually be protonated as the negative imidazolate ion is very unusual in protein structures, though we do observe this form in the mechanism of triose phosphate isomerase (which unfortunately is not included in this dataset). As a general rule however, histidine primed using a hydrogen bond donor (or positive charge) will not engage in acid/base catalysis by losing its single proton. It is capable of transition state stabilisation catalysis through donation of a hydrogen bond however (as is observed in a number of cases).

Priming through a negative charge or hydrogen bond acceptor ensures the nitrogen being primed is protonated at all times. This allows the other nitrogen to be either protonated (forming the positive imidazolate ion) or unprotonated. A histidine primed in this way can engage in acid/base chemistry and charge stabilisation via the positive charge or hydrogen bond.

- pK_a Modification. The histidine sidechain usually has a pK_a of 6-7. However the placement of charged/uncharged residues nearby can greatly alter this by stabilising/destabilising the positive form. A nearby negative charge will raise the pK_a of the histidine by stabilising the positive imidazole. In contrast a positive charge or hydrophobic group will destabilise the positive form and reduce the pK_a. Charged groups are used in mechanisms where pK_a modulation is required for catalysis.

From the survey of priming groups we see that hydrogen bond acceptors are by far the most common group. This is as expected from the rationale given above, as these groups allow the histidine to take part in acid/base chemistry, which is the most common function for histidine. Priming by hydrogen bond donors is seen, but is usually restricted to those cases where the histidine is merely involved in stabilising a negative charge on another group or substrate.

Of the hydrogen bond acceptors we see that the carboxylates, particularly of aspartate, are seen most often in priming. In contrast, the hydroxyl groups of serine, threonine and tyrosine are seen less often. This confirms the importance of pK_a modification in priming histidine since the negatively charged carboxylate will significantly alter the histidine's pK_a, unlike a hydroxyl group.

6.4.2 Interaction Geometries

We find that the distribution of the geometries of the catalytic aspartate-histidine interactions are significantly different to the non-catalytic interactions. Both distributions show the expected clustering of carboxylates around the imidazole nitrogens, but the catalytic distribution showed a distinct preference for the Nδ atom rather than the Nϵ atom. In contrast, the non-catalytic set shows a slight preference for Nϵ over Nδ. Given that aspartate is used so often as a priming group, compared to how often the histidine is used to prime an aspartate, we can assume that this distinction is related to how the aspartate is priming the histidine.

One possible explanation for this observation is that it is related to the pK_a differences between the two imidazole nitrogens. Making a histidine interact with an aspartate will tend to raise its pK_a and hence make it a stronger base. The pK_a of Nϵ is slightly higher than that of Nδ (6.73 compared to 6.12)[176] and so Nϵ is the more basic of the two imidazole nitrogens. So, given that after interacting with an aspartate, the histidine is likely to be used as a base, it seems logical that evolution will choose the more basic nitrogen to act as that base. Therefore, Nϵ is preferred as the atom that acts as the base, whilst Nδ is the atom at which the aspartate interacts.

It has been noted previously that many catalytic residues, including, but not restricted to, histidine, have perturbed pK_a values[177], and these perturbations have also been used to predict the identity of catalytic residues in enzyme structures of unknown function by using *in silico* calculations[65, 178], so the idea of priming being related to pK_a modification seems reasonable.

The same preference for priming at the Nδ has also been observed in metal ligands. In these cases the priming residue interacts at one nitrogen, whilst the other nitrogen ligates a metal atom. Chakrabarti first made this observation[179] and Alberts *et al* confirmed it using a larger dataset, finding that ~70% of zinc ligating atoms did so at the Nϵ atom[180]. This preference for priming at Nδ seems very similar to our observation. Alberts *et al* do not comment further on this effect, but Chakrabarti suggests that it is related to steric constraints that prevent metal ligation at Nδ, because the metal is then too close to the main chain atoms. The same reasoning could be applied in catalytic systems to bulky substrates, which may find it easier to interact at the Nϵ, rather than Nδ. However, this does not seem to explain the preference for Nδ priming in systems such as the catalytic triad where the histidine interacts with a serine side chain rather than a substrate directly. In these cases there does not seem to be a steric reason why evolution would choose aspartate for Nδ and serine for Nϵ.

The other residue interaction that we have investigated is the serine-histidine interaction. The small number of histidine-serine interactions in our dataset makes the conclusions from this data very tentative. However, there are a number of interesting observations we make from this data. From the point of view of the histidine, the serine does not seem to prefer any particular nitrogen, though we do note that three out of the four histidine-serine interactions involved in a catalytic triad prefer to interact at Nϵ. As mentioned above, this nitrogen is more basic, and so may be better able to extract protons from the nucleophilic serine. Interestingly, we note that the only Ser-His-Asp triad in this dataset that has a Nϵ-aspartate priming interaction (the reverse of the usually observed interaction) is the methylesterase CheB[181]. This enzyme has a similar mechanism to the serine proteases, but cleaves an ester bond rather than an amide bond. Ester bonds are much more reactive than amide bonds and do not require as strong a nucleophile as amide bonds to cleave. This may explain why the less basic histidine nitrogen is used in this case to activate the serine.

We also note that the histidine seems to interact with the serine in a particular way from the point of view of the serine. The imidazole group is aligned such that when a hydrogen bond is formed between the hydroxyl and the imidazole the proton is eclipsed with the Cβ hydrogens of the serine side chain. The effect of this will be to weaken the bond from the serine to the proton and encourage its loss to the histidine. This is of course the first step in the mechanism of serine proteases and we speculate that this is a deliberate evolutionary design for encouraging the formation of the serine nucleophile.

Chapter 7

Improved Non-covalent Bonding Within Enzyme Structures Upon Binding of Ligands

7.1 Introduction

In this chapter, we investigate a recent theory to explain the high rates of catalysis achieved by certain enzymes. The theory uses the effect of entropy-enthalpy compensation upon substrate (and transition state) binding to provide an explanation as to how enzymes reduce the energy of the bound transition state.

In enzyme catalysed reactions, like any reaction, the free energy change that drives the direction of the reaction is described by the Gibbs equation (7.1).

$$\Delta G = \Delta H - T\Delta S \tag{7.1}$$

Where ΔG is the change in free energy, ΔH is the enthalpy change, T is the temperature and ΔS is the entropy change of the reaction. A spontaneous reaction has a negative ΔG and can either be driven by a negative ΔH (an exothermic reaction driven by the formation of bonds) or a positive ΔS (an endothermic reaction driven by an increase in the entropy of the system) or both.

The enthalpy and entropy changes in a reaction are related, since improved bonding between two groups (enthalpy) opposes their relative motion (entropy). In enzymes, the binding of ligand molecules is an example of this effect. The formation of bonds between the enzyme and the substrate favours ligand binding (by making ΔH increasingly negative), but the reduction of the dynamics of both the ligand and the enzyme disfavours binding by decreasing the entropy of the system. It has been recently proposed by Williams *et al*[182] that one of the effects of this reduction in the dynamics of the ligand bound complex is to reduce the length of non-covalent bonds throughout the enzyme. This serves to counteract the disfavourable entropy change by providing benefits in enthalpy (improved bonding) in the ligand bound state, over and above that provided by the bonds formed between the enzyme and the ligand.

In enzymes, one of the goals of the catalyst is to reduce the free energy of the transition state and thereby enhance the rate of reaction. Transition state theory holds that the main mechanism for this reduction of energy is preferential bonding of the enzyme to the transition state rather than the substrate. In addition to this, Williams *et al* suggest that improved non-covalent bonding *within* the enzyme can play an important role in reducing the free energy of the transition state and hence may be a previously unrecognised source of catalytic power.

The hypothesis predicts that protein structures will show both a reduction in dynamics in the ligand bound state and improved packing of the enzyme core. A reduction in dynamics has been observed from hydrogen/deuterium (H/D) exchange studies on streptavidin[183] showing that backbone amide groups exchange protons more slowly with the solvent in ligand bound states than the apo state. However, bond length shortening has not been documented to date.

The magnitude of non-covalent bond changes may be very small and yet still significant, because of the large number of non-covalent bonds within an enzyme and the additive effect of strengthening many of them. Bond length changes of just 1% have been proposed as being significant. This would correspond to a change of 0.03Å for a hydrogen bond of 3Å. We have attempted to observe this effect in crystal structures of enzymes where the structure of the apo and ligand bound forms are known at a high resolution.

7.2 Methods

We use a set of enzyme structures extracted from the Protein Data Bank (PDB) using the Macromolecular Structure Database (MSD)[184]. In the main dataset, we look at structures that have been solved to a resolution less than or equal to 1.8Å. The structures found are placed into groups that have identical sequences and crystallographic space groups. All structures were also checked using the WHATCHECK protein structure verification system[185]. Any structures with problems in unit cell scaling were rejected, as these errors may skew measurements of distances in these structures.

The groups of structures are then filtered to remove redundancy by ensuring that no two groups share the same CATH identifier (to the superfamily level). We then select those groups containing at least one structure of the apo form of the enzyme, and at least one suitable for representing a ligand bound form of the enzyme, preferably a transition state or intermediate form.

We also analyse two enzymes, shown in Table 7.2, with only lower resolution structures available, but where reduced dynamics have been observed in previous experiments. The hypothesis states that reduced dynamics is linked to bond shortening, so we might expect to see an effect in these cases even if the resolution is poor.

We also look at two proteins, shown in Table 7.3, that function as receptors rather than enzymes but have particular binding properties that make them interesting to investigate: Streptavidin is a very strong binder of biotin and reduced dynamics have been observed upon binding, so it is predicted to show a large bond length reduction. In contrast, the oxygen binding molecule hemoglobin is predicted to expand upon the binding of oxygen because of the co-operative nature of the four subunits in a hemoglobin tetramer. This co-operative binding makes the affinity of the hemoglobin tetramer go down once one subunit has bound oxygen.

We only consider main chain hydrogen bonds in this study (NH-O). These bonds are the most numerous type of hydrogen bond in protein structures and are responsible for holding both 2^o and 3^o structures together, making them the easiest and most likely place to observe bond shortening, though the strengthening of other non-covalent bonds is predicted to occur as well. We do not expect length changes in covalent bonds, because, at room temperature, they are too strong for their thermal vibrations to contribute to the overall entropy. We also ignore any larger scale changes in the enzyme structure, though in no case in this dataset are significant larger scale conformational changes observed.

HBPLUS[186] was run with default settings to identify all the main chain hydrogen bonds within each structure. For each apo/ligand-bound pair all such bonds present in both structures were identified. For each bond the following parameters were measured: The N-O distance for both apo and ligand bound states, the shortest distance from either N or O to the closest ligand atom and the average temperature

Enzyme Name	EC	PDB Codes				Ligand Type
	Number	Apo		Ligand		
Biliverdin-β Reductase	1.5.1.30	1HDO	(1.15Å)	1HE2	(1.20Å)	Substrate Analog
				1HE3	(1.40Å)	Substrate
				1HE4	(1.40Å)	Substrate
				1HE5	(1.50Å)	Substrate Analog
Spermidine Synthase	2.5.1.16	1INL	(1.50Å)	1JQ3	(1.80Å)	TS Analog
Endoglucanase A	3.2.1.4	1IS9	(1.03Å)	1KWF	(0.94Å)	Substrate
Endoglucanase Cel5A	3.2.1.4	1A3H	(1.57Å)	6A3H	(1.68Å)	Intermediate
Xylanase	3.2.1.8	2BVV	(1.50Å)	1BVV	(1.80Å)	Intermediate
Endopolygalacturonase	3.2.1.15	1K5C	(0.96Å)	1KCC	(1.00Å)	Product
				1KCD	(1.15Å)	Product
Lysozyme	3.2.1.17	1JSE	(1.12Å)	1LJN	(1.19Å)	Substrate Analog
Elastase	3.4.21.36	1QNJ	(1.10Å)	1GVK	(0.94Å)	Intermediate
Proteinase K	3.4.21.64	1IC6	(0.98Å)	1P7V	(1.08Å)	Substrate Analog
Pyrophosphatase	3.6.1.1	1I40	(1.10Å)	1I6T	(1.20Å)	Substrate
Aldolase	4.1.2.4	1P1X	(0.99Å)	1JCJ	(1.10Å)	Intermediate
Xylose Isomerase	5.3.1.5	1MUW	(0.86Å)	1S5M	(0.98Å)	Substrate Analog
				1S5N	(0.95Å)	Intermediate

Table 7.1: High resolution set of enzyme structures and their PDB codes used in this study. TS = transition state

Enzyme Name	EC	PDB Codes				Ligand Type
	Number	Apo		Ligand		
Glyceraldehyde-3-Phosphate Dehydrogenase	1.2.1.12	1DC5	(2.00Å)	1DC6	(2.00Å)	Cofactor
Carboxypeptidase A	3.4.17.1	5CPA	(1.54Å)	7CPA	(2.00Å)	Transition State Analog

Table 7.2: Enzymes and PDB codes used in this study where reductions in dynamics have been previously observed.

Enzyme Name	EC	PDB Codes				Ligand Type
	Number	Apo		Ligand		
Streptavidin	-	1SWB	(1.90Å)	1MK5	(1.40Å)	Biotin
Hemoglobin	-	1BZ0	(1.50Å)	1YHE	(2.10Å)	Oxygen

Table 7.3: Receptors and binding proteins used in this study used in this study.

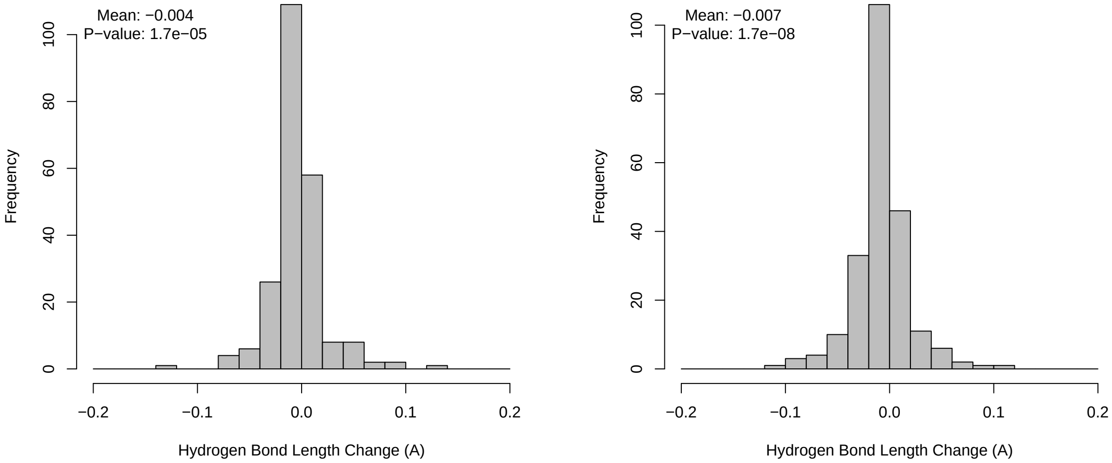
factor of the N and O atoms involved in the bond.

If bond length shrinking is observed then the change in bond length from apo to ligand bound states should be negative. To test whether the mean of the bond length changes was significantly different from zero, a one sample Wilcoxon signed rank test was performed on the data from each enzyme and on the data obtained from all enzymes. The null hypothesis for the test is that the average bond length change is zero.

