The geometry and evolution of catalytic sites and metal binding sites.

James William Torrance

Robinson College, Cambridge

European Bioinformatics Institute

This dissertation is submitted for the degree of Doctor of Philosophy.

March 2008

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 limit specified by the Graduate School of Biological, Medical and Veterinary Sciences.

Abstract

Analysing the geometry of functional sites in proteins can shed light on the evolution of these functional sites, explore the relationship between active site geometry and chemistry, and work towards methods for predicting protein function from structure. This thesis describes the analysis of manually annotated datasets of catalytic residues, biologically relevant metal binding sites, and catalytic mechanisms.

The principal source of data for this thesis was the Catalytic Site Atlas, a database of catalytic sites in proteins of known structure. The author has supervised an expansion of the coverage and level of detail of this database. The expanded database has been analysed to discover trends in the catalytic roles played by residues and cofactors.

A comparison of the structures of catalytic sites in homologous enzymes showed that these mostly differ by less than 1 Å root mean square deviation, even when the sequence similarity between the proteins is low. As a consequence of this structural conservation, structural templates representing catalytic sites have the potential to succeed at function prediction in cases where methods based on sequence or overall structure fail. Templates were found to discriminate between matches to related proteins and random matches with over 85% sensitivity and predictive accuracy. Templates based on protein backbone positions were found to be more discriminating than those based on sidechain atoms.

This approach to analysing structural variation can also be applied to other functional sites in proteins, such as metal binding sites. An analysis of a set of of well-documented structural calcium and zinc binding sites found that, like catalytic sites, these are highly conserved between distant relatives. Structural templates representing these conserved calcium and zinc binding sites were used to search the Protein Data Bank for cases where unrelated proteins have converged upon the same residue selection and geometry for metal

binding. This allowed the identification of "archetypal" metal binding sites, which had independently evolved on a number of occasions. Relatives of these metal binding proteins sometimes do not bind metal. For most of the calcium binding sites studied, the lack of metal binding in relatives was due to point mutation of the metal-binding residues, whilst for zinc binding sites, lack of metal binding in relatives always involved more extensive changes.

As a complement to the analysis of overall structural variation in catalytic sites described above, statistics were gathered describing the typical distances and angles of individual catalytic residues with regard to the substrate and one another. The geometry of residues whose function involves the transfer or sharing of hydrogens was found to closely resemble the geometry of non-catalytic hydrogen bonds.

Acknowledgements

First of all, thanks are due to my supervisor Janet Thornton. She has kept me focused on the big picture, the positive side, and the schedule. Without her advice and encouragement, this thesis would have been a big pile of blank paper sitting inside a printer. I also thank my co-supervisor in the Chemistry Department, John Mitchell, who was always happy to have as much or as little involvement as was necessary at different stages, and who supplied an important chemical perspective.

Many people have passed through the Thornton group over the last four years, and all of them have provided some combination of technical advice and/or moral support; those who are named here are just the first among those many. Craig Porter, Gail Bartlett and Alex Gutteridge introduced me to the Catalytic Site Atlas. Jonathan Barker provided assistance with his template matching program Jess, as well as amusement through his mechanical ingenuity and artistic talents. Malcolm MacArthur furnished me with his dataset of metal binding sites, and acted as a patient guide to the world of metalloproteins. Gemma Holliday explained the workings of the MACiE database, and endured my questions on chemical topics. All of the above, along with Gabby Reeves, James Watson and Roman Laskowski, kindly gave up their time to proofread portions of this thesis. I'm also grateful to the various summer students who suffered under my tutelage.

I am indebted to other members of the group for non-academic reasons. My various office-mates down the years tolerated my nervous tics, muttering, and attempts to fill the room with houseplants. Matthew Bashton recklessly offered me a room in his flat despite having previously been my office-mate. Gabby and James dragged me to Steve "drill sergeant" Russen's circuit training sessions, resulting in the Arnold-Schwarzenegger-like physique that I rejoice in today. Tim Massingham bravely protected me from the goths

down at the Kambar.

Light relief was provided by an assortment of friends, acquaintances, cronies, hangers-on and ne'er-do-wells, including the EBI PhD student mob, the Robinson PhD student mob, the old York University gang from antediluvian times, and the even older Robinson gang from the times before that. Depeche Mode, Iron Maiden and the Sisters of Mercy permitted me to trade away some of my hearing in order to retain some of my sanity. These bands are fairly unlikely to read this thesis, but I suppose it could help them pass the time on a tour bus.

Finally, I'd like to thank my parents. Not only have they been a consistent source of entertainment, education, and spirited yet amicable political debate, they have also borne the brunt of my whingeing with extraordinary patience.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The role and importance of enzymes

Life would be impossible without catalysts to accelerate the rates of specific chemical reactions. Enzymes fulfil this role, and they are capable of enormous rate enhancement and specificity. Orotidine 5′-monophosphate spontaneously undergoes decarboxylation (Figure 1.1) with a half-life of 78 million years; at the active site of orotidine 5′-phosphate decarboxylase, the same reaction occurs with a half-life of 18 milliseconds (Miller *et al.*, 2000). Whilst the 10^{17}-fold rate enhancement achieved by orotidine 5′-phosphate decarboxylase is the greatest known, enzymes routinely achieve rate enhancements of many orders of magnitude. Enzymes are also able to discriminate between highly similar substrates, including the ability to distinguish between enantiomers.

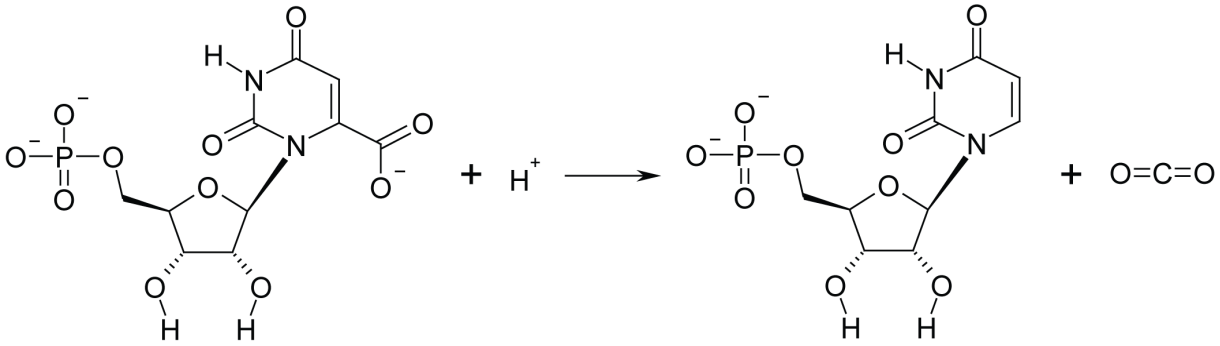


Figure 1.1: Decarboxylation of orotidine 5′-phosphate.