7.3 Results

7.3.1 Xylose Isomerase

Xylose isomerase catalyses the interconversion of aldose and ketose sugars (primarily xylose to xylulose and glucose to fructose). This involves a ring opening step followed by isomerisation. In this dataset we find two ligand bound structures: the first complexed with glucose (1S5M), representing a substrate bound prior to ring opening, and the second complexed with xylitol (1S5N), representing a substrate analog in the ring opened state[187]. Histograms of the hydrogen bond length changes between the apo structure (1MUW) and the two ligand bound forms are shown in Figure 7.1. Both liganded forms show a small average shortening of hydrogen bond lengths. The glucose bound structure shrinks each hydrogen bond by 0.004Å on average. The Wilcoxon p-value is 1.7×10^{-5}. The xylitol bound structure has an average shrinkage of 0.007Å and a Wilcoxon p-value of 1.7×10^{-8} suggesting both these results are statistically significant.



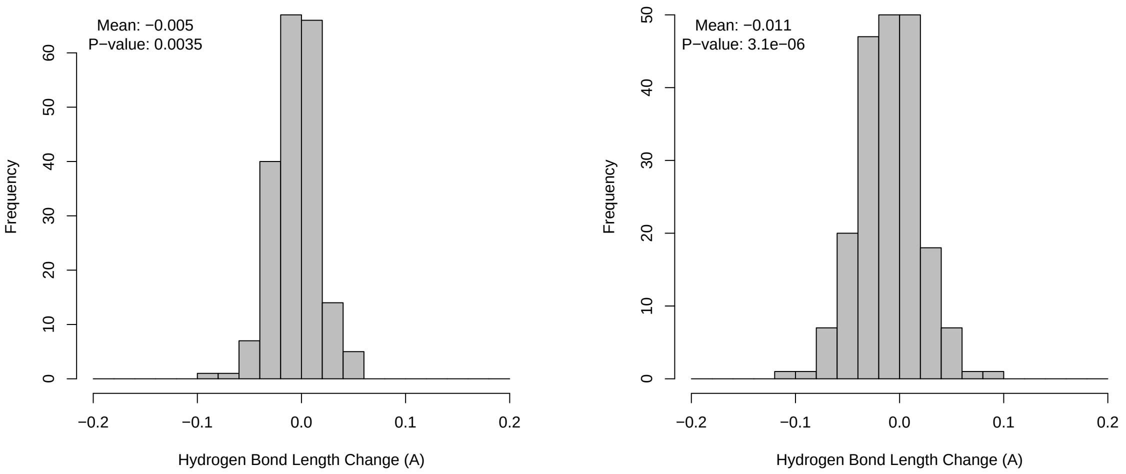
(a) Xylose Isomerase Bound to Glucose (1S5M).

(b) Xylose Isomerase Bound to Xylitol (1S5N).

Figure 7.1: Histograms of the change in hydrogen bond lengths observed in xylose isomerase upon binding of glucose and xylitol. Unless noted, all graphs are drawn with x-axis limits of -0.2Å and 0.2Å for clarity, a few outliers have changes outside these limits. Outliers are included in all other calculations.

7.3.2 Endopolygalacturonase

Endopolygalacturonase catalyses the degradation of pectin by hydrolysing the α-1,4-glycosidic bonds between α-D-galacturonic acid residues in the pectin main chain. Two ligand bound structures are available, one complexed with a single galacturonic acid residue (1KCC), representing half of the product bound form, and one with two galacturonic acid residues (1KCD), representing a full product bound form[188]. Histograms of the hydrogen bond length changes between the apo structure (1K5C) and the two ligand bound forms are shown in Figure 7.2. Both structures show a small average shortening of hydrogen bond lengths. The structure complexed with one galacturonic acid shrinks each hydrogen bond by 0.005Å (p-value=0.0035), the structure with two galacturonic acid residues shrinks each hydrogen bond by 0.011Å (p-value=3.1×10^{-6}).



(a) Endopolygalacturonase bound to single galacturonic acid residue (1KCC).

(b) Endopolygalacturonase bound to two galacturonic acid residues (1KCD).

Figure 7.2: The change in hydrogen bond lengths observed in endopolygalacturonase upon binding of one or two galacturonic acid residues.

7.3.3 Biliverdin-IXβ Reductase

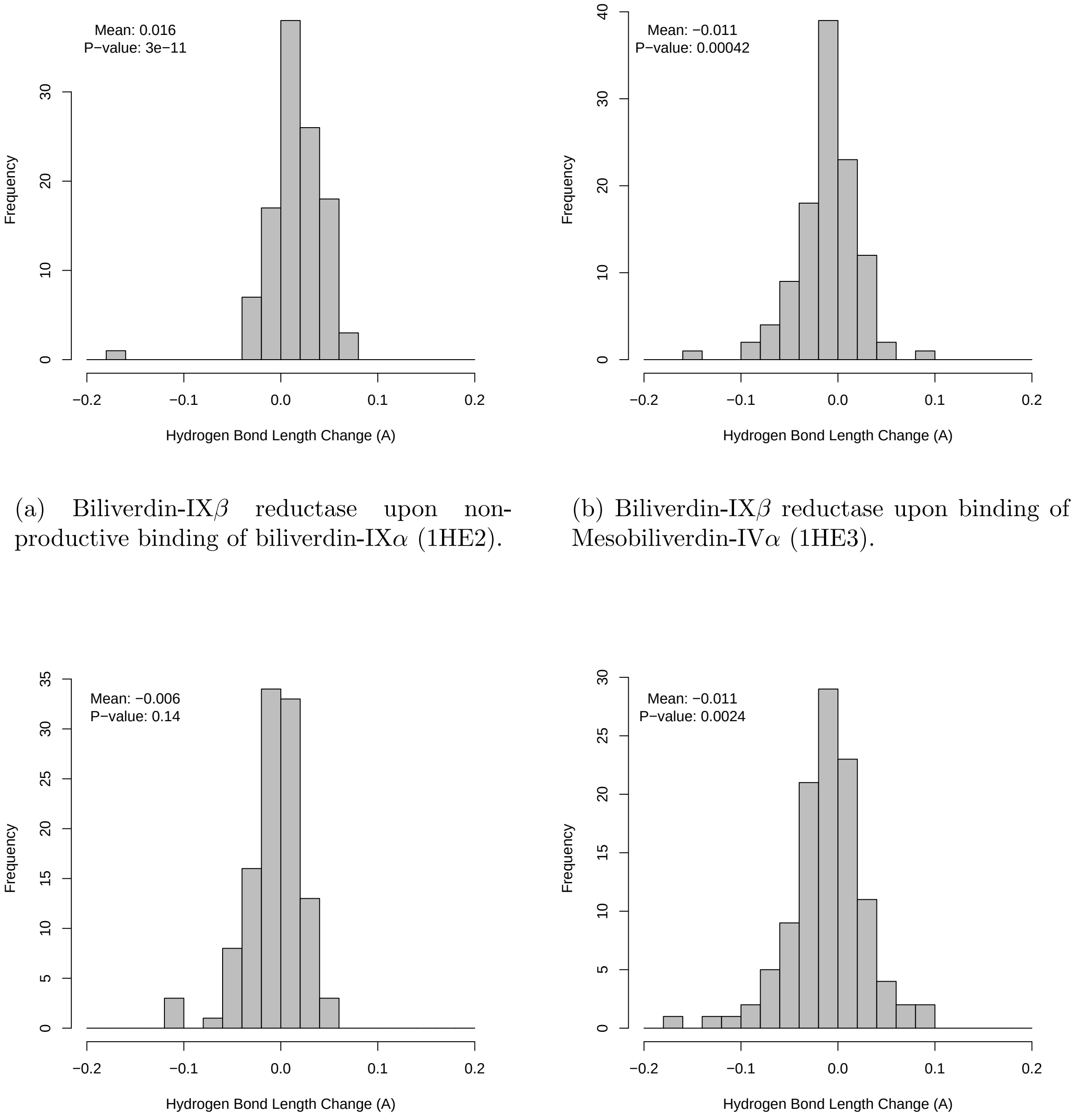
Biliverdin-IXβ reductase (BVR-B) primarily catalyses the production of bilirubin-IXβ from biliverdin-IXβ, though it can catalyse the reduction of a wide range of substrates via an NADP cofactor. All the structures of BVR-B have NADP bound; one structure with no other ligands is available (1HDO), that we use as the apo form, and several other structures with NADP and bound ligands/inhibitors are also available[189].

The first ligand bound structure (1HE2) is bound to biliverdin-IXα, a biliverdin isomer that inhibits BVR-B. This ligand binds to the active site of BVR-B in a non-productive mode rotated 90^o relative to the binding of the true substrates, due to steric hindrance from propionate groups attached to the biliverdin.

This non-productively bound form shows an increase in average hydrogen bond length of 0.016Å (p-value=3×10^{-11}). In contrast, the two productively bound substrates: mesobiliverdin-IVα (1HE3) and FMN (1HE4) lead to a decrease in average hydrogen bond length of 0.011Å (p-value=4.2×10^{-4}) and 0.006Å (p-value=0.14) respectively. A third ligand bound structure with lumichrome bound (1HE5) also shows a shrinking of 0.011Å (p-value=0.0024). Lumichrome is an inhibitor of BVR-B that binds in the productive mode. Histograms of the hydrogen bond length changes are shown in Figures 7.3.

7.3.4 Pyrophosphatase

Pyrophosphatase catalyses the hydrolysis of pyrophosphate (PP_i) into two molecules of orthophosphate (P_i). Structures of the apo (1I40) and substrate bound states (1I6T) have been solved in the presence of Ca^{2+} ions which are a strong inhibitor of pyrophosphatase[190]. The histogram of hydrogen bond length changes is shown in Figure 7.4(a). Pyrophosphatase shows a strong shrinking effect with an average hydrogen bond length change of -0.021 (p-value = 0.0019). The relatively high p-value reflects the small size of pyrophosphatase leading to it having a structure with fewer hydrogen bonds than some of the other enzymes.



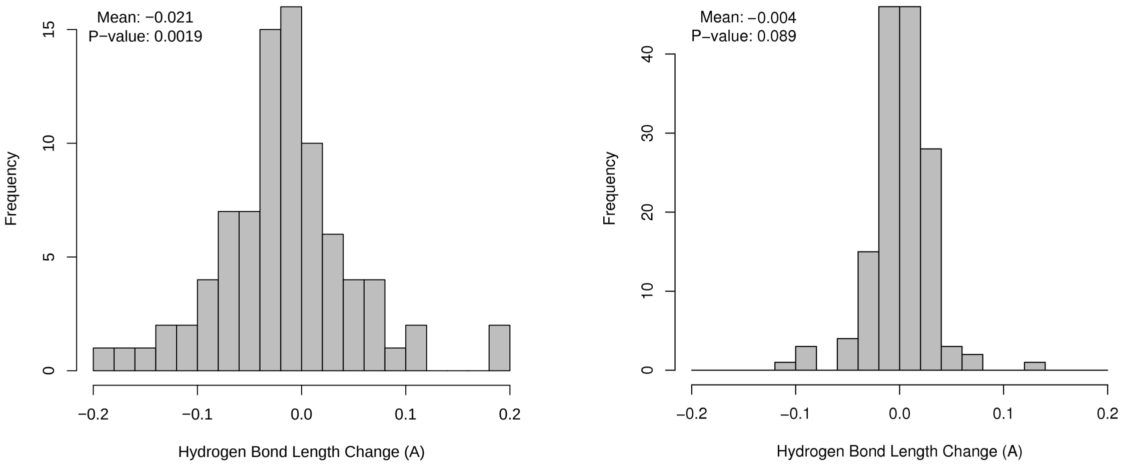
(a) Biliverdin-IXβ reductase upon non-productive binding of biliverdin-IXα (1HE2).

(b) Biliverdin-IXβ reductase upon binding of Mesobiliverdin-IVα (1HE3).

(c) Biliverdin-IXβ reductase upon binding of FMN (1HE4).

(d) Biliverdin-IXβ reductase upon binding of Lumichrome (1HE5).

Figure 7.3: Histograms of the change in hydrogen bond lengths observed in Biliverdin-IXβ reductase upon binding of ligands.



(a) Pyrophosphatase upon binding to pyrophosphate (1I6T)

(b) Proteinase K upon binding to heptapeptide inhibitor.

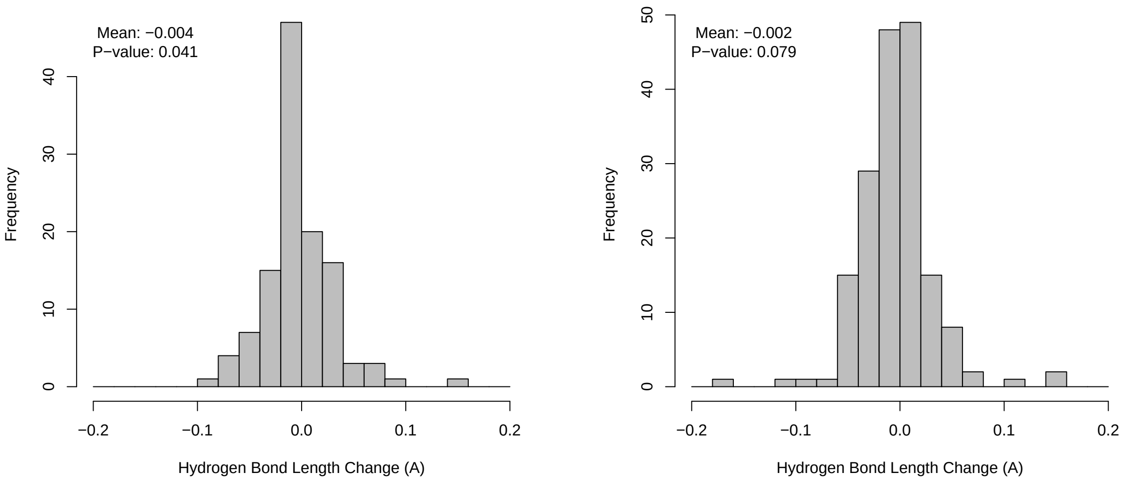
Figure 7.4: Histograms of the change in hydrogen bond lengths observed in pyrophosphatase and proteinase K upon binding of ligands.

7.3.5 Proteinase K

Proteinase K is a subtilisin-like serine protease that catalyse the hydrolysis of peptide bonds in proteins. Structures of the apo enzyme (1IC6) and the enzyme complexed with a heptapeptide inhibitor (1P7V) have been solved[191]. We do not observe any significant change in hydrogen bond length upon binding of the inhibitor. The average change in hydrogen bond length is +0.004 (p-value=0.089). Figure 7.4(b) shows the histogram of the hydrogen bond length changes.