The existence of specific substances that catalyse biological reactions was discovered in the late 19th century. In 1897 Eduard Buchner demonstrated that cell-free extracts of yeast could carry out fermentation; in 1894 Emil Fischer proposed the "lock and key" hypothesis to explain how enzymes interact with their substrates (Fischer, 1894). Research over the course of more than a century since then has developed our knowledge about how enzymes operate at the chemical level (Buchner, 1894). Technological developments over the last couple of decades have greatly multiplied the number of enzymes whose chemical mechanism is understood— especially the routine use of X-ray crystallography to determine the three-dimensional structure of enzymes and the routine use of site-directed mutagenesis to dissect the contributions made by individual amino acid residues.

Despite this progress, there remain controversies and uncertainties concerning how enzymes operate. These include such fundamental questions as which different chemical aspects of enzymatic catalysis make the greatest quantitative contribution to the rate enhancement achieved by enzymes (Bugg, 2001; Kraut *et al.*, 2003).

A better understanding of how enzymes evolve and function at the molecular level has the intrinsic benefit of providing an insight into the most fundamental workings of living things. It also has a number of pragmatic uses. Many existing drugs are enzyme inhibitors, from natural products such as penicillin (Spratt, 1975) to the antiretroviral drugs that inhibit HIV protease (Flexner, 1998) and reverse transcriptase (Esnouf *et al.*, 1995). Rational design of enzyme inhibitors can produce more effective drugs. Enzymes are also employed in a range of industrial applications (Kirk *et al.*, 2002). These include the use of proteases in cleaning applications, the use of various hydrolases in the food industry, and the use of various enzymes in synthetic chemistry where their substrate specificity (particularly stereospecificity) is useful. Enzyme engineering can be useful in improving the specificity and robustness of these enzymes. In both drug design and protein engineering, rational design approaches contend with random screening methods (Tao & Cornish, 2002). However, rational and random approaches are not mutually exclusive, and a greater understanding of enzyme function can hope to aid both drug design and enzyme engineering.

1.1.1 Classifying enzymes

A classification system for enzyme activities facilitates comparisons of differences in function between homologous enzymes and similarities in function in non-homologous enzymes. It also aids computational analyses of enzyme function, and simplifies the transfer of functional annotation between homologous proteins.

By far the most commonly used classification of enzyme activities is the Enzyme Commission (EC) classification created by the International Union of Pure and Applied Chemistry (IUPAC) (Webb, 1992). This is a numerical, hierarchical classification with four levels. It divides all enzymes into six numbered classes, described in Table 1.1. Each of these classes is further broken down into subclasses, or "second-level EC numbers". The number and meaning of these second-level classifications is different for each first-level classification; this is also true of the further subdivisions of the EC classification. Each second-level classification is broken down into third-level classifications, and each of these third-level classification is subdivided into fourth-level classifications. The third level of the classification often specifies the overall chemical change carried out by the enzyme, whilst the fourth level generally specifies the precise substrate (which can sometimes be a class of compounds, such as DNA or peptides).

Table 1.1: Classes of the Enzyme Commission (EC) classification.

First EC number	Function
1	Oxidoreductases
2	Transferases
3	Hydrolases
4	Lyases
5	Isomerases
6	Ligases

Each level of the classification is expressed as a number, and a complete classification gives these numbers separated by dots. For example, β-lactamases (reaction shown in Figure 1.2) have the EC number 3.5.2.6. The first-level classification is 3, which signifies that this is a hydrolase. The second-level classification, 3.5, means that this enzyme acts to cleave a carbon-nitrogen bond that is not a peptide bond. The third-level classification,

3.5.2, signifies that this carbon-nitrogen bond is in a cyclic amide. The fourth-level classification, 3.5.2.6, identifies the substrate as belonging to the β-lactam class.

The EC classification is concerned with enzyme activities rather than individual enzymes. It generally only describes the substrates and products of the reaction. Enzymes which convert the same substrates to the same products using entirely different reaction mechanisms will normally have the same EC classification. Furthermore, unrelated enzymes which convert the same substrates to the same products have the same EC classification. For example, the β-lactamase classification referred to in the previous paragraph applies to a range of enzymes including several groups that are not homologous to one another and have entirely different mechanisms.

An alternative classification of enzymes called RLCP is used by the EzCatDb database (Nagano, 2005). Unlike the EC, this classification classifies enzymes according to reaction mechanism and the residues employed by the enzyme. Like the EC, it has four digits; these correspond to basic reaction (R), ligand group involved in catalysis (L), catalytic mechanism (C), and residues/cofactors located on Proteins (P). This classification scheme does not currently cover all enzymes.

There are also several classification schemes developed for particular groups of enzymes. These include the classification of enzymes acting on glycosidic bonds which is associated with the CAZy database (Henrissat & Davies, 1997) and the classification of eukaryotic protein kinases developed by Hanks & Hunter (1995).

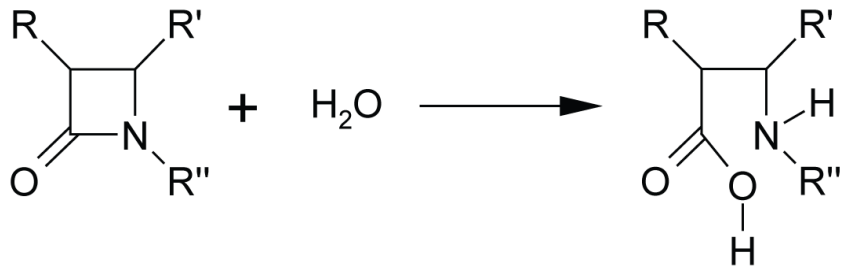


Figure 1.2: β-lactamase reaction.

1.1.2 Fundamentals of the thermodynamics of enzymatic catalysis

The thermodynamics of an enzyme-catalysed reaction can be understood in terms of transition state theory (Pauling, 1946; Garcia-Viloca *et al.*, 2004). The final concentration of substrates and products once the reaction reaches equilibrium (and thus the direction that the reaction takes from its initial conditions) is determined by the free energy difference between substrate(s) and product(s), ΔG. This quantity is not affected by catalysts, and consequently catalysts (including enzymes) have no effect on reaction equilibria. The way that the free energy changes over the course of a reaction can be shown in a diagram that plots the free energy against the progress of the reaction. An uncatalysed reaction is shown in this way in Figure 1.3a. As the reaction proceeds, the free energy of the system first increases, and then decreases. The highest energy, least stable state that the reaction must pass through is known as the transition state.

The turnover number for an enzyme-catalysed reaction, k_{cat}, is determined by the *activation energy* (E_a): the difference in free energy between the substrate(s) and the

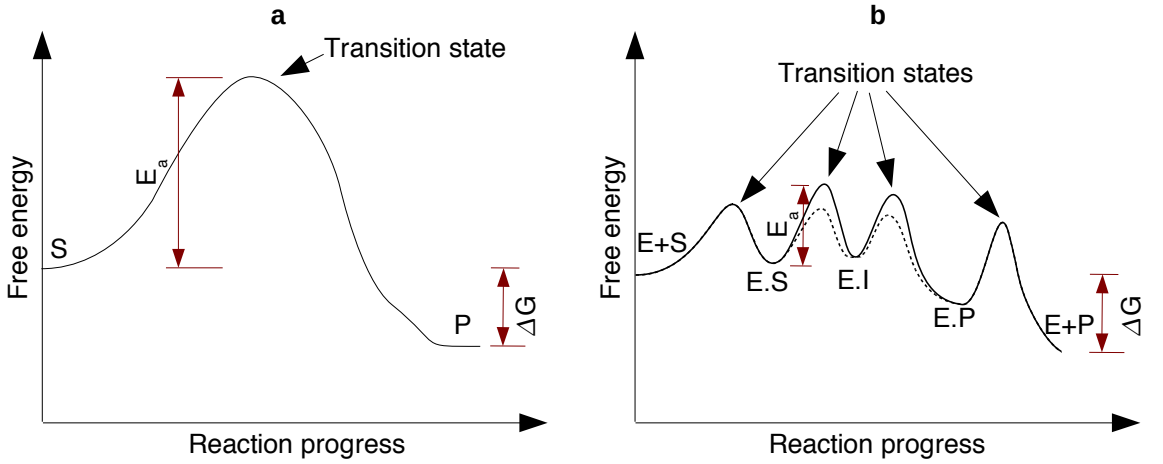


Figure 1.3: Free energy diagrams for an uncatalysed reaction and an enzyme-catalysed reaction.
a. Uncatalysed reaction. Shows conversion of substrate S into product P with activation energy E_a, and free energy change ΔG. b. Enzyme-catalysed reaction. Shows conversion of separate enzyme and substrate (E+S) into enzyme-substrate complex (E.S), enzyme-intermediate complex (E.I), enzyme-product complex (E.P), and finally separate enzyme and product (E+P). The dashed line shows an alternate possible free energy profile for an enzyme where association of enzyme and substrate is the rate-determining step.

transition state, according to the following equation:

$$k_{cat} = Ae^{\frac{-E_a}{RT}}$$

Where A is a constant for the reaction, T is the temperature, and R is the gas constant.

Because of this exponential relationship between reaction rate and activation energy, it follows that a small change in activation energy can bring about a large change in reaction rate. Catalysts, including enzymes, increase the rate of reactions by reducing the activation energy (Figure 1.3b). This involves either stabilising the transition state which would occur in the uncatalysed reaction, or else permitting a different reaction mechanism with a different, lower-energy transition state.

In an enzyme-catalysed reaction, the substrate must first bind to the enzyme to create an enzyme-substrate complex. Enzyme-catalysed reactions (like other reactions) may involve several transition states, separated by stable intermediates; one transition state may present the predominant energy barrier, or there may be several of similar magnitude. Ultimately an enzyme-product complex is formed, which dissociates to leave product, and free enzyme which can begin another cycle. As shown in Figure 1.3b, there is an energy barrier to the assembly of the enzyme-substrate complex, and a barrier to the release of the product. For some enzymes, this association or product release can involve the largest energy barrier (Trentham, 1971; Albery & Knowles, 1976). This possibility is shown by the dashed line in Figure 1.3b. In some cases, such as triose phosphate isomerase, the reaction rate is limited only by the rate at which the substrate can diffuse into the active site (Albery & Knowles, 1976).

1.2 Functions of catalytic residues

The work described in this thesis is concerned with how enzymes operate at the chemical level. Specifically, it examines the geometry and evolution of the individual amino acid residues which contribute to catalysis. This section of the introduction discusses the definition of a catalytic residue, describes those aspects of enzyme function which cannot be ascribed to individual residues, and details the functions which catalytic residues perform.

1.2.1 The definition of a catalytic residue

Individual residues can contribute to enzyme function by binding the substrate, or by being involved in catalysis. Residues can also contribute in more subtle ways: by binding to cofactors, and by maintaining the structure of the active site. A given residue may contribute to both binding and catalysis.

The concept of a catalytic residue is not a clear-cut one, and there is no consistent definition employed in the scientific literature. The process of binding itself makes a contribution to catalysis (as discussed below, and in Jencks & Page (1974)), and some catalytic effects, such as putting steric strain on the substrate or creating a hydrophobic environment, may be spread diffusely over a large number of residues. Furthermore, for those catalytic effects which do not involve the formation or breaking of covalent bonds, contributions range along a continuous scale from large to small, and small contributions may not be experimentally detectable. More strictly speaking, it is not possible to assign a precise, quantitative value to the energetic contribution made to catalysis by any given residue; the effects of catalytic residues are not independent of one another, so it is not possible to dissect their individual contributions (Kraut *et al.*, 2003).

Despite these qualifications, it is possible to identify important catalytic residues, and there are now many enzymes for which the identity and function of key residues contributing to catalysis is known with a good degree of certainty. The work described in this thesis adopts a set of definitions adapted from those set out by Bartlett *et al.* (2002). Residues are defined as catalytic if they play one or more of the following roles:

1. Forming or breaking a covalent bond as part of the catalytic mechanism.
2. Gaining or losing an electron, or acting as a medium for electron tunnelling.
3. Altering the pK_a of a residue or water molecule directly involved in the catalytic mechanism.
4. Stabilising a transition state or intermediate to a greater extent than the residue in question stabilises the enzyme-substrate complex.
5. Activating the substrate in some way, such as by polarising a bond to be broken, or exerting steric strain.

6. Sterically preventing nonproductive chemical reactions.

There are enzymes which have no catalytic residues in the sense defined above. Some rely entirely on cofactors: deacetoxycephalosporin-C synthase (reaction shown in Figure 1.4) catalyses a complex, multi-step redox reaction using only an iron cofactor (Valegard *et al.*, 2004).