7.3.6 Elastase

Elastase is a trypsin-like serine protease that catalyse the hydrolysis of peptide bonds in proteins. Structures of the apo enzyme (1QNJ)[192] and the acyl-enzyme complex with an inhibitor (1GVK) have been solved[193]. Considering all the data, we do not observe any significant change in hydrogen bond length upon formation of the acyl-enzyme intermediate; the average change in hydrogen bond length is +0.0001Å (p-value=0.063). However, this data is skewed by the inclusion of an outlier that has a change in hydrogen bond length of 0.5Å. This bond is between two adjacent residues in a tight β-turn. In the apo structure the distance is given as 2.65Å, very short for a hydrogen bond, and further investigation reveals that this bond is in a region that 'could not be localised completely' according to the authors[192]. If this single data point is excluded, the average hydrogen bond length is found to be -0.004Å (p-value=0.041). Figure 7.5(a) shows the histogram of the hydrogen bond length changes.



(a) Elastase upon formation of a covalent complex (1GVK). One outlying bond shows a length increase of 0.5Å, which is off the scale of this graph.

(b) Deoxyribose-phosphate aldolase upon formation of a Schiff Base intermediate (1JCJ).

Figure 7.5: Histograms of the change in hydrogen bond lengths observed in elastase and deoxyribose-phosphate aldolase upon binding ligands.

7.3.7 Deoxyribose-phosphate Aldolase

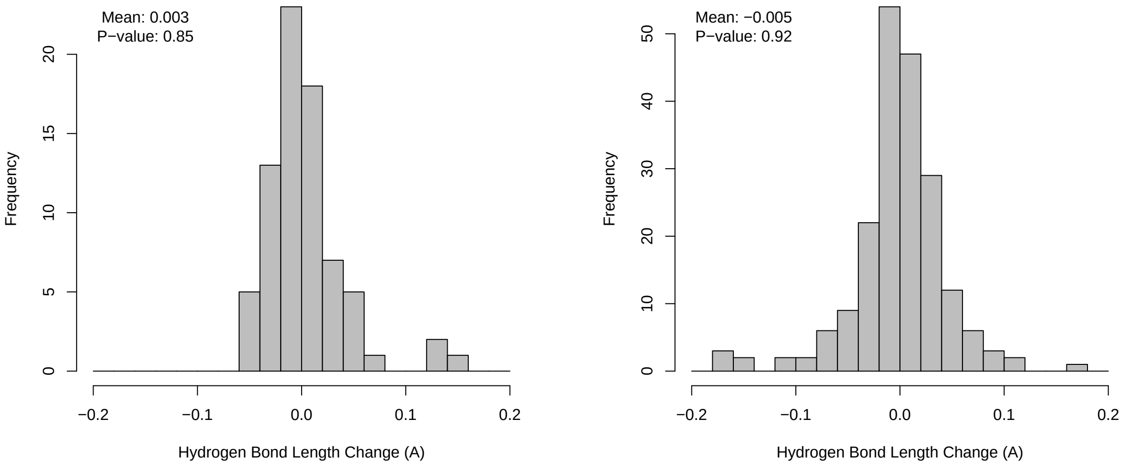
Deoxyribose-phosphate aldolase catalyses the addition of acetyldehyde to D-glyceraldehyde-3-phosphate to form D-2-deoxyribose-5-phosphate. It performs catalysis by forming a Schiff base intermediate where the substrate is bound covalently to an active site lysine. Structures of the apo (1P1X) and Schiff base form (1JCL) have been solved[194]. No significant change in hydrogen bond length is observed when we compare the apo and Schiff base forms. The average change in hydrogen bond length is -0.002Å (p-value = 0.079). Figure 7.5(b) shows the histogram of the hydrogen bond length changes.

7.3.8 Lysozyme

Lysozyme catalyses the hydrolysis of glycosidic linkages between residues in peptidoglycan polymers. A structure of the apo form has been solved, as well as a structure with Di(n-acetyl-d-glucosamine) (NAG2) bound[195]. No significant change in hydrogen bond length is observed between these two structures. The average hydrogen bond length change is +0.003Å (p-value = 0.85). Figure 7.6(a) shows the histogram of the hydrogen bond length changes.

7.3.9 Endoglucanase A

Endoglucanase A catalyses the hydrolysis of glycosidic linkages in oligosaccharides. The structures of the apo enzyme (1IS9) and the enzyme complexed with an oligosaccharide (1KWF) have been solved[196]. No significant change in hydrogen bond length is observed in the substrate bound structure. The average change in hydrogen bond length is -0.005Å (p-value = 0.92). Figure 7.6(b) shows the histogram of the hydrogen bond length changes.



(a) Lysozyme upon binding di(n-acetyl-d-glucosamine) (1LJN)

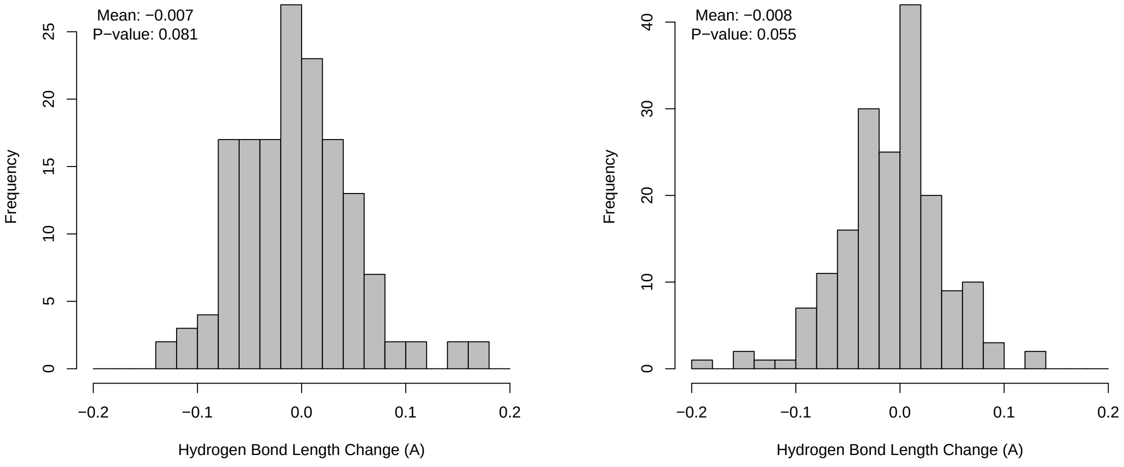
(b) Endoglucanase A upon binding oligosaccharide (1KWF)

Figure 7.6: Histograms of the change in hydrogen bond lengths observed in lysozyme and endoglucanase A upon binding ligands.

7.3.10 Spermidine Synthase

Spermidine Synthase (PAPT) catalyses the synthesis of spermidine from putrescine and S-adenosyl-L-methionine. The structure of PAPT from Thermotoga maritima has been solved in both apo forms and bound to a mixed substrate-product analog (AdoDATO)[197].

The average hydrogen bond length change is -0.007Å (p-value 0.08), although this change is larger than some other changes, the lower resolution of the spermidine synthase structures makes it harder to determine the significance of the change. Figure 7.7(a) shows the histogram of the hydrogen bond length changes.



(a) Spermidine synthase upon binding to AdoDATO (1JQ3)

(b) Endoglucanase Cel5A upon formation of a covalent intermediate.

Figure 7.7: Histograms of the change in hydrogen bond lengths observed in spermidine synthase and endoglucanase cel5A upon binding ligands.

7.3.11 Endoglucanase Cel5A

Endoglucanase Cel5A catalyses the hydrolysis of glycosidic linkages in oligosaccharides. The use of slow-turnover fluoridated substrates at low pH has allowed the structures of the apo[198] and covalently bound intermediate forms to be solved[158].

The covalent intermediate (which we might expect to resemble the transition state) shows an average hydrogen bond length change of -0.008Å (p-value: 0.055) compared to the apo form. As with spermidine synthase, the lower resolution of the structures makes determining significance harder. Figure 7.7(b) shows the histogram of the hydrogen bond length changes.

7.3.12 Xylanase

Xylanase catalyses the hydrolysis of xylan, a hemi-cellulose sugar. As with Cel5A (see above) the use of fluoridated substrate analogs allow the capture of covalent intermediates[199].

The average hydrogen bond length change observed in forming the covalent intermediate is +0.009Å (Wilcoxon p-value: 0.19). Figure 7.8 shows the histogram of the hydrogen bond length changes.

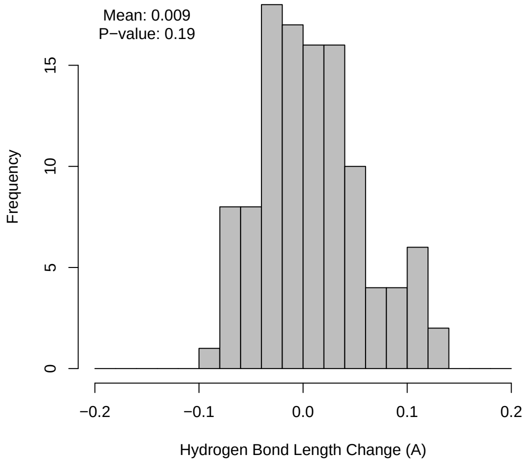
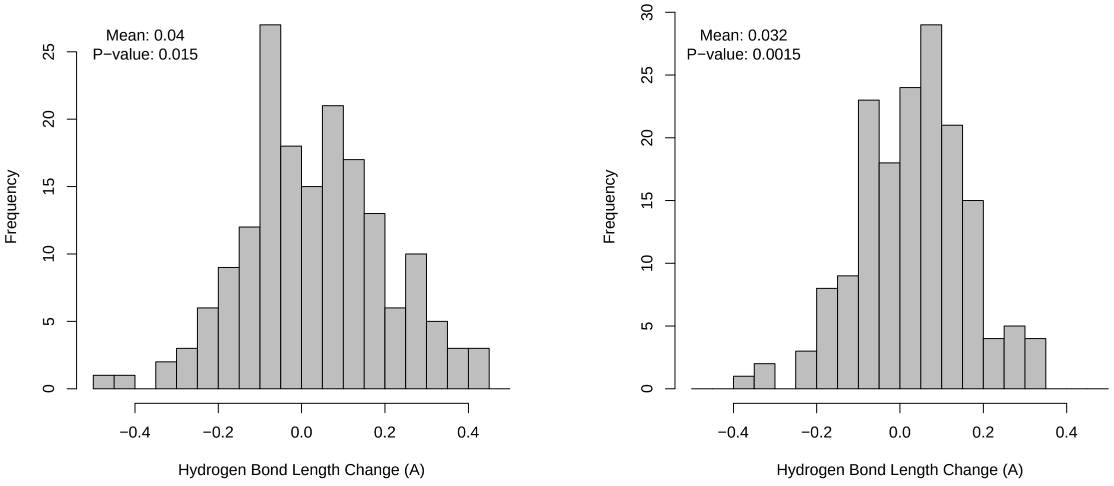


Figure 7.8: Histogram of the change in hydrogen bond lengths observed in Xylanase upon formation of the covalent intermediate.

7.3.13 Glyceraldehyde phosphate dehydrogenase

Glyceraldehyde-3-phosphate dehydrogenase catalyses the phosphorylation of glyceraldehyde-3-phosphate with concomitant reduction of NAD$^+$. Structures of the apo and NADH bound forms have been solved[22]. An observation has been made of reduced dynamics in the NADH bound structure [Williams *et al*, unpublished], so it is predicted that there will be hydrogen bond length changes to compensate for this. However, the only available structures are solved at 2Å resolution - substantially lower than the resolution of the structures used earlier.

We find an average hydrogen bond length change of +0.040Å (p-value: 0.015). This is a large increase relative to the changes seen previously, and the p-value suggests it is statistically significant. However, the low resolution of the structures make making firm conclusions from this data difficult and is reflected in the spread of changes shown in Figure 7.9(a). Note that the x-axis in this plot has a different scale from previous plots.



(a) GAPDH upon binding of NADH (1DC6)

(b) Carboxypeptidase A upon binding of a transition state analog (7CPA).

Figure 7.9: Histograms of the change in hydrogen bond lengths observed in GAPDH and carboxypeptidase A upon binding ligands. Note that the x axis has been expanded to accommodate the range of changes observed in these lower resolution structures.

7.3.14 Carboxypeptidase A

Carboxypeptidase A catalyses the hydrolysis of the C-terminal peptide bond of peptides and proteins, with a preference for hydrophobic side chains. The structures of the apo and transition state analog bound forms have been solved[200, 201]. Again, reduction in dynamics has been reported within the transition state, but the transition state structure is only solved to 2Å resolution, which is low compared to the previous structures.

We find an average hydrogen bond length change of +0.032Å (p-value: 0.0015). As with GAPDH, the low resolution of the available structures gives a large spread of values, but the hydrogen bond length increase in this transition state still appears to be significant and in direct contradiction of the change predicted by the hypothesis. Figure 7.9(b) shows the histogram of the hydrogen bond length changes.

7.3.15 Streptavidin

Streptavidin is not an enzyme, but it is note worthy for binding biotin with extremely high affinity[202]. The binding is also a highly exothermic event and is accompanied by a loss of dynamics and hence entropy in the receptor tetramer[203]. It is predicted therefore, that streptavidin will show a large decrease in hydrogen bond length upon binding biotin. Crystal structures of the bound and unbound forms of streptavidin have been solved[204].

We find an average hydrogen bond length change of -0.054Å (p-value: 0.0011). This represents a significant bond length decrease throughout the structure. Again however, the resolution of the structures used is not as high as in the previous examples (1.9Å) so care must be taken in analysing this result.

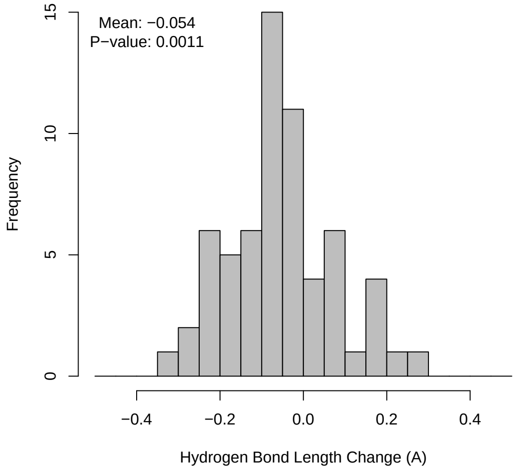
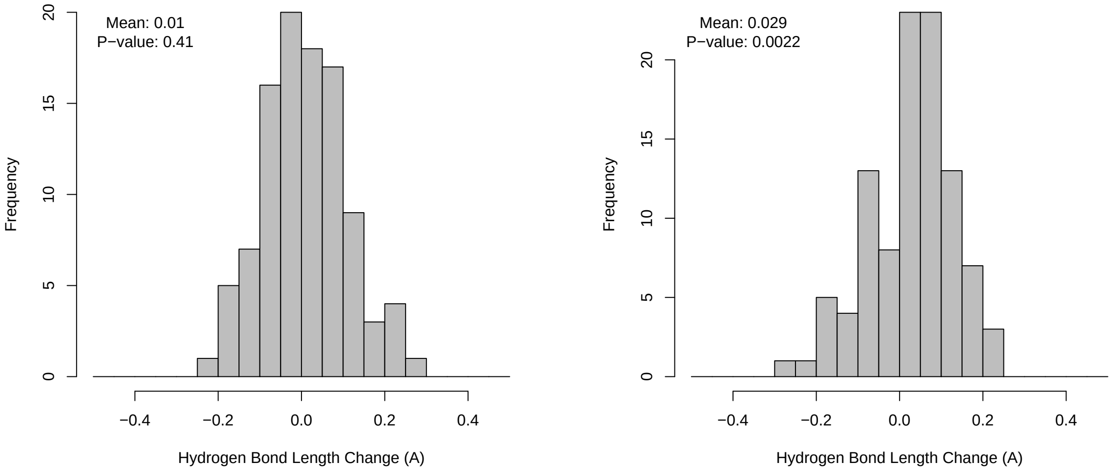


Figure 7.10: Histogram of the change in hydrogen bond lengths observed in streptavidin upon binding of biotin.