Residues which are only involved in substrate or cofactor binding are not regarded as "catalytic residues" for the purposes of this thesis. Despite this, it should be noted that binding of the substrate itself contributes to catalysis in a number of ways. It brings the substrate into an appropriate orientation to interact with the "catalytic" residues, in the sense defined above. Furthermore, where the reaction involves two or more substrates, the enzyme serves to bring them into proximity with one another, thus greatly increasing their effective concentration. Enzyme binding also brings these multiple reactants into an appropriate orientation with regard to one another for reaction to occur. This combination of increased concentration and appropriate orientation is known as the proximity effect (Jencks & Page, 1974).Considered in thermodynamic terms, the enzyme is reducing the negative entropy cost of achieving the transition state, and thus lowering the activation energy. Experiments with equivalent small molecule systems indicate that the proximity and orientation effects each contribute a rate enhancement of around 10^4, for a total enhancement of 10^8 (Page & Jencks, 1971; Jencks & Page, 1974). It has been proposed that enzymes further contribute to catalysis by very precisely positioning the electronic orbitals of the substrate into a suitable conformation for catalysis (Storm & Koshland, 1970); however, the consensus is that the entropic effects described above are sufficient to account for the rate enhancement due to the binding of enzyme to substrate (Jencks & Page, 1974; Fersht, 1999).

1.2.2 Roles played by catalytic residues

This section discusses the chemical roles played by catalytic residues. Non-residue cofactors are critical to the function of many enzymes. However, the work described in this thesis mainly focuses on protein residues, so the functions of cofactors are not discussed here. A given catalytic residue can play a number of roles over the course of a reaction.

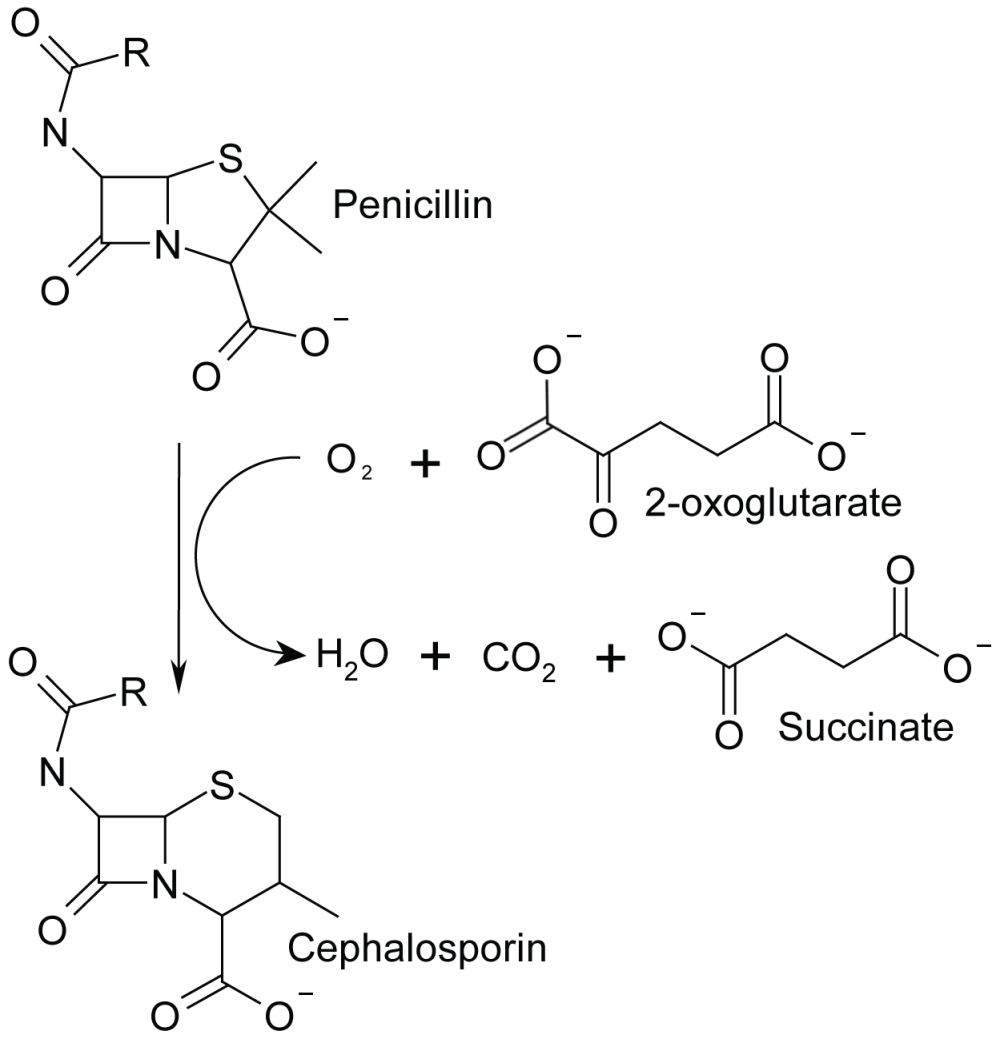


Figure 1.4: Deacetoxycephalosporin-C synthase reaction.

1.2.2.1 Residues forming and breaking covalent bonds

Those catalytic residues which undergo formation and cleavage of covalent bonds are generally more important to catalysis and easier to unambiguously identify experimentally than those catalytic residues that do not undergo any change in covalent bond order. Residues undergoing changes in covalent bond order include those acting as nucleophiles, those acting as electrophiles, those acting as acids or bases, and those which form radicals.

Residues which carry out a **nucleophilic** attack on the substrate produce an intermediate which is covalently bound to the protein. This intermediate must be broken up at a subsequent stage of the reaction. The strength, or "nucleophilicity", of a nucleophile depends on several factors; one of the most important is the basicity of the group (Jencks & Gilchrist, 1968).Nucleophilic residues are often deprotonated by another residue immediately prior to carrying out their nucleophilic attack; this deprotonation creates an unstable, highly basic group. The classic example of a catalytic nucleophile is the serine in hydrolases featuring a Ser-His-Asp catalytic triad, such as chmyotrypsin (Kraut, 1977). This serine is deprotonated by the histidine, priming it for a nucleophilic attack (Figure 1.5). In the case of proteases like chymotrypsin, the serine attacks the electrophilic carbon atom of the carbonyl group in a peptide bond (Hartley, 1964). This results in an intermediate which is covalently bound to the serine nucleophile. This intermediate is then hydrolysed by a water carrying out a nucleophilic attack on the carbon in the intermediate which is directly covalently bound to the serine (Kraut, 1977).

Residues seldom act as **electrophiles**, although some positively charged cofactors such as metal cations and pyridoxal phosphate (Karpeisky & Ivanov, 1966) may do so. There are cases where a residue acts as an electrophile because an intermediate covalently bound to the residue is broken up *via* the nucleophilic attack of another molecule upon the residue. In the case of 4-chlorobenzoyl-coenzyme A dehalogenase (Yang *et al.*, 1996), the residue in question is an aspartate (Figure 1.6). A water molecule makes a nucleophilic attack on the electropositive γ-carbon of the sidechain, which at that stage of the reaction forms part of an ester linkage to the covalently bound intermediate. The γ-carbon is acting as an electrophile.

When residues act as **acids or bases**, this reaction is also nucleophilic in nature.

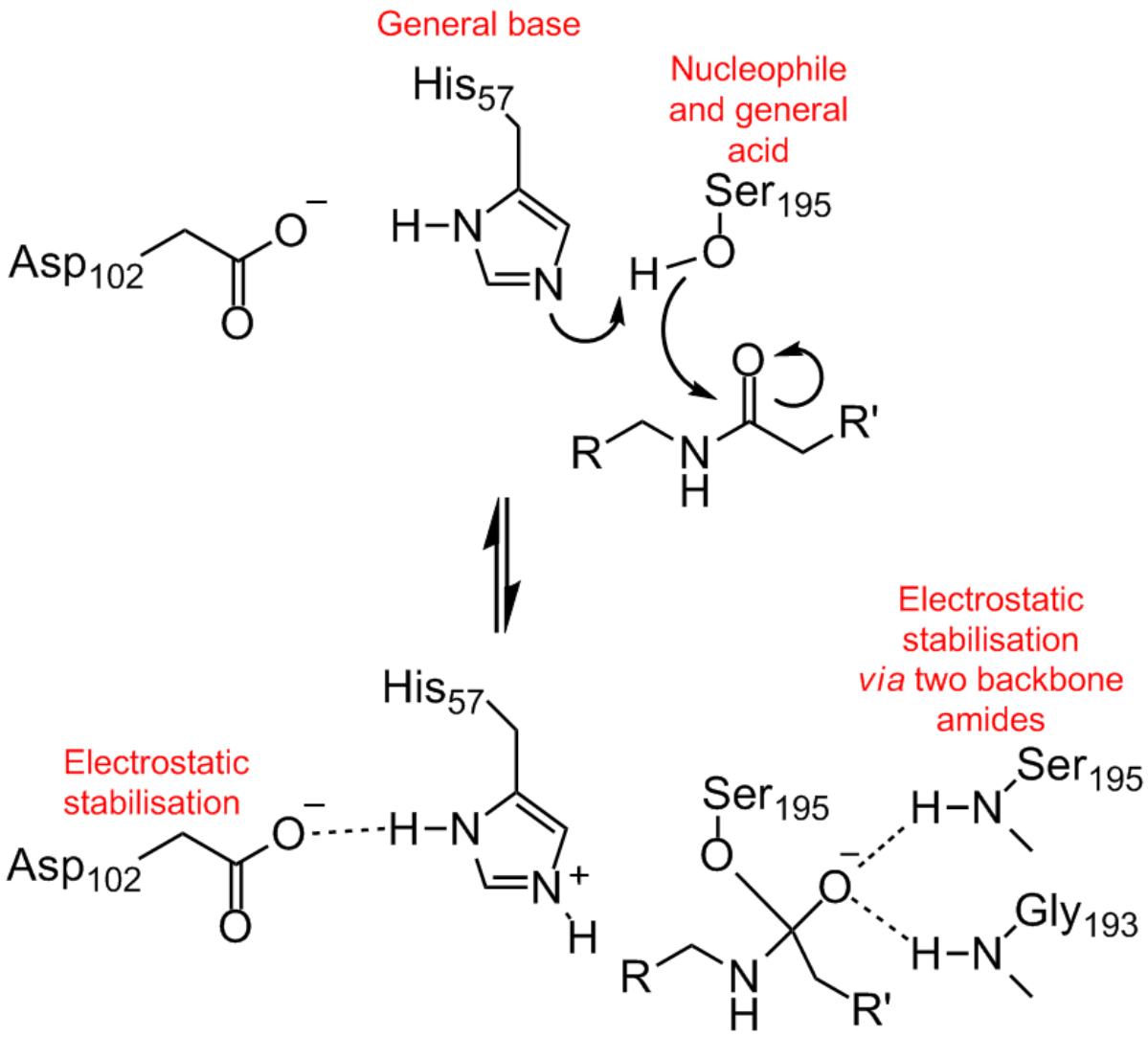


Figure 1.5: Roles of residues in the reaction mechanism of α-chymotrypsin. Only the first step of the reaction is shown, in which a covalent intermediate is formed.

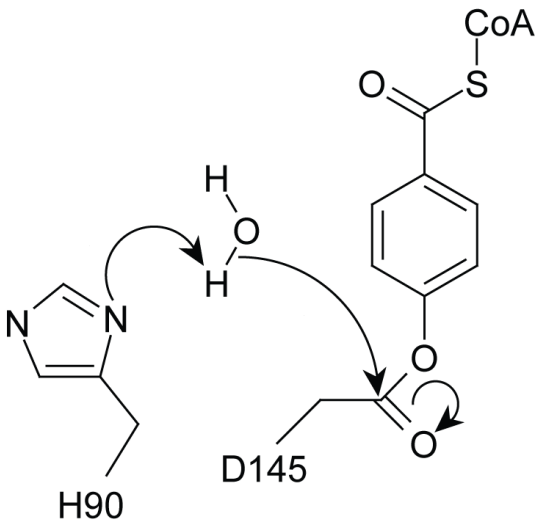


Figure 1.6: Example of a residue acting as an electrophile.
In 4-chlorobenzoyl-coenzyme A dehalogenase, D145 acts as an electrophile in the hydrolysis of a covalent intermediate (Yang *et al.*, 1996).