7.3.16 Hemoglobin

Hemoglobin binds oxygen for transport around the body in the blood stream. Hemoglobin is a tetramer made from two related chains (each of which occurs twice in the tetramer). The crystal structures of the full deoxy and full oxygenated forms have been solved[205, 206]. It is predicted that because of the negative cooperativity shown by hemoglobin (binding oxygen to one chain changes the structure of the tetramer such that further binding to the other tetramers is more favourable) hemoglobin should expand on binding oxygen.

We observe an average hydrogen bond length change of +0.01 (p-value: 0.41) in hemoglobin chain A and +0.029 (p-value: 0.0022) in chain B upon binding of oxygen to each subunit of the tetramer. As predicted therefore, both chains show an increase in hydrogen bond length, and there is also a suggestion that the two chains do not respond equally to binding oxygen. The hydrogen bond length changes are shown in Figure 7.11.



(a) Hemoglobin A chain upon binding of oxygen.

(b) Hemoglobin B chain upon binding of oxygen.

Figure 7.11: The change in hydrogen bond lengths observed in hemoglobin upon binding of oxygen.

7.3.17 Group Results

Table 7.4 shows all the average hydrogen bond length changes and significance values calculated for each enzyme. Those pairs highlighted in bold are those chosen as the closest to the transition state form (and hence we might expect to see greatest shrinking). Figure 7.12 shows a histogram of all the hydrogen bond length changes for these transitions in the high resolution enzyme structures (GAPDH, carboxypeptidase A, streptavidin and hemoglobin are excluded). The average hydrogen bond length change in this subset of the data is -0.005Å (p-value = 3×10^{-6}). The standard deviation of this distribution is 0.05Å, and the number of bonds is 1793, leading to a standard error of the mean of 0.001Å.

We might expect the bond length changes to be related to the distance of each bond from the active site (with the largest changes occurring closest to the ligand).

Enzyme Name	EC Number	PDB (Apo)	PDB (Ligand)	Number of Hydrogen Bonds	Hydrogen Bond Length Change	P-value	Ligand
High Res. Structures							
Biliverdin-β Reductase	1.5.1.30	1HDO	1HE2	110	+0.016	6×10^{-8}	Biliverdin$IX-\alpha$
			1HE3	111	**-0.011**	3×10^{-4}	Mesobiliverdin$IV-\alpha$
			1HE4	111	-0.006	0.03	FMN
			1HE5	111	-0.011	0.003	Lumichrome
Spermidine Synthase	2.5.1.16	1INL	1JQ3	156	**-0.007**	0.08	Transition state analog
Endoglucanase A	3.2.1.4	1IS9	1KWF	202	**-0.005**	0.2	Oligosaccharide
Endoglucanase Cel5A	3.2.1.4	1A3H	6A3H	180	**-0.008**	0.055	Covalent Intermediate
Xylanase	3.2.1.8	2BVV	1BVV	110	**+0.009**	0.19	Covalent intermediate
Endopolygalacturonase	3.2.1.15	1K5C	1KCC	203	-0.005	0.03	Galacturonate
			1KCD	204	**-0.011**	2×10^{-5}	2 × Galacturonate
Lysozyme	3.2.1.17	1JSE	1LJN	75	**+0.003**	0.4	Di-n-acetylchitobiose
Elastase	3.4.21.36	1QNJ	1GVK	119	0	1	Acetyl-peptide
			1GVK	119	**-0.005**	0.06	(outlier removed)
Proteinase K	3.4.21.64	1IC6	1P7V	150	**+0.003**	0.2	Heptapeptide
Pyrophosphatase	3.6.1.1	1I40	1I6T	87	**-0.021**	0.01	Pyrophosphate
Aldolase	4.1.2.4	1P1X	1JCL	172	**-0.002**	0.6	2-deoxy D-ribose 5-P
Xylose Isomerase	5.3.1.5	1MUW	1S5M	225	-0.004	0.02	Glucose
			1S5N	225	**-0.007**	0.002	Xylitol
Low Res. Structures							
GAPDH	1.2.1.12	1DC5	1DC6	175	**+0.040**	0.015	Cofactor
Carboxypeptidase A	3.4.17.1	5CPA	7CPA	166	**+0.032**	0.0015	Transition state analog
Receptors							
Streptavidin	-	1SWB	1MK5	63	**-0.054**	0.0011	Biotin
Hemoglobin A chain	-	1BZ0	1YHE	101	**+0.010**	0.41	Oxygen
Hemoglobin B chain	-	1BZ0	1YHE	101	**+0.029**	0.0022	Oxygen

Table 7.4: Average hydrogen bond length changes and significance values for all the enzymes in the dataset. Data used in the group results are shown in bold face.

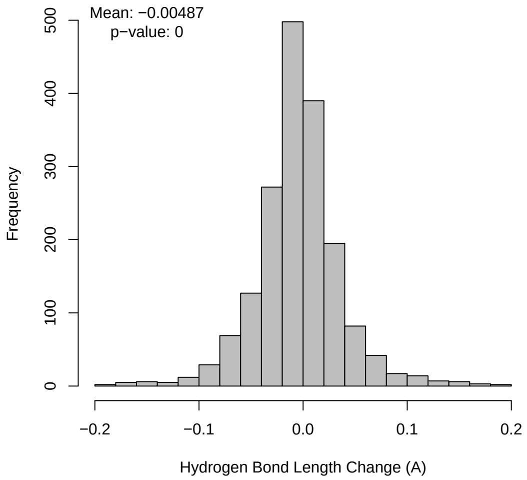


Figure 7.12: Histogram of the change in hydrogen bond lengths observed in all high resolution enzymes.

In Figure 7.13 we have binned these data by the distance from the ligand and averaged the hydrogen bond length change in each bin. We see that, in the range 0-25Å, the magnitude of the observed shrinking decreases from 0.010Å in those hydrogen bonds within 5Å of the ligand to 0.004Å in those hydrogen bonds 25Å from the ligand. Error bars are drawn representing the standard error of the mean for each bin, the relatively low number of bonds in direct contact with the ligand (which the 0-5Å bin represents) mean that the error is particularly high in this bin.

We also might expect to see different changes in bonds of different starting lengths. For instance, longer bonds will have more capacity to shrink than shorter bonds. Figure 7.14(a) shows the data for each bond binned by their length in the apo form of the enzyme. For each bin the average length change is calculated. We can observe that the very short hydrogen bonds in the apo form tend to lengthen when the ligand binds whilst the very long hydrogen bonds tend to shrink. Hydrogen bonds in the middle range (2.7Å-3.2Å) show little change relative to these

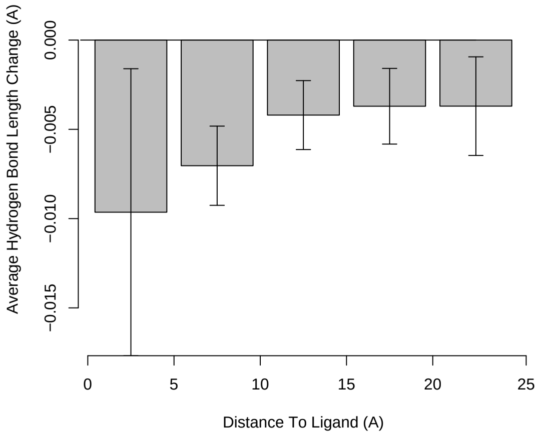
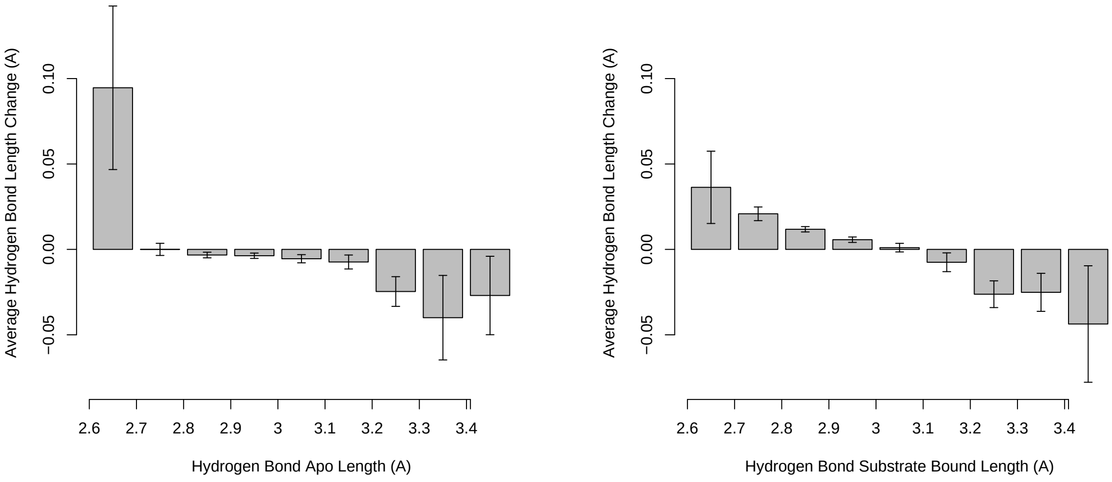


Figure 7.13: Histogram of the change in hydrogen bond lengths observed for hydrogen bonds at different distances from the ligand. Error bars represent the standard error of the mean for each bin.

outliers, but a small overall shrinkage is observed. The small number of bonds at the extremes of the range mean that the errors in the averages of these measurements are much higher than for the hydrogen bonds of 'normal' length.

To see whether these changes are really due to ligand binding, we also look at the changes in the opposite direction (from ligand bound to apo). Figure 7.14(b) shows the same data as Figure 7.14(a) but plotted with the ligand bound hydrogen bond length and with the signs of the hydrogen bond length changes reversed. We can see that the same general trend is present: long hydrogen bonds in the ligand bound structure tend to be shorter in the apo and short hydrogen bonds tend to be longer. This suggests that these changes simply reflect the fact that we are sampling at the extremes of the bond length distribution when we look at these bonds. *i.e.* when we measure a very long bond in the apo form it is more likely to be shorter, than longer, when we measure it a second time in the ligand bound state. We note however, that, in the apo to ligand bound set, there is a consistent small shortening



(a) Histogram of the change in hydrogen bond lengths observed for hydrogen bonds of different lengths in the apo form. Error bars represent the standard error of the mean of each bin.

(b) Histogram of the change in hydrogen bond lengths observed for hydrogen bonds of different lengths in the ligand bound form. Error bars represent the standard error of the mean of each bin.

Figure 7.14: The relationship between hydrogen bond length and the change in hydrogen bond length observed.

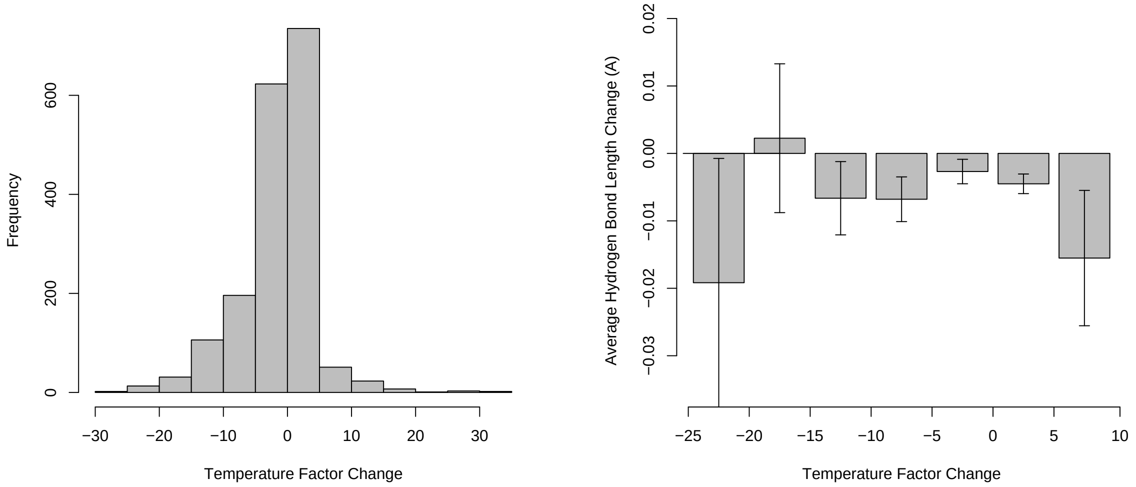
in the 'normal' length bonds (2.7Å-3.2Å) not observed in the ligand bound to apo set.

Figure 7.15(a) shows a histogram of the change in the average temperature factor of the atoms in each bond when going from apo to ligand bound states. The average temperature factor change is -2, and Figure 7.15(a) confirms that ligand binding does reduce the dynamics of most parts of most of the enzymes. Figure 7.15(b) shows the data for hydrogen bond length changes binned by temperature factor change. Surprisingly, there appears to be no correlation between the change in temperature factor and the hydrogen bond length change. Since most temperature factor normalising procedures only work within a structure and not between two structures it is difficult to compare temperature factors between structures which may explain why no correlation is observed.

7.4 Discussion

The enzymes within this dataset give a mixed result as far as the hypothesis of contracting non-covalent bonds is concerned. Only five of the 12 high resolution enzymes: biliverdin-β reductase, endoglucanase Cel5A, endopolygalacturonase, pyrophosphatase and xylose isomerase show significant shrinkage. With average hydrogen bond length changes of the order of 0.2%-0.4% of the length of an average hydrogen bond (0.005Å-0.01Å). Elastase shows some shrinkage with borderline significance once a single outlying hydrogen bond is removed. However, none of the enzymes in this group show significant bond length increases.

The two enzymes GAPDH and carboxypeptidase A are both examples where reductions in dynamics have been observed upon binding of ligands, and so we would expect both enzymes to show significant reduction in bond lengths going from apo



(a) Histogram of the temperature factor changes observed.

(b) Histogram of the change in hydrogen bond lengths observed for hydrogen bonds of different changes in B-factor. Error bars represent the standard error of the mean of each bin.

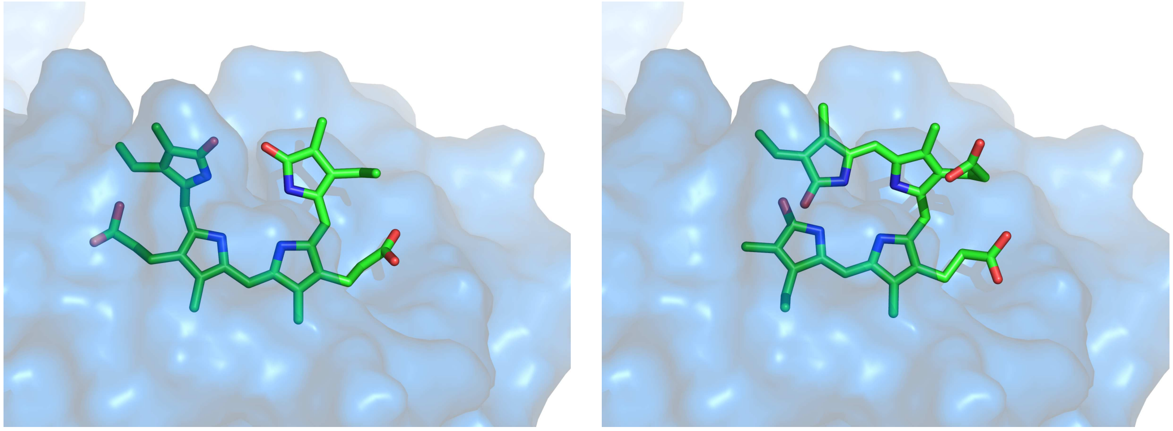
Figure 7.15: The relationship between the change in the temperature factor of the hydrogen bond and the change in hydrogen bond length observed.

to ligand bound structures. However, we see significant bond length increases in both cases, directly contradicting the hypothesis. It is noticeable though, that these structures, because of their lower resolution have a much larger co-ordinate error. This is reflected in the spread of hydrogen bond length changes observed (shown in Figure 7.9), which suggests that the results obtained from these examples are less reliable than for the high resolution structures.

There are two possible explanations for the variation in average hydrogen bond length changes between enzymes. Firstly, the shrinking may truly occur in all enzymes, but in some cases crystallisation artifacts make observing the changes impossible. This could be because the ligands used are not perfect substrate or transition state analogs, or crystal packing prevents the changes taking place. Secondly, it is possible that shrinking only occurs in some enzymes, and that these have evolved in such a way that they can take advantage of hydrogen bond strengthening in catalysis, whilst others have not, either because their structure makes it difficult, or because the particular details of their mechanism means it is not needed for efficient catalysis.

Given that we observe a range of responses to ligand binding, and the nature of these ligands is quite heterogeneous (for instance, not all the ligands are transition state analogs) we have to take care when treating these enzymes as a single group. However, the hypothesis is that shrinking is a common mechanism in enzymes, so it seems appropriate to perform analysis on these enzymes as a whole. When we consider all eleven enzymes there is a very slight overall shrinking of 0.005Å representing an average hydrogen bond length change of just 0.2%. This change is almost an order of magnitude less than the 1% bond length change that has been predicted, and so seems to suggest against the bond shrinking effect being a significant general mechanism of enzyme catalysis.

We do see a number of interesting individual cases, most striking being that of BVR binding to different ligands. The binding of biliverdin-IX-α (Figure 7.16(b)) causes a significant lengthening of hydrogen bonds relative to the apo form, whilst binding mesobiliverdin-IV-α (Figure 7.16(a)) causes a significant shortening of hydrogen bonds (+0.016Å compared to -0.011Å). The binding of these two ligands occurs at the same pocket on the surface, but the biliverdin-IX-α is bound unproductively in a conformation rotated 90^o from the productive conformation represented by mesobiliverdin-IV-α and the other substrates. We could explain these results by hypothesing that the enzyme has evolved so that unproductive binding is energetically disfavoured compared to productive binding by collectively shrinking non-covalent bonds in the productive form. However, what features of the enzymes structure could account for this subtle effect is a mystery.



(a) BVR bound productively to Mesobiliverdin IV-α.

(b) BVR bound unproductively to Biliverdin IX-α.

Figure 7.16: Two different binding modes of BVR. Binding to Mesobiliverdin IV-α (binding productively) leads to a significant shrinking in the structure, whilst Biliverdin IX-α (binding unproductively) leads to a significant expansion. The unproductive binding mode is rotated 90^o anti-clockwise relative to the productive binding mode in this figure. However, neither shrinking nor expansion is on a scale that would be visible in this figure.

The hydrogen bond contraction hypothesis suggests that hydrogen bond contractions propagate throughout the structure because stabilisation of the ligand binding residue leads to improved hydrogen bonding of those residues with the second shell residues, which in turn leads to improved hydrogen bonding of those residues with third shell residues and so on. We see that the shrinking effect does drop off with increased distance to the ligand, and that shrinking is particularly evident in those residues in direct contact with the ligand, though it must be acknowledged that the errors in these measurements are high. This weak correlation suggests that the shrinking is a genuine effect due to ligand binding, but it would seem that the complexity of protein structure means that propagation of the shrinking effect is not straight-forward.

We also find a correlation between hydrogen bond length changes and the original hydrogen bond length. Short hydrogen bonds tend to lengthen, whilst long hydrogen bonds tend to contract. Hydrogen bonds have an optimum length below which the hydrogen bonding atoms are too close together and above which they are too far apart. It is possible, even in structures which are at their global energy minimum, for single hydrogen bonds to be shorter than their optimum length because they are forced into that conformation by the overall fold of the protein. Put another way, if several hydrogen bonds can form at their optimum length by forcing one hydrogen bond to be shorter then they will force it into that conformation because the overall energy is lower in that state. Ligand binding may change the overall energy minimum and allow the short hydrogen bonds to relax into longer ones.

However, the fact that we see a similar pattern of changes when we consider the ligand bound structures changing into apo structures suggests that this is simply a sampling artefact - samples at the edge of a distribution (very long and short bonds in this case) will tend to move towards the mean rather than away when measured repeatedly. However, there are differences between the two plots, particularly in the range occupied by most ‘normal’ hydrogen bonds (2.7Å-3.1Å), where bonds tend to shorten (if minutely) going from apo to ligand bound, whereas they lengthen going from ligand bound to apo.

We would expect to see a correlation of hydrogen bond length changes with temperature factor changes, since those residues showing the greatest reduction in dynamics should also show the largest enthalpy change and hence decrease in bond length. However, we see no significant correlation, perhaps due to the difficulties in using un-normalised temperature factors between two different structures.

The two receptors streptavidin and hemoglobin are not enzymes, and are not included in the analysis above. However, the binding of biotin by streptavidin is noteworthy for its strength, and highly exothermic character. It may be significant therefore, that, as predicted, streptavidin shows a large reduction in bond length (0.05Å). In contrast, hemoglobin shows an increase in bond length, again as predicted, due to the negative co-operativity displayed by hemoglobin binding oxygen. Interestingly the two chains in hemoglobin respond differently, though the difference may not be statistically significant.

In conclusion, we find evidence of hydrogen bond length changes in some enzymes and in streptavidin. However, these hydrogen bond length changes are extremely small, and are at the borderline of what can be reliably measured using crystallography. The conclusion is uncertain therefore, with respect to the proposed hypothesis and enzyme catalysis. We do find however, that the hydrogen bond length changes are correlated with the distance to the ligand, suggesting that the effect is due to ligand binding, but that propagation of this effect is not easy within enzyme structure. Clearly a larger number of ultra high resolution crystal structures (resolution <1.2Å), preferably with good transition state analogs bound, would be helpful in further study of this effect. However, it may be that techniques such as NMR are more sensitive to these types of small, dynamic changes rather than crystallography.

Chapter 8

Concluding Remarks

8.1 Annotating Enzymes and Analysing Structures

Every piece of work in this thesis relies on the annotation of catalytic residues in enzyme structures made in the CSA, and the subsequent analysis of those structures. Therefore, there are a number of caveats which must be kept in mind when we make conclusions from these analyses.

Considering first the CSA, it is important to remember that a dataset such as this can only be assembled by expert human annotators. This immediately leads to a certain amount of variation and error in the annotations made. The use of a strict definition of what constitutes a 'catalytic residue', has helped to lessen this variation, but terms such as 'transition state stabilisation' and 'substrate activation' are always open to a certain amount of individual interpretation. This is particularly noticeable in the primary literature itself, from which all the annotation is derived from. Different authors may emphasise the roles of different residues depending on the focus of their own interests. This produces a difficult dilemma for the annotator, who is aiming to stay true to the annotation given in the literature, whilst also attempting to maintain consistency between annotations. Furthermore, particularly in an evolving subject such as the study of enzyme mechanisms, there will be cases where the annotation, however well made, is later proved wrong because of new experimental results.

As well as basic errors in the annotation, a second potential source of error in the CSA comes from the transfer of annotation across to homologous enzymes. It may be the case that these homologous enzymes use different catalytic residues to the originally annotated enzyme. This can be true even when the two enzymes are closely related, and the catalytic residues are conserved in both cases. In fact, the transfer of annotation to enzymes with different EC classifications is relatively rare (occurring around 10% of the time), so this is not likely to be major source of error. Every effort has been made to ensure that the data in the CSA is up-to-date and accurate, and we do not believe that errors in annotation will be significant to any of the analyses presented here.

Further annotation has been made to certain enzyme structures in the work presented in Chapters 4 and 7. In these studies, the type of ligand bound to certain structures has been classified according to whether it represents a substrate, transition state or product. There is an inherent difficulty in deciding this, due to the often similar chemical nature of these different ligands, and the use of analogs which are not identical to any of them. Again, every effort has been made to ensure that these annotations are as accurate as possible, but errors are inevitable.

We must also mention certain caveats relevant to any analysis of crystal structures. Firstly, the crystal form of the enzyme may differ from the structure of the functional enzyme in solution, and even where the solution and crystal structures closely match, there will also be errors in the position of the atoms. Recent estimates suggest that ∼3% of residues are incorrectly modelled in many structures. All the analyses presented here use resolution and temperature factor cutoffs, so that those atoms with large errors are excluded[207, 208].

Errors in crystal structures are especially relevant to the analyses performed in Chapter 4, where motions of binding residues of a few angstroms are measured, and Chapter 7, where we measure motions an order of magnitude smaller than this. Such motions are far smaller than the error in the position of each atom. However we believe that by looking at many such measurements (hundreds in a typical structure) we can derive useful information from them. However, this is clearly pushing such analyses to the limits of the detail that they can reveal, and the conclusions from the work in Chapter 7 are tentative for this reason.

8.2 The Structural Mechanics of Enzyme Catalysis

This thesis concerns itself with trying to understand how the structures of enzymes help to facilitate their catalytic activity. In doing so, we have looked at a large range of different enzyme structures in various contexts:

8.2.1 Substrate Binding

The first step of an enzyme catalysed reaction is the binding of the substrates. Binding must be carefully controlled, because binding too tightly prevents catalysis (by stabilising the substrate bound state relative to the transition state), and binding too loosely lowers the efficiency of the enzyme. The dominant theory of substrate binding is induced fit, whereby the enzyme exists in at least two separate states: an ‘open’ state prior to binding, and a ‘closed’ substrate bound state. The change from one to the other is characterised by a closing of the active site around the substrate. However, prior to the work presented here, there was no large scale study of the magnitude and type of these changes observed in crystal structures. Reviews of the subject tend to understandably focus on those, not necessarily representative, examples which undergo extreme or unusual conformational changes.

We find that conformational changes in enzymes are usually surprisingly small upon substrate binding, with RMSDs of less than 1Å between the apo and substrate bound structures being the rule rather than the exception. This is not to imply that these changes are insignificant, since even small changes can fundamentally alter the way in which an enzyme functions. However, many of these changes are less than that observed between two separate apo structures of the same enzyme.

We also find that the catalytic residues are somewhat distinct from the binding residues in terms of the magnitude of the motions observed, with less backbone motion observed in particular. However, we do find that they undergo a similar amount of side chain rearrangement. We suggest that this lack of motion (relative to the binding residues) in the catalytic site, allows the catalytic groups to be largely pre-formed. This has entropic benefits, since the catalytic groups do not have to be restrained when forming the transition state, and presumably also allows them to be more precisely positioned relative to each other and the substrate. We have also examined how enzymes change between the substrate and product bound states in a small set of enzymes. We find that the changes are even smaller than those observed on substrate binding.

8.2.2 Conversion of Substrate to Product

The ways in which enzymes use different residues to catalyse the conversion of substrate to product have, similarly to induced fit motions, been studied in great detail, but generally, studies have looked at single enzymes or family of enzymes at a time. In this work, we have tried to find general trends across all enzymes, to show how catalytic residues are used in enzyme catalysis.

The importance of short range interactions (often hydrogen bonds) between polar and charged residues seems clear from the fact that certain patterns (most obviously the catalytic triad) have evolved multiple times. We find that a number of other important ‘catalytic units’, are found repeatedly in different enzymes. These interactions can provide charges for transition state stabilisation and also modify the properties (particularly the pK_a) of the residues involved. By associating each unit with a particular enzyme function, it may be possible to use the units to predict the function of novel enzymes, and even to design new functions that use the units to perform catalysis.

The sensitivity of some of these units to the precise nature of the interactions is shown by those interactions involving the imidazole ring of histidine. We find patterns in the geometry of interactions between histidine, aspartate and serine, which seem to have functional significance, particularly given the way these residues are used to prime each other, and the role of histidine-aspartate pairs in acid/base chemistry.

This work has focused on local structural arrangements, such as the catalytic triad, but much of an enzymes power is due to global structural effects, which provide the correct environment in the active site for catalysis to take place. These global features include hydrophobic pockets, long range electrostatic effects and the whole network of weak, non-covalent interactions that hold together the enzyme structure.

In Chapter 7, we address the predictions made by one recent hypothesis, suggesting that the non-covalent interactions found throughout an enzyme structure are of functional importance for catalysis, in a way not previously appreciated. The prediction is made, that these bonds should shrink in length, leading to an energetically favourable tightening of an enzyme's structure in the transition state. We have attempted to look for these changes in very high resolution crystal structures, but have found mostly inconclusive evidence for the hypothesis. Very small shrinking is observed in a few cases, but there does appear to be a correlation with the magnitude of shrinking and the distance from the binding site. However, the magnitude of the changes is such that it would be impossible to make firm conclusions on the basis of the results to date. More ultra high resolution structures, specifically of transition state structures, will be required to make further progress in this area, though the use of different experimental data, such as NMR, could also be fruitful.

One aspect of enzyme catalysis that has not been addressed in this study is how catalytic residues interact with the substrates. A similar approach to that used in Chapters 5 and 6 could be applied to these interactions. The difficulty with this is that the substrate analogs found in most crystal structures vary considerably in their similarity to the real substrate. This may make it difficult to define a consistent data set. However, it would be interesting to see how different residues, with different functions, approach substrate groups in different ways.

8.2.3 Release of Products

One aspect of the catalytic cycle that we have not addressed in this thesis, is the release of the product, necessary to free the enzyme ready for the next substrate to bind. In Chapter 4 we see that some enzymes use the binding of fresh substrate to cause the expulsion of product, sometimes with the product moving to a secondary binding site. Whether this is a general mechanism, and whether or not enzymes actively try to expel products, rather than allowing them to simply diffuse away, are largely open questions.

8.3 Prediction of Function From Structure

One of the important questions of recent years has been how to exploit the wealth of information coming from structural genomics projects when performing functional annotation. Part of the difficulty with addressing this problem is the many different layers of functional annotation that can be made, from the molecular level dealt with in this thesis, to the function an enzyme plays within the larger scale networks that make up the metabolism of a cell.