However, it cannot be described accurately by labelling residues as "nucleophiles" or "electrophiles", and is thus best considered separately. When a residue acts as a Brønsted-Lowry acid/base (proton donor or acceptor), this is referred to as "general acid/base catalysis", as distinct from "specific acid/base catalysis", which signifies the direct action of water in the form of hydronium (H_3O^+) and hydroxide (OH^-) ions. The stronger an acid, the more powerful it will be as a general acid catalyst, and the stronger a base, the more powerful it will be as a general base catalyst Fife (1972); Fersht (1999). However, powerful acids and bases will not be present in their catalytic ionisation state in large concentrations at physiological pH. This means that enzymes generally use sidechains with pK_a values between around 4 and 10 as general acid/base catalysts: aspartate, glutamate, histidine, cysteine, tyrosine, lysine. However, the pK_a values of residues can be considerably altered by their environment in the protein (Copeland, 2000). Residues playing acid/base roles may also act on other residues to prime them for interactions with the substrate; the classic example is the histidine in Ser-His-Asp triads (Figure 1.5), which deprotonates the neighbouring serine, activating this serine for carrying out a nucleophilic attack.

A few enzymes operate *via* **free radical** mechanisms. This generally only involves cofactors, but sometimes residues are used for radical generation and (more frequently)

propagation. For example, in formate C-acetyltransferase (Figure 1.7) a glycine C_α is the source of a radical which then propagates *via* a pair of cysteine sidechains (Leppanen *et al.*, 1999).

Residues seldom undergo **electron transfer** in enzymes catalysing redox processes; this task is usually undertaken by cofactors. Somewhat more frequently, residues act as a medium through which electrons pass when transferring between redox centres by means of quantum tunneling (Gray & Winkler, 1996).

1.2.2.2 Residues which stabilise or destabilise

Residues which do not form or break any covalent bonds can still contribute to catalysis by stabilising transition states and intermediates (to a greater extent than the extent to which they stabilise the enzyme-substrate complex), and by destabilising the substrate and blocking the formation of unwanted products. This may be achieved electrostatically or sterically.

Transition states (and intermediates) often involve unbalanced charges; charged or polar residues can counterbalance these charges, and lower the activation energy (Warshel, 1978). Polar residues often stabilise charge through hydrogen bonding; backbone carbonyl and amide groups can play the same role. Aromatic residues can also provide electrostatic stabilisation via cation-pi interactions (Ordentlich *et al.*, 1995). Catalytic residues which interact electrostatically with other catalytic residues may also play an important role

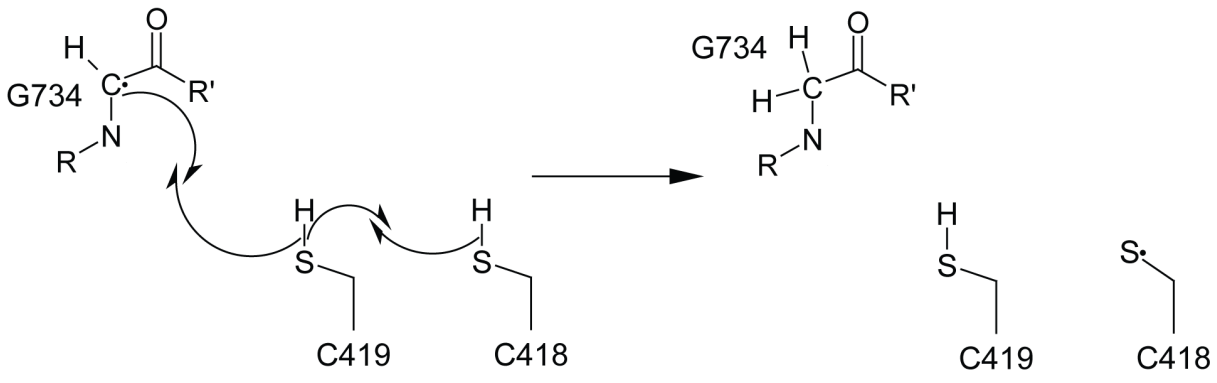


Figure 1.7: Example of residues participating in a free radical mechanism.
In formate-c-acetyltransferase, G734 is the source of a radical that propagates *via* a pair of cysteine sidechains (Leppanen *et al.*, 1999).

in catalysis by altering the pK_a of the other residue, typically making it more able to engage in general acid/base catalysis. The classic example of this is the aspartate in Ser-His-Asp triads (Figure 1.5), which raises the pK_a of the adjacent histidine, making this histidine a better general base (Blow *et al.*, 1969). Non-polar residues may also affect electrostatic catalysis by creating an environment with a lower dielectric constant, altering the behaviour of nearby charged groups (Price & Stevens, 1999).

As described above, enzymes can achieve catalysis in part by binding the transition state more strongly than the substrate. In some cases, the active site exerts **steric strain** to force the substrate to adopt a conformation similar to the transition state. In thermodynamic terms, this reduces the difference in energy between the bound substrate and the transition state, decreasing the activation energy. This steric strain will make substrate binding more difficult, but if the substrate is sufficiently large, the binding energy due to other interactions with the substrate will offset the energetically unfavourable effect of the strain (Jencks, 1975). Steric strain is exerted by the active site as a whole, and is therefore less easy to localise to a single residue than the other catalytic activities described above; however, there are cases where it can be ascribed to one or a few residues (Benning *et al.*, 2000).

There are also some cases where a residue acts to **sterically hinder** the formation of an undesired alternative product (Mancia *et al.*, 1999). It could be argued that this kind of steric hindrance is effectively a variation on the specificity of substrate binding, rather than part of catalysis; it is included here for completeness.

1.3 Experimentally determining catalytic residues and enzyme mechanisms

Determining the identity and roles of an enzyme's catalytic residues is part of the broader task of discovering its catalytic mechanism. A range of experimental techniques can be brought to bear on this problem. These are described below.

1.3.1 Non-structural methods

Experimental methods for identifying catalytic residues and determining the reaction mechanism involve manipulating the enzyme, its substrate, or the reaction conditions, and studying the effects of this manipulation on the kinetics of the reaction. The kinetics can be measured in the steady state, which provides information on the catalytic turnover rate (k_{cat}) and the substrate concentration (K_M) at which the enzyme achieves half of its maximum rate, which provides a rough indication of the affinity of the enzyme for its substrate (L. & Menten, 1913; Fersht, 1999). Further information about the kinetics of individual steps in the reaction can be obtained by analysing the kinetics of the enzyme-catalysed reaction in the short time period before it reaches a steady state: "pre-steady-state kinetics" (Fersht, 1999). This can be studied by rapidly mixing the enzyme and substrate (Hartridge & Roughton, 1923; Roughton, 1934; Fersht, 1999), by using unreactive substrates that can be rapidly activated by laser irradiation ("flash photolysis") (Kaplan *et al.*, 1978), or by relaxation methods Gutfreund (1971), where a reaction at equilibrium is perturbed by a sudden change in temperature, pH, or some other parameter, and then "relaxes" to a new equilibrium.