8.3.1 Annotating Enzyme Structures

Prior to predicting function in unannotated structures, it is important that we gather the information known on well studied enzymes into a coherent and consistent framework. The CSA is an ongoing effort to do this, by annotating enzyme structures with their known catalytic residues and the functions of these residues. This information is often well known and published, but is not in a form that is searchable or amenable to large scale analyses. Increasing the size of the CSA database allows us to understand more about enzyme catalysis and therefore, to make better predictions for unannotated enzymes.

8.3.2 Designing New Enzymes

As well as structural genomics providing novel structures, one of the other important developments in recent years is our increasing ability to design novel protein functions, including enzymes. Experiments have been performed to increase the thermostability of enzymes[209], design novel ligand affinities for binding and receptor proteins[210, 211] and even to modify a protein to perform a new enzymatic function[78].

In this later case, an enzyme that already catalyses the ‘novel’ function, in this case TIM, is required to form a template for the new active site. The work presented in this thesis, particularly with respect to the catalytic units, may suggest ways to further develop these partially *de novo* techniques, by suggesting combinations of residues which may work well together. As well as to help develop completely *de novo* techniques which could design enzymes to catalyse previously uncatalysed reactions without requiring a template enzyme.

The reaction catalysed by TIM for instance, requires a carboxylate group acting as an acid/base, and a histidine capable of forming the rare, doubly unprotonated, negatively charged form of the imidazole. The use of another acidic residue close to the carboxylate could help raise it’s pK_a, thus enabling it to act as an acid/base. In contrast, a positively charged residue could help stabilise the negatively charged histidine.

8.3.3 Prediction of Catalytic Residues From Structure

We have found that a combination of sequence conservation and a few basic structural parameters can identify active site locations quite effectively. 69% of the sites in our test set were correctly predicted, and a further 25% partially correct. However, deriving the identity of the catalytic residues within the active site remains a difficult challenge, since many functional, but non-catalytic residues, share the same characteristics as catalytic residues. We would suggest that *ab initio* methods that can predict residues based on features such as unusual pK_a values, and possibly testing for the presence of known catalytic units, in conjunction with conservation approaches, will help to improve this.

One of the most important things in these types of predictions, is knowing where, and how, any ligands bind. If we can predict a cognate ligand for a structure, and successfully dock that ligand into the putative active site, then the number of possible candidate catalytic residues decreases greatly, because we can see which residues are positioned appropriately relative to the ligand. Consensus approaches that combine electrostatics, conservation, ligand matching and docking will be the most likely to prove successful in this area.

8.3.4 Prediction of Catalytic Function From Structure

Once potential catalytic residues are identified in a structure, the next step to improve annotation is to assign chemical functions to each of them. This is not a problem that we have addressed, though ways towards this are suggested by some of our results. Encouragingly, the number of different functions performed by most residues seems to be quite low, and different residues are often quite specialised in the roles they perform. When we also consider the way in which certain residue combinations are often associated with particular functions (histidine-aspartate with acid/base for instance), it seems that making predictions of this nature may be possible.

The final level of prediction once the role of each residue is known, is to predict the overall chemical reaction an enzyme performs, and from there to understand the place the enzyme takes within the metabolic system of a cell. The ability to make predictions such as these is some way in the future, but by improving our basic understanding of how Nature performs catalysis, we have made progress towards this goal.

Publications Arising From This Work:

1. A. Gutteridge, GJ Bartlett, JM Thornton, Using a neural network and spatial clustering to predict the location of active sites in enzymes. J. Mol. Biol. (2003), 330(4), 719-34.

2. A. Gutteridge, JM Thornton, Conformational change in substrate binding, catalysis and product release: An open and shut case? FEBS Letters (2004), 567(1), 67-73.

3. A. Gutteridge, JM Thornton, Conformational changes observed in enzyme crystal structures upon substrate binding. J. Mol. Biol. (2004), 346(1), 21-8.

4. RA George, RV Spriggs, GJ Bartlett, A Gutteridge, MW MacArthur, CT Porter, B Al-Lazikani, JM Thornton, MB Swindells, Effective function annotation through catalytic residue conservation. Proc. Natl. Acad. Sci. USA (2005), Epub ahead of print.

5. A Gutteridge, JM Thornton, Understanding Nature's catalytic toolkit, TiBS, Submitted.

Bibliography

[1] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, M. Hattori, The KEGG resource for deciphering the genome., Nucleic Acids Res 32 (2004) D277–80.

[2] A. Radzicka, R. Wolfenden, A proficient enzyme., Science 267 (1995) 90–3.

[3] T. Bugg, The development of mechanistic enzymology in the 20th century., Nat Prod Rep 18 (2001) 465–93.

[4] Y. Cha, C. Murray, J. Klinman, Hydrogen tunneling in enzyme reactions., Science 243 (1989) 1325–30.

[5] T. Bruice, S. Benkovic, Chemical basis for enzyme catalysis., Biochemistry 39 (2000) 6267–74.

[6] G. Kartha, J. Bello, D. Harker, Tertiary structure of ribonuclease., Nature 213 (1967) 862–5.

[7] D. Findlay, D. Herries, A. Mathias, B. Rabin, C. Ross, The active site and mechanism of action of bovine pancreatic ribonuclease., Nature 190 (1961) 781–84.

[8] R. Wilmouth, K. Edman, R. Neutze, P. Wright, I. Clifton, T. Schneider, C. Schofield, J. Hajdu, X-ray snapshots of serine protease catalysis reveal a tetrahedral intermediate., Nat Struct Biol 8 (2001) 689–94.

[9] J. Judice, T. Gamble, E. Murphy, A. de Vos, P. Schultz, Probing the mechanism of staphylococcal nuclease with unnatural amino acids: kinetic and structural studies., Science 261 (1993) 1578–81.

[10] N. Hooper, Families of zinc metalloproteases., FEBS Lett 354 (1994) 1–6.

[11] A. Eliot, J. Kirsch, Pyridoxal phosphate enzymes: mechanistic, structural, and evolutionary considerations., Annu Rev Biochem 73 (2004) 383–415.

[12] M. Pohl, G. Sprenger, M. Mller, A new perspective on thiamine catalysis., Curr Opin Biotechnol 15 (2004) 335–42.

[13] A. Fersht, Structure and mechanism in protein science., Freeman, 1999.

[14] C. Blake, D. Koenig, G. Mair, A. North, D. Phillips, V. Sarma, Structure of hen egg-white lysozyme. A three-dimensional Fourier synthesis at 2 Angstrom resolution., Nature 206 (1965) 757–61.

[15] L. Johnson, D. Phillips, Structure of some crystalline lysozyme-inhibitor complexes determined by X-ray analysis at 6 Angstrom resolution., Nature 206 (1965) 761–3.

[16] Z. Dauter, V. Lamzin, K. Wilson, The benefits of atomic resolution., Curr Opin Struct Biol 7 (1997) 681–8.

[17] G. Kleywegt, H. Hoier, T. Jones, A re-evaluation of the crystal structure of chloromuconate cycloisomerase., Acta Crystallogr D Biol Crystallogr 52 (1996) 858–63.

[18] G. Kleywegt, K. Henrick, E. Dodson, D. van Aalten, Pound-wise but penny-foolish: How well do micromolecules fare in macromolecular refinement?, Structure (Camb) 11 (2003) 1051–9.

[19] R. Laskowski, M. MacArthur, J. Thornton, Validation of protein models derived from experiment., Curr Opin Struct Biol 8 (1998) 631–9.

[20] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, The Protein Data Bank., Nucleic Acids Res 28 (2000) 235–42.

[21] M. Rossmann, D. Moras, K. Olsen, Chemical and biological evolution of nucleotide-binding protein., Nature 250 (1974) 194–9.

[22] M. Yun, C. Park, J. Kim, H. Park, Structural analysis of glyceraldehyde 3-phosphate dehydrogenase from Escherichia coli: direct evidence of substrate binding and cofactor-induced conformational changes., Biochemistry 39 (2000) 10702–10.

[23] K. Reuter, S. Sanderbrand, H. Jomaa, J. Wiesner, I. Steinbrecher, E. Beck, M. Hintz, G. Klebe, M. Stubbs, Crystal structure of 1-deoxy-D-xylulose-5-phosphate reductoisomerase, a crucial enzyme in the non-mevalonate pathway of isoprenoid biosynthesis., J Biol Chem 277 (2002) 5378–84.

[24] R. A. Laskowski, N. M. Luscombe, M. B. Swindells, J. M. Thornton, Protein clefts in molecular recognition and function., Protein Sci 5 (1996) 2438–52.

[25] A. Monzingo, A. Breksa, S. Ernst, D. Appling, J. Robertus, The X-ray structure of the NAD-dependent 5,10-methylenetetrahydrofolate dehydrogenase from Saccharomyces cerevisiae., Protein Sci 9 (2000) 1374–81.

[26] K. Aktories, Identification of the catalytic site of clostridial ADP-ribosyltransferases., Adv Exp Med Biol 419 (1997) 53–60.

[27] L. Hedstrom, Serine protease mechanism and specificity., Chem Rev 102 (2002) 4501–24.

[28] K. L. Morrison, G. A. Weiss, Combinatorial alanine-scanning., Curr Opin Chem Biol 5 (2001) 302–7.

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

[30] Z. Zhang, Chemical and mechanistic approaches to the study of protein tyrosine phosphatases., Acc Chem Res 36 (2003) 385–92.

[31] W. Albery, J. Knowles, Evolution of enzyme function and the development of catalytic efficiency., Biochemistry 15 (1976) 5631–40.

[32] A. Bairoch, The data bank., Nucleic Acids Res 21 (1993) 3155–6.

[33] N. Nagano, C. T. Porter, J. M. Thornton, The (beta/alpha)(8) glycosidases: sequence and structure analyses suggest distant evolutionary relationships., Protein Eng 14 (2001) 845–55.

[34] A. E. Todd, C. A. Orengo, J. M. Thornton, Evolution of function in protein superfamilies, from a structural perspective., J Mol Biol 307 (2001) 1113–43.

[35] L. Holm, C. Sander, Mapping the protein universe., Science 273 (1996) 595–603.

[36] E. Krissinel, K. Henrick, Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions., Acta Crystallogr D Biol Crystallogr 60 (2004) 2256–68.

[37] F. Pearl, A. Todd, I. Sillitoe, M. Dibley, O. Redfern, T. Lewis, C. Bennett, R. Marsden, A. Grant, D. Lee, A. Akpor, M. Maibaum, A. Harrison, T. Dallman, G. Reeves, I. Diboun, S. Addou, S. Lise, C. Johnston, A. Sillero, J. Thornton, C. Orengo, The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis., Nucleic Acids Res 33 (2005) D247–51.

[38] A. Andreeva, D. Howorth, S. Brenner, T. Hubbard, C. Chothia, A. Murzin, SCOP database in 2004: refinements integrate structure and sequence family data., Nucleic Acids Res 32 (2004) D226–9.

[39] W. Taylor, C. Orengo, Protein structure alignment., J Mol Biol 208 (1989) 1–22.

[40] A. Bairoch, The ENZYME database in 2000., Nucleic Acids Res 28 (2000) 304–5.

[41] A. Matagne, J. Lamotte-Brasseur, J. Frre, Catalytic properties of class A beta-lactamases: efficiency and diversity., Biochem J 330 (Pt 2) (1998) 581–98.

[42] Z. Wang, W. Fast, A. Valentine, S. Benkovic, Metallo-beta-lactamase: structure and mechanism., Curr Opin Chem Biol 3 (1999) 614–22.

[43] B. Matthews, P. Sigler, R. Henderson, D. Blow, Three-dimensional structure of tosyl-alpha-chymotrypsin., Nature 214 (1967) 652–6.

[44] C. Wright, R. Alden, J. Kraut, Structure of subtilisin BPN' at 2.5 angstrm resolution., Nature 221 (1969) 235–42.

[45] A. Kossiakoff, S. Spencer, Direct determination of the protonation states of aspartic acid-102 and histidine-57 in the tetrahedral intermediate of the serine proteases: neutron structure of trypsin., Biochemistry 20 (1981) 6462–74.

[46] W. Cleland, M. Kreevoy, Low-barrier hydrogen bonds and enzymic catalysis., Science 264 (1994) 1887–90.

[47] A. Warshel, A. Papazyan, P. Kollman, On low-barrier hydrogen bonds and enzyme catalysis., Science 269 (1995) 102–6.

[48] S. K. Burley, S. C. Almo, J. B. Bonanno, M. Capel, M. R. Chance, T. Gaasterland, D. Lin, A. Sali, F. W. Studier, S. Swaminathan, Structural genomics: beyond the human genome project., Nat Genet 23 (1999) 151–7.

[49] L. Shapiro, T. Harris, Finding function through structural genomics., Curr Opin Biotechnol 11 (2000) 31–5.

[50] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs., Nucleic Acids Res 25 (1997) 3389–402.

[51] P. D. Karp, What we do not know about sequence analysis and sequence databases., Bioinformatics 14 (1998) 753–4.

[52] D. Devos, A. Valencia, Practical limits of function prediction., Proteins 41 (2000) 98–107.

[53] C. A. Wilson, J. Kreychman, M. Gerstein, Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores., J Mol Biol 297 (2000) 233–49.

[54] N. Hulo, C. Sigrist, V. L. Saux, P. Langendijk-Genevaux, L. Bordoli, A. Gattiker, E. D. Castro, P. Bucher, A. Bairoch, Recent improvements to the PROSITE database., Nucleic Acids Res 32 (2004) D134–7.

[55] A. Bateman, E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S. R. Eddy, S. Griffiths-Jones, K. L. Howe, M. Marshall, E. L. Sonnhammer, The Pfam protein families database., Nucleic Acids Res 30 (2002) 276–80.

[56] A. C. Wallace, N. Borkakoti, J. M. Thornton, TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites., Protein Sci 6 (1997) 2308–23.

[57] J. Barker, J. Thornton, An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis., Bioinformatics 19 (2003) 1644–9.

[58] G. J. Kleywegt, Recognition of spatial motifs in protein structures., J Mol Biol 285 (1999) 1887–97.

[59] J. S. Fetrow, A. Godzik, J. Skolnick, Functional analysis of the Escherichia coli genome using the sequence-to-structure-to-function paradigm: identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxidoreductase activity., J Mol Biol 282 (1998) 703–11.

[60] J. Torrance, G. Bartlett, C. Porter, J. Thornton, Using a library of structural templates to recognise catalytic sites and explore their evolution in homologous families., J Mol Biol 347 (2005) 565–81.

[61] R. Laskowski, J. Watson, J. Thornton, Protein Function Prediction Using Local 3D Templates., J Mol Biol 351 (2005) 614–26.

[62] M. Landau, I. Mayrose, Y. Rosenberg, F. Glaser, E. Martz, T. Pupko, N. Ben-Tal, ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures., Nucleic Acids Res 33 (2005) W299–302.

[63] O. Lichtarge, H. R. Bourne, F. E. Cohen, An evolutionary trace method defines binding surfaces common to protein families., J Mol Biol 257 (1996) 342–58.

[64] S. Madabushi, H. Yao, M. Marsh, D. M. Kristensen, A. Philippi, M. E. Sowa, O. Lichtarge, Structural Clusters of Evolutionary Trace Residues are Statistically Significant and Common in Proteins., J Mol Biol 316 (2002) 139–154.

[65] M. J. Ondrechen, J. G. Clifton, D. Ringe, THEMATICS: a simple computational predictor of enzyme function from structure., Proc Natl Acad Sci U S A 98 (2001) 12473–8.

[66] A. H. Elcock, Prediction of functionally important residues based solely on the computed energetics of protein structure, J Mol Biol 312 (2001) 885–896.

[67] R. Laskowski, J. Watson, J. Thornton, ProFunc: a server for predicting protein function from 3D structure., Nucleic Acids Res 33 (2005) W89–93.

[68] R. Friesner, V. Guallar, Ab initio quantum chemical and mixed quantum mechanics/molecular mechanics (QM/MM) methods for studying enzymatic catalysis., Annu Rev Phys Chem 56 (2005) 389–427.

[69] A. Tousignant, J. Pelletier, Protein motions promote catalysis., Chem Biol 11 (2004) 1037–42.

[70] N. M. Luscombe, R. A. Laskowski, J. M. Thornton, Amino acid-base interactions: a three-dimensional analysis of protein interactions at an atomic level., Nucleic Acids Res 29 (2001) 2860–74.

[71] J. Ippolito, R. Alexander, D. Christianson, Hydrogen bond stereochemistry in protein structure and function., J Mol Biol 215 (1990) 457–71.

[72] J. Singh, J. Thornton, M. Snarey, S. Campbell, The geometries of interacting arginine-carboxyls in proteins., FEBS Lett 224 (1987) 161–71.

[73] J. Mitchell, J. Thornton, J. Singh, S. Price, Towards an understanding of the arginine-aspartate interaction., J Mol Biol 226 (1992) 251–62.

[74] S. Karlin, M. Zuker, L. Brocchieri, Measuring residue associations in protein structures. Possible implications for protein folding., J Mol Biol 239 (1994) 227–48.

[75] C. Porter, G. Bartlett, J. Thornton, The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data., Nucleic Acids Res 32 Database issue (2004) D129–33.

[76] G. J. Bartlett, C. T. Porter, N. Borkakoti, J. M. Thornton, Analysis of Catalytic Residues in Enzyme Active Sites, J Mol Biol 324 (2002) 105–121.

[77] G. Bartlett, N. Borkakoti, J. Thornton, Catalysing new reactions during evolution: economy of residues and mechanism., J Mol Biol 331 (2003) 829–60.

[78] M. Dwyer, L. Looger, H. Hellinga, Computational design of a biologically active enzyme., Science 304 (2004) 1967–71.

[79] C. Burge, S. Karlin, Prediction of complete gene structures in human genomic DNA., J Mol Biol 268 (1997) 78–94.

[80] E. Birney, R. Durbin, Using GeneWise in the Drosophila annotation experiment., Genome Res 10 (2000) 547–8.

[81] X. Zhou, M. D. Toney, pH studies on the mechanism of the pyridoxal phosphate-dependent dialkylglycine decarboxylase., Biochemistry 38 (1999) 311–20.

[82] P. Aloy, E. Querol, F. X. Aviles, M. J. Sternberg, Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking., J Mol Biol 311 (2001) 395–408.

[83] R. Landgraf, I. Xenarios, D. Eisenberg, Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins., J Mol Biol 307 (2001) 1487–502.

[84] T. Pupko, R. E. Bell, I. Mayrose, F. Glaser, N. Ben-Tal, Rate4Site: An algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues., Bioinformatics 18 (2002) S71–S77.

[85] A. Armon, D. Graur, N. Ben-Tal, ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information., J Mol Biol 307 (2001) 447–463.

[86] W. S. Valdar, Scoring residue conservation., Proteins 48 (2002) 227–41.

[87] S. J. Hubbard, J. M. Thornton, NACCESS (1993).

[88] W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features., Biopolymers 22 (1983) 2577–637.

[89] R. A. Laskowski, : a program for visualizing molecular surfaces, cavities, and intermolecular interactions., J Mol Graph 13 (1995) 323–30,.

[90] A. J. Shepherd, D. Gorse, J. M. Thornton, Prediction of the location and type of beta-turns in proteins using neural networks., Protein Sci 8 (1999) 1045–55.

[91] J. Min, X. Zhang, X. Cheng, S. I. Grewal, R. M. Xu, Structure of the domain histone lysine methyltransferase Clr4., Nat Struct Biol 9 (2002) 828–32.

[92] S. A. Jacobs, J. M. Harp, S. Devarakonda, Y. Kim, F. Rastinejad, S. Khorasanizadeh, The active site of the domain is constructed on a knot., Nat Struct Biol 9 (2002) 833–8.

[93] J. Wilson, C. Jing, P. Walker, S. Martin, S. Howell, G. Blackburn, S. Gamblin, B. Xiao, Crystal Structure and Functional Analysis of the Histone Methyltransferase SET7/9., Cell 111.

[94] X. Zhang, H. Tamaru, S. Khan, J. Horton, L. Keefe, E. Selker, X. Cheng, Structure of the Neurospora Domain Protein-5, a Histone H3 Lysine Methyltransferase., Cell 111.

[95] R. Trievel, B. Beach, L. Dirk, R. Houtz, J. Hurley, Structure and Catalytic Mechanism of a Domain Protein Methyltransferase., Cell 111.

[96] B. Xiao, C. Jing, J. R. Wilson, P. A. Walker, N. Vasisht, G. Kelly, S. Howell, I. A. Taylor, G. M. Blackburn, S. J. Gamblin, Structure and catalytic mechanism of the human histone methyltransferase SET7/9, Nature 42 (2003) 652–656.

[97] S. Rea, F. E. O., D. 'Carroll, B. D. Strahl, Z. W. Sun, M. Schmid, S. Opravil, K. Mechtler, C. P. Ponting, C. D. Allis, T. Jenuwein, Regulation of chromatin

structure by site-specific histone H3 methyltransferases., Nature 406 (2000) 593–9.

[98] J. Nakayama, J. C. Rice, B. D. Strahl, C. D. Allis, S. I. Grewal, Role of histone H3 lysine 9 methylation in epigenetic control of heterochromatin assembly., Science 292 (2001) 110–3.

[99] V. P. Roey, L. Meehan, J. C. Kowalski, M. Belfort, V. Derbyshire, Catalytic domain structure and hypothesis for function of intron endonuclease-TevI., Nat Struct Biol 9 (2002) 806–11.

[100] V. Derbyshire, J. C. Kowalski, J. T. Dansereau, C. R. Hauer, M. Belfort, Two-domain structure of the td intron-encoded endonuclease-TevI correlates with the two-domain configuration of the homing site., J Mol Biol 265 (1997) 494–506.

[101] D. Nurizzo, J. P. Turkenburg, S. J. Charnock, S. M. Roberts, E. J. Dodson, V. A. McKie, E. J. Taylor, H. J. Gilbert, G. J. Davies, Cellvibrio japonicus alpha-arabinanase 43A has a novel five-blade beta-propeller fold., Nat Struct Biol 9 (2002) 665–8.

[102] G. Davies, M. L. Sinnott, S. G. Withers, Comprehensive Biological Catalysis, Vol. 1, 1997.

[103] B. Berger-Bachi, L. Barberis-Maino, A. Strassle, F. H. Kayser, FemA, a host-mediated factor essential for methicillin resistance in Staphylococcus aureus: molecular cloning and characterization., Mol Gen Genet 219(1-2) (1989) 263–9.

[104] B. Berger-Bachi, M. Tschierske, Role of Fem factors in methicillin resistance, Drug Resist. Updat. 1 (1998) 325–335.

[105] T. Benson, D. Prince, V. Mutchler, K. Curry, A. Ho, R. Sarver, J. Hagadorn, G. Choi, R. Garlick, -Ray Crystal Structure of Staphylococcus aureus FemA., Structure (Camb) 10.

[106] J. M. Lovgren, P. M. Wikstrom, The rlmB gene is essential for formation of Gm2251 in 23S rRNA but not for ribosome maturation in Escherichia coli., J Bacteriol 183 (2001) 6957–60.

[107] G. Michel, V. Sauve, R. Larocque, Y. Li, A. Matte, M. Cygler, The structure of the RlmB 23S rRNA methyltransferase reveals a new methyltransferase fold with a unique knot., Structure (Camb) 10 (2002) 1303–15.

[108] C. Gustafsson, R. Reid, P. J. Greene, D. V. Santi, Identification of new modifying enzymes by iterative genome search using known modifying enzymes as probes., Nucleic Acids Res 24 (1996) 3756–62.

[109] B. C. Persson, G. Jager, C. Gustafsson, The spoU gene of Escherichia coli, the fourth gene of the spoT operon, is essential for tRNA (Gm18) 2'-methyltransferase activity., Nucleic Acids Res 25 (1997) 4093–7.

[110] M. J. Zvelebil, M. J. Sternberg, Analysis and prediction of the location of catalytic residues in enzymes., Protein Eng 2 (1988) 127–38.

[111] R. Chelli, F. Gervasio, P. Procacci, V. Schettino, Inter-residue and solvent-residue interactions in proteins: a statistical study on experimental structures., Proteins 55 (2004) 139–51.

[112] A. Bairoch, R. Apweiler, C. Wu, W. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. Martin, D. Natale, C. O'Donovan, N. Redaschi, L. Yeh, The Universal Protein Resource (UniProt)., Nucleic Acids Res 33 (2005) D154–9.

[113] D. Koshland, Correlation of structure and function in enzyme action., Science 142 (1963) 1533–41.

[114] G. Hammes, Multiple conformational changes in enzyme catalysis., Biochemistry 41 (2002) 8221–8.

[115] N. Echols, D. Milburn, M. Gerstein, MolMovDB: analysis and visualization of conformational change and structural flexibility., Nucleic Acids Res 31 (2003) 478–82.

[116] M. Gerstein, W. Krebs, A database of macromolecular motions., Nucleic Acids Res 26 (1998) 4280–90.

[117] D. Joseph, G. Petsko, M. Karplus, Anatomy of a conformational change: hinged ”lid” motion of the triosephosphate isomerase loop., Science 249 (1990) 1425–8.

[118] C. Anderson, F. Zucker, T. Steitz, Space-filling models of kinase clefts and conformation changes., Science 204 (1979) 375–80.

[119] C. McPhalen, M. Vincent, D. Picot, J. Jansonius, A. Lesk, C. Chothia, Domain closure in mitochondrial aspartate aminotransferase., J Mol Biol 227 (1992) 197–213.

[120] M. Sawaya, J. Kraut, Loop and subdomain movements in the mechanism of Escherichia coli dihydrofolate reductase: crystallographic evidence., Biochemistry 36 (1997) 586–603.

[121] J. Taylor, F. Takusagawa, G. Markham, The active site loop of S-adenosylmethionine synthetase modulates catalytic efficiency., Biochemistry 41 (2002) 9358–69.

[122] B. Shoichet, S. McGovern, B. Wei, J. Irwin, Lead discovery using molecular docking., Curr Opin Chem Biol 6 (2002) 439–46.

[123] S. Teague, Implications of protein flexibility for drug discovery., Nat Rev Drug Discov 2 (2003) 527–41.

[124] A. C. Wallace, R. A. Laskowski, J. M. Thornton, Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases., Protein Sci 5 (1996) 1001–13.

[125] T. Skarzynski, A. Mistry, A. Wonacott, S. Hutchinson, V. Kelly, K. Duncan, Structure of UDP-N-acetylglucosamine enolpyruvyl transferase, an enzyme essential for the synthesis of bacterial peptidoglycan, complexed with substrate UDP-N-acetylglucosamine and the drug fosfomycin., Structure 4 (1996) 1465–74.

[126] S. Eschenburg, E. Schnbrunn, Comparative X-ray analysis of the un-liganded fosfomycin-target murA., Proteins 40 (2000) 290–8.

[127] A. C. R. Martin, ProFit - http://www.bioinf.org.uk/software/profit/ (1996).

[128] A. Karlsson, J. Parales, R. Parales, D. Gibson, H. Eklund, S. Ramaswamy, Crystal structure of naphthalene dioxygenase: side-on binding of dioxygen to iron., Science 299 (2003) 1039–42.

[129] A. Albert, L. Yenush, M. Gil-Mascarell, P. Rodriguez, S. Patel, M. Martnez-Ripoll, T. Blundell, R. Serrano, X-ray structure of yeast Hal2p, a major target of lithium and sodium toxicity, and identification of framework interactions determining cation sensitivity., J Mol Biol 295 (2000) 927–38.

[130] S. Patel, M. Martnez-Ripoll, T. Blundell, A. Albert, Structural enzymology of Li(+)-sensitive/Mg(2+)-dependent phosphatases., J Mol Biol 320 (2002) 1087–94.

[131] H. Lauble, C. Stout, Steric and conformational features of the aconitase mechanism., Proteins 22 (1995) 1–11.

[132] S. Lloyd, H. Lauble, G. Prasad, C. Stout, The mechanism of aconitase: 1.8 A resolution crystal structure of the S642a:citrate complex., Protein Sci 8 (1999) 2655–62.

[133] J. Ferrer, J. Jez, M. Bowman, R. Dixon, J. Noel, Structure of chalcone synthase and the molecular basis of plant polyketide biosynthesis., Nat Struct Biol 6 (1999) 775–84.

[134] A. Becker, I. Schlichting, W. Kabsch, D. Groche, S. Schultz, A. Wagner, Iron center, substrate recognition and mechanism of peptide deformylase., Nat Struct Biol 5 (1998) 1053–8.

[135] J. Guilloteau, M. Mathieu, C. Giglione, V. Blanc, A. Dupuy, M. Chevrier, P. Gil, A. Famechon, T. Meinnel, V. Mikol, The crystal structures of four peptide deformylases bound to the antibiotic actinonin reveal two distinct types: a platform for the structure-based design of antibacterial agents., J Mol Biol 320 (2002) 951–62.

[136] I. Schlichting, J. Berendzen, K. Chu, A. Stock, S. Maves, D. Benson, R. Sweet, D. Ringe, G. Petsko, S. Sligar, The catalytic pathway of cytochrome p450cam at atomic resolution., Science 287 (2000) 1615–22.

[137] S. Long, P. Casey, L. Beese, Reaction path of protein farnesyltransferase at atomic resolution., Nature 419 (2002) 645–50.

[138] T. Stout, C. Sage, R. Stroud, The additivity of substrate fragments in enzyme-ligand binding., Structure 6 (1998) 839–48.

[139] C. Sage, M. Michelitsch, T. Stout, D. Biermann, R. Nissen, J. Finer-Moore, R. Stroud, D221 in thymidylate synthase controls conformation change, and thereby opening of the imidazolidine., Biochemistry 37 (1998) 13893–901.

[140] W. Montfort, K. Perry, E. Fauman, J. Finer-Moore, G. Maley, L. Hardy, F. Maley, R. Stroud, Structure, multiple site binding, and segmental accommodation in thymidylate synthase on binding dUMP and an anti-folate., Biochemistry 29 (1990) 6964–77.

[141] E. Fauman, E. Rutenber, G. Maley, F. Maley, R. Stroud, Water-mediated substrate/product discrimination: the product complex of thymidylate synthase at 1.83 A., Biochemistry 33 (1994) 1502–11.

[142] R. Stroud, J. Finer-Moore, Conformational dynamics along an enzymatic reaction pathway: thymidylate synthase, ”the movie”., Biochemistry 42 (2003) 239–47.