Perhaps the most commonly employed and most definitive means of testing whether a residue plays a role in catalysis is site-directed mutagenesis. A single residue in the enzyme is mutated in order to discover the effect of altering this residue on function, generally by means of comparing the kinetics of the mutant enzyme with those of the wild-type enzyme. There are a number of methods for achieving the mutation; most are based on oligodeoxynucleotide-directed mutagenesis (Shortle *et al.*, 1981). If a residue is involved in catalysis, then mutating it should affect the rate of catalysis. It is possible for a mutation to affect catalysis by reducing substrate binding, or by disrupting catalysis; kinetic information can discriminate between these possibilities. In brief, if the catalytic turnover rate k_{cat} is reduced, this indicates that the mutation of the residue has affected catalysis rather than binding Plapp (1995); Fersht (1999). Even if a residue is not involved in catalysis, mutating it can affect the rate of catalysis if the mutation disrupts the structure of the enzyme. For this reason, the mutation carried out will generally be one that eliminates the proposed functional group of the residue, whilst making the minimum

possible alteration to the size and polarity (Plapp, 1995; Brannigan & Wilkinson, 2002). For example, replacing Asp with Asn removes its charge without altering its steric bulk, and replacing Tyr with Phe removes its hydroxyl group whilst leaving its phenyl ring in place. A broad assessment of whether the mutation has altered the protein structure can be obtained using spectral techniques such as circular dichroism. Ideally, the structure of the mutant protein would be determined, and compared to the structure of the wild-type protein; however, this is not always possible.

For each of the residue sidechains commonly involved in catalysis, there are one or more compounds available which will react with it in a specific manner that modifies its chemically active moiety and prevents its involvement in catalysis. For example, tetranitromethane will react with the phenol group of tyrosine, nitrating it (Sokolovsky *et al.*, 1966). If an enzyme is inactivated by tetranitromethane, this suggests that it may have a catalytically essential tyrosine residue. However, this may not be a catalytic residue in the sense defined above; it may simply be involved in substrate binding, or it may be that it simply lies near the active site and sterically blocks substrate binding when modified (Bugg, 1997).

These chemical modifications can potentially act at any point on the protein. A more specific chemical modification method, known as "affinity labelling" (Wofsy *et al.*, 1962) involves attaching a chemically reactive group to a substrate analogue. This substrate analogue is bound in the active site, and then chemically modifies a residue at the active site, physically blocking catalysis.

Residues with acidic or basic sidechains need to be in a specific protonation state for effective catalysis to occur. Studying the variation in enzyme activity with pH may reveal a sudden change in activity at a particular pH level, suggesting that there is a critical catalytic residue whose sidechain pK_a has this value (Hammond & Gutfreund, 1955; Copeland, 2000).

Where a rate-limiting step in the reaction involves a group transfer, the rate of the reaction will be slowed if an atom in that group is replaced with a heavier isotope. This kinetic isotope effect can therefore be used to establish which substrate atoms are transferred during the course of a reaction (Northrop, 1975). This effect can be used in concert with the type of pH manipulations described in the previous paragraph to identify residues

acting as acids or bases (Cook, 1991).

1.3.2 Protein structure as a source of information on enzymes

Protein structure cannot always provide definitive information about enzyme function, but it serves as a framework for the interpretation of all other evidence and a basis for the formulation of hypotheses which can be confirmed by other means. This section discusses the essentials of protein structure determination, and then describes the means by which (and extent to which) information about enzyme function can be determined from structure. The reliability of speculations about enzyme function based on structural information depends on the magnitude of errors and uncertainties in protein structures; this topic is therefore also discussed below.

1.3.2.1 Overview of enzyme structure determination using X-ray crystallography

The most common method for determining protein structures is X-ray crystallography. Nuclear magnetic resonance (NMR), neutron diffraction, and electron microscopy can also be used for this purpose; however, all of the structures analysed in detail in the structural analysis chapters of this thesis (Chapters 3–5) were determined using X-ray crystallography.

The electron clouds of atoms scatter X-rays. These X-rays can be derived from a heated cathode source Drenth (1999), or more powerful synchrotron sources (Moffat & Ren, 1997). If a crystal of a macromolecule is produced and an X-ray beam is directed through it, the X-rays will be diffracted by the crystal, creating a diffraction pattern. This diffraction pattern contains information that can be used to reconstruct the details of the electron density in the protein. This diffraction pattern cannot be used in isolation to deduce an electron density map of the protein, because the pattern lacks information on the phases of the scattered X-rays; this phase information can be supplied by a number of methods, including isomorphous replacement, multiwavelength anomalous dispersion, and molecular replacement. Once the electron density map has been determined, an initial model of the atom positions and covalent bond orders in the structure is fitted to

this electron density. It is possible to calculate the diffraction pattern which this model would produce if it were the true structure; this calculated diffraction pattern can then be compared to the actual diffraction pattern. The results of this comparison can be used to improve the model in an iterative process known as refinement (Drenth, 1999).

1.3.2.2 Obtaining crystals of enzymes and enzyme-substrate complexes

The structure of an enzyme is considerably more informative if it features the substrate(s) bound in the active site. It is difficult to obtain structures of enzymes complexed with their substrates, because the enzyme will convert substrate to product on a much more rapid timescale than the collection of X-ray diffraction data. It is generally necessary to use a complex with the product, or to sabotage catalysis in some manner (Fersht, 1999; Price & Stevens, 1999).

Catalysis can be prevented by some modification of the substrates or cofactors. One substrate can be omitted (where there are several substrates) (Eklund *et al.*, 1984), a cofactor can be omitted, a cofactor can be used which is in the wrong oxidation state for the reaction to proceed (Oubrie *et al.*, 1999), or a catalytic metal ion can be replaced by a metal ion that does not facilitate the reaction (Regni *et al.*, 2004). A poor substrate or a competitive inhibitor can be used; these will bind the active site, but will undergo reaction slowly or not at all (Eklund *et al.*, 1984). This inhibitor may be unreactive because it corresponds to only one portion of the substrate, or because one or more of the reactive bonds has been modified.

Alternatively, the conditions under which diffraction data are collected can be altered to slow or prevent catalysis. Low temperatures can be used to slow the reaction (Ding *et al.*, 1994). It is possible to use a pH level where the enzyme is only weakly catalytic because a key catalytic residue is in the wrong protonation state (Fersht, 1999). The use of powerful synchrotron sources for X-ray radiation greatly reduces the time required for data collection; this can be used in concert with techniques for slowing a reaction to obtain a structure of the enzyme-substrate complex.