[143] T. Sandalova, G. Schneider, H. Kck, Y. Lindqvist, Structure of dethiobiotin synthetase at 0.97 A resolution., Acta Crystallogr D Biol Crystallogr 55 (Pt 3) (1999) 610–24.

[144] W. Huang, J. Jia, K. Gibson, W. Taylor, A. Rendina, G. Schneider, Y. Lindqvist, Mechanism of an ATP-dependent carboxylase, dethiobiotin synthetase, based on crystallographic studies of complexes with substrates and a reaction intermediate., Biochemistry 34 (1995) 10985–95.

[145] F. Mancia, P. Evans, Conformational changes on substrate binding to methylmalonyl CoA mutase and new insights into the free radical mechanism., Structure 6 (1998) 711–20.

[146] F. Mancia, G. Smith, P. Evans, Crystal structure of substrate complexes of methylmalonyl-CoA mutase., Biochemistry 38 (1999) 7999–8005.

[147] D. Koshland, Conformational changes: how small is big enough?, Nat Med 4 (1998) 1112–4.

[148] A. Mesecar, B. Stoddard, D. Koshland, Orbital steering in the catalytic power of enzymes: small structural changes with large catalytic consequences., Science 277 (1997) 202–6.

[149] S. Burley, G. Petsko, Aromatic-aromatic interaction: a mechanism of protein structure stabilization., Science 229 (1985) 23–8.

[150] E. Baker, R. Hubbard, Hydrogen bonding in globular proteins., Prog Biophys Mol Biol 44 (1984) 97–179.

[151] W. Wedemeyer, E. Welker, M. Narayan, H. Scheraga, Disulfide bonds and protein folding., Biochemistry 39 (2000) 4207–16.

[152] L. Sawyer, M. James, Carboxyl-carboxylate interactions in proteins., Nature 295 (1982) 79–80.

[153] M. Flocco, S. Mowbray, Strange bedfellows: interactions between acidic sidechains in proteins., J Mol Biol 254 (1995) 96–105.

[154] D. Schmidt, F. Westheimer, PK of the lysine amino group at the active site of acetoacetate decarboxylase., Biochemistry 10 (1971) 1249–53.

[155] F. Hollfelder, A. Kirby, D. Tawfik, On the magnitude and specificity of medium effects in enzyme-like catalysts for proton transfer., J Org Chem 66 (2001) 5866–74.

[156] M. Berry, G. Phillips, Crystal structures of Bacillus stearothermophilus adenylate kinase with bound Ap5A, Mg2+ Ap5A, and Mn2+ Ap5A reveal an intermediate lid position and six coordinate octahedral geometry for bound Mg2+ and Mn2+., Proteins 32 (1998) 276–88.

[157] C. Frazao, I. Bento, J. Costa, C. Soares, P. Verissimo, C. Faro, E. Pires, J. Cooper, M. Carrondo, Crystal structure of cardosin A, a glycosylated and Arg-Gly-Asp-containing aspartic proteinase from the flowers of Cynara cardunculus L., J Biol Chem 274 (1999) 27694–701.

[158] A. Varrot, G. Davies, Direct experimental observation of the hydrogen-bonding network of a glycosidase along its reaction coordinate revealed by atomic resolution analyses of endoglucanase Cel5A., Acta Crystallogr D Biol Crystallogr 59 (2003) 447–52.

[159] D. Sproge, L. van den Broek, O. Mirza, J. Kastrup, A. Voragen, M. Gajhede, L. Skov, Crystal structure of sucrose phosphorylase from Bifidobacterium adolescentis., Biochemistry 43 (2004) 1156–62.

[160] J. Lubkowski, M. Dauter, K. Aghaiypour, A. Wlodawer, Z. Dauter, Atomic resolution structure of Erwinia chrysanthemi L-asparaginase., Acta Crystallogr D Biol Crystallogr 59 (2003) 84–92.

[161] A. Ehrensberger, D. Wilson, Structural and catalytic diversity in the two family 11 aldo-keto reductases., J Mol Biol 337 (2004) 661–73.

[162] M. Hennig, B. Darimont, J. Jansonius, K. Kirschner, The catalytic mechanism of indole-3-glycerol phosphate synthase: crystal structures of complexes of the enzyme from Sulfolobus solfataricus with substrate analogue, substrate, and product., J Mol Biol 319 (2002) 757–66.

[163] A. Teplyakov, G. Obmolova, M. Badet-Denisot, B. Badet, I. Polikarpov, Involvement of the C terminus in intramolecular nitrogen channeling in glucosamine 6-phosphate synthase: evidence from a 1.6 A crystal structure of the isomerase domain., Structure 6 (1998) 1047–55.

[164] A. Teplyakov, G. Obmolova, M. Badet-Denisot, B. Badet, The mechanism of sugar phosphate isomerization by glucosamine 6-phosphate synthase., Protein Sci 8 (1999) 596–602.

[165] H. Song, S. Suh, Kunitz-type soybean trypsin inhibitor revisited: refined structure of its complex with porcine trypsin reveals an insight into the interaction between a homologous inhibitor from Erythrina caffra and tissue-type plasminogen activator., J Mol Biol 275 (1998) 347–63.

[166] D. Liao, G. Kapadia, P. Reddy, M. Saier, J. Reizer, O. Herzberg, Structure of the IIA domain of the glucose permease of Bacillus subtilis at 2.2-A resolution., Biochemistry 30 (1991) 9583–94.

[167] A. Shaw, R. Bott, C. Vonrhein, G. Bricogne, S. Power, A. G. Day, A novel combination of two classic catalytic schemes., J Mol Biol 320 (2002) 303–9.

[168] J. Knowles, Enzyme catalysis: not different, just better., Nature 350 (1991) 121–4.

[169] D. Walters, A. Allerhand, Tautomeric states of the histidine residues of bovine pancreatic ribonuclease A. Application of carbon 13 nuclear magnetic resonance spectroscopy., J Biol Chem 255 (1980) 6200–4.

[170] G. Stranzl, K. Gruber, G. Steinkellner, K. Zangger, H. Schwab, C. Kratky, Observation of a short, strong hydrogen bond in the active site of hydroxynitrile lyase from Hevea brasiliensis explains a large pKa shift of the catalytic base induced by the reaction intermediate., J Biol Chem 279 (2004) 3699–707.

[171] C. Goward, D. Nicholls, Malate dehydrogenase: a model for structure, evolution, and catalysis., Protein Sci 3 (1994) 1883–8.

[172] T. Steiner, G. Koellner, Hydrogen bonds with pi-acceptors in proteins: frequencies and role in stabilizing local 3D structures., J Mol Biol 305 (2001) 535–57.

[173] J. B. Mitchell, R. A. Laskowski, J. M. Thornton, Non-randomness in side-chain packing: the distribution of interplanar angles., Proteins 29 (1997) 370–80.

[174] L. Brocchieri, S. Karlin, Geometry of interplanar residue contacts in protein structures., Proc Natl Acad Sci U S A 91 (1994) 9297–301.

[175] R. Bhattacharyya, R. Saha, U. Samanta, P. Chakrabarti, Geometry of interaction of the histidine ring with other planar and basic residues., J Proteome Res 2 (2003) 255–63.

[176] M. Tanokura, 1H-NMR study on the tautomerism of the imidazole ring of histidine residues. I. Microscopic pK values and molar ratios of tautomers in histidine-containing peptides., Biochim Biophys Acta 742 (1983) 576–85.

[177] T. Harris, G. Turner, Structural basis of perturbed pKa values of catalytic groups in enzyme active sites., IUBMB Life 53 (2002) 85–98.

[178] J. Ko, L. Murga, P. Andr, H. Yang, M. Ondrechen, R. Williams, A. Agunwamba, D. Budil, Statistical criteria for the identification of protein active sites using Theoretical Microscopic Titration Curves., Proteins 59 (2005) 183–95.

[179] P. Chakrabarti, Geometry of interaction of metal ions with histidine residues in protein structures., Protein Eng 4 (1990) 57–63.

[180] I. Alberts, K. Nadassy, S. Wodak, Analysis of zinc binding sites in protein crystal structures., Protein Sci 7 (1998) 1700–16.

[181] A. West, E. Martinez-Hackert, A. Stock, Crystal structure of the catalytic domain of the chemotaxis receptor methylesterase, CheB., J Mol Biol 250 (1995) 276–90.

[182] D. Williams, E. Stephens, D. O'Brien, M. Zhou, Understanding noncovalent interactions: ligand binding energy and catalytic efficiency from ligand-induced reductions in motion within receptors and enzymes., Angew Chem Int Ed Engl 43 (2004) 6596–616.

[183] D. Williams, E. Stephens, M. Zhou, Ligand binding energy and catalytic efficiency from improved packing within receptors and enzymes., J Mol Biol 329 (2003) 389–99.

[184] S. Velankar, P. McNeil, V. Mittard-Runte, A. Suarez, D. Barrell, R. Apweiler, K. Henrick, E-MSD: an integrated data resource for bioinformatics., Nucleic Acids Res 33 Database Issue (2005) D262–5.

[185] R. Hooft, G. Vriend, C. Sander, E. Abola, Errors in protein structures., Nature 381.

[186] I. McDonald, J. Thornton, Satisfying hydrogen bonding potential in proteins., J Mol Biol 238 (1994) 777–93.

[187] T. Fenn, D. Ringe, G. Petsko, Xylose isomerase in substrate and inhibitor michaelis states: atomic resolution studies of a metal-mediated hydride shift., Biochemistry 43 (2004) 6464–74.

[188] T. Shimizu, T. Nakatsu, K. Miyairi, T. Okuno, H. Kato, Active-site architecture of endopolygalacturonase I from Stereum purpureum revealed by crystal structures in native and ligand-bound forms at atomic resolution., Biochemistry 41 (2002) 6651–9.

[189] P. Pereira, S. Macedo-Ribeiro, A. Prraga, R. Prez-Luque, O. Cunningham, K. Darcy, T. Mantle, M. Coll, Structure of human biliverdin IXbeta reductase, an early fetal bilirubin IXbeta producing enzyme., Nat Struct Biol 8 (2001) 215–20.

[190] V. Samygina, A. Popov, E. Rodina, N. Vorobyeva, V. Lamzin, K. Polyakov, S. Kurilova, T. Nazarova, S. Avaeva, The structures of Escherichia coli inorganic pyrophosphatase complexed with Ca(2+) or CaPP(i) at atomic resolution and their mechanistic implications., J Mol Biol 314 (2001) 633–45.

[191] C. Betzel, S. Gourinath, P. Kumar, P. Kaur, M. Perbandt, S. Eschenburg, T. Singh, Structure of a serine protease proteinase K from Tritirachium album limber at 0.98 A resolution., Biochemistry 40 (2001) 3080–8.

[192] M. Wrtele, M. Hahn, K. Hilpert, W. Hhne, Atomic resolution structure of native porcine pancreatic elastase at 1.1 A., Acta Crystallogr D Biol Crystallogr 56 (Pt 4) (2000) 520–3.

[193] G. Katona, R. Wilmouth, P. Wright, G. Berglund, J. Hajdu, R. Neutze, C. Schofield, X-ray structure of a serine protease acyl-enzyme complex at 0.95-A resolution., J Biol Chem 277 (2002) 21962–70.

[194] A. Heine, G. DeSantis, J. Luz, M. Mitchell, C. Wong, I. Wilson, Observation of covalent intermediates in an enzyme mechanism at atomic resolution., Science 294 (2001) 369–74.

[195] K. Harata, R. Kanai, Crystallographic dissection of the thermal motion of protein-sugar complex., Proteins 48 (2002) 53–62.

[196] D. Gurin, M. Lascombe, M. Costabel, H. Souchon, V. Lamzin, P. Bguin, P. Alzari, Atomic (0.94 A) resolution structure of an inverting glycosidase in complex with substrate., J Mol Biol 316 (2002) 1061–9.

[197] S. Korolev, Y. Ikeguchi, T. Skarina, S. Beasley, C. Arrowsmith, A. Edwards, A. Joachimiak, A. Pegg, A. Savchenko, The crystal structure of spermidine synthase with a multisubstrate adduct inhibitor., Nat Struct Biol 9 (2002) 27–31.

[198] G. Davies, L. Mackenzie, A. Varrot, M. Dauter, A. Brzozowski, M. Schlein, S. Withers, Snapshots along an enzymatic reaction coordinate: analysis of a retaining beta-glycoside hydrolase., Biochemistry 37 (1998) 11707–13.

[199] G. Sidhu, S. Withers, N. Nguyen, L. McIntosh, L. Ziser, G. Brayer, Sugar ring distortion in the glycosyl-enzyme intermediate of a family G/11 xylanase., Biochemistry 38 (1999) 5346–54.

[200] D. Rees, M. Lewis, W. Lipscomb, Refined crystal structure of carboxypeptidase A at 1.54 A resolution., J Mol Biol 168 (1983) 367–87.

[201] H. Kim, W. Lipscomb, Comparison of the structures of three carboxypeptidase A-phosphonate complexes determined by X-ray crystallography., Biochemistry 30 (1991) 8171–80.

[202] I. Kuntz, K. Chen, K. Sharp, P. Kollman, The maximal affinity of ligands., Proc Natl Acad Sci U S A 96 (1999) 9997–10002.

[203] D. Hyre, I. L. Trong, S. Freitag, R. Stenkamp, P. Stayton, Ser45 plays an important role in managing both the equilibrium and transition state energetics of the streptavidin-biotin system., Protein Sci 9 (2000) 878–85.

[204] S. Freitag, I. L. Trong, L. Klumb, P. Stayton, R. Stenkamp, Structural studies of the streptavidin binding loop., Protein Sci 6 (1997) 1157–66.

[205] J. Kavanaugh, W. Moo-Penn, A. Arnone, Accommodation of insertions in helices: the mutation in hemoglobin Catonsville (Pro 37 alpha-Glu-Thr 38 alpha) generates a 3(10)–¿alpha bulge., Biochemistry 32 (1993) 2509–13.

[206] J. Kavanaugh, P. Rogers, A. Arnone, Crystallographic evidence for a new ensemble of ligand-induced allosteric transitions in hemoglobin: the T-to-T(high) quaternary transitions., Biochemistry 44 (2005) 6101–21.

[207] K. Acharya, M. Lloyd, The advantages and limitations of protein crystal structures., Trends Pharmacol Sci 26 (2005) 10–4.

[208] J. Badger, J. Hendle, Reliable quality-control methods for protein crystal structures., Acta Crystallogr D Biol Crystallogr 58 (2002) 284–91.

[209] A. Korkegian, M. Black, D. Baker, B. Stoddard, Computational thermostabilization of an enzyme., Science 308 (2005) 857–60.

[210] L. Looger, M. Dwyer, J. Smith, H. Hellinga, Computational design of receptor and sensor proteins with novel functions., Nature 423 (2003) 185–90.

[211] W. Yang, A. Wilkins, Y. Ye, Z. Liu, S. Li, J. Urbauer, H. Hellinga, A. Kearney, P. van der Merwe, J. Yang, Design of a calcium-binding protein with desired structure in a cell adhesion molecule., J Am Chem Soc 127 (2005) 2085–93.