Finally, it is possible to employ a catalytically inactive mutant form of the enzyme (Campbell *et al.*, 2000), created using site-directed mutagenesis as described in Section 1.3.1.

1.3.2.3 Storage and classification of protein structure data

The repository of all protein structure data is the Protein Data Bank (PDB). This is archive is administered by an organisation called the Worldwide Protein Data Bank (wwPDB), which is a collaboration between several databases which store the information: the Research Collaboratory for Structural Bioinformatics PDB (RCSB PDB, based in the USA), the Biological Magnetic Resonance Data Bank (BMRB, based in the USA), the Macromolecular Structure Database at the European Bioinformatics Institute (MSD-EBI, based in Europe) and the Protein Data Bank Japan (PDBj (Berman *et al.*, 2007)).

The files in the PDB describing protein structures include coordinate data for the atoms in the protein structure, the various parameters described below for expressing uncertainty concerning the protein structure, and a range of other information. This other information include SITE records, which are an optional record of those residues in the protein which are judged by the depositors to be part of important sites in the protein. The concept of "important sites" is not closely defined, so these records may or may not include catalytic residues in enzymes.

The structure in these files typically corresponds to the asymmetric unit of the crystal. The asymmetric unit is the smallest portion of the crystal lattice which can be used to recreate the unit cell by crystallographic symmetry operations. The unit cell is, in turn, the smallest unit which can be translated to recreate the entire crystal. This asymmetric unit may be larger or smaller than the biologically occurring oligomeric state of the protein. The biological oligomeric state may sometimes not be known.

The PQS server (part of the MSD-EBI) attempts to reconstruct the biological oligomeric states of the structures in the PDB. These oligomeric states are predicted by looking at each of the interfaces occurring between protein chains in the crystal, and assessing which ones are specific, biologically relevant interfaces, and which ones are non-specific interfaces corresponding to crystal packing. The assessment of whether a contact is biologically meaningful or not is based on an empirically-weighted score with contributions from solvent-accessible surface area buried in the contact, the number of buried residues at the interface, the estimated change in the solvation free energy of folding due to the interface, the number of salt bridges at the interface, and whether there are disulphide

bridges between the chains (Henrick & Thornton, 1998).

There are two major structural classifications of the data in the PDB: the Structural Classification of Proteins (SCOP (Andreeva *et al.*, 2004)) and CATH (Pearl *et al.*, 2005), the name of which derives from its use of the structural classification levels Class, Architecture, Topology and Homology. These are both classifications of protein domains rather than entire proteins. CATH defines a domain in structural terms as a compact unit capable of independent folding, whilst SCOP defines it as an evolutionary unit observed either in isolation or in multiple contexts in multidomain proteins. Despite this difference in definitions, only around 17% of domain boundary definitions in the two classifications disagree (Orengo *et al.*, 2003).

Both the SCOP and CATH classifications are hierarchical, with higher levels of the hierarchy corresponding to purely structural features of the protein fold, and lower levels corresponding to a classification of structures on the basis of homology. In both cases, the classification is semi-automated and periodically updated, but both inevitably lag slightly behind the expansion of the PDB, and as a result some structures in the PDB at any given time are unclassified by one or both classifications.

The SCOP classification has as its highest level of classification the structural **class**: whether a domain is composed of all α-helices, all β-sheets, an alternating pattern of the two (α/β), a non-alternating combination of the two ($\alpha+\beta$), small domains with little secondary structure, and a few other minor classifications. These classes are subdivided into **folds**: sets of domains with the same secondary structural elements in the same three-dimensional arrangement with the same topology. Unrelated domains can come to have the same fold through evolution converging on solutions to protein structure which are favourable in terms of physics and chemistry; folds are therefore subdivided into **superfamilies** of homologous proteins. These superfamilies are subdivided into families consisting of domains which either have sequence identity levels of 30% or more, or else have very similar structures and functions.

The major levels of classification in CATH are Class, Architecture, Topology and Homologous Superfamily. **Class** broadly corresponds to the class classification in SCOP (although α/β and $\alpha+\beta$ domains are grouped together in CATH). **Architecture** is a classification level falling between SCOP's class and fold levels, denoting proteins which

have the same three-dimensional arrangement of secondary structural elements, but which do not necessarily share the same topology. **Topology** corresponds to fold in SCOP, and **Homologous superfamily** corresponds to superfamily in SCOP. These homologous superfamilies are further subdivided into sequence families with various threshold levels of sequence similarity.

A comparison of the SCOP and CATH classifications found that the classifications were largely in agreement with one another, with the discrepancies between the two largely arising naturally from the different guidelines used for classification (Hadley & Jones, 1999).

1.3.2.4 Obtaining functional information from structure

Knowing the structure of an enzyme allows hypotheses about catalytic mechanism and the roles of residues to be constructed. These hypotheses require confirmation by the use of the experimental techniques described above. In practice, not every residue has its role in the proposed mechanism experimentally confirmed. Although residues which are proposed to make or break covalent bonds in the course of the reaction (mainly nucleophiles and general acids/bases) tend to have their function confirmed by site-directed mutagenesis or other methods, residues which are proposed to play electrostatic roles are less likely to be tested in this manner.

Detailed speculation about the enzyme mechanism generally requires the enzyme substrate, or product (or an analogue of these) to be present in the structure. It is also possible to obtain a complex of the enzyme with a stable compound thought to resemble the transition state: a "transition state analogue" (Schramm, 1998). A complex with a transition state analogue can provide further mechanistic information, including but not limited to those residues involved in stabilising the transition state, and whether the protein undergoes any structural changes in the transition state. Structures can be obtained with a trapped covalent intermediate by various methods, including the use of substrate analogues (Burmeister *et al.*, 1997) or low temperatures (Modis & Wierenga, 2000). This confirms the identity of the enzyme residue responsible for forming a covalent bond with the substrate, and suggests which residues may be responsible for stabilising the intermediate.

The above discussion assumes that the overall function of the enzyme is known. Traditionally, structures would only be determined for proteins whose function was already well-studied, but structural genomics projects are now producing considerable numbers of structures whose function is unknown. As of the end of 2004, major structural genomics consortia had deposited 1540 structures in the PDB; a substantial minority of these are of unknown function (Todd *et al.*, 2005). In most cases, direct functional speculation about these structures is not possible, although it may be possible to apply bioinformatics methods, as described below in Section 1.5.2. In some cases, the structure may have a compound bound which was present by chance in the crystallisation buffer. This compound may be the true substrate; if not, it may nonetheless indicate the general location of the active site (Kim *et al.*, 2003).

1.3.2.5 Positional uncertainty in protein structures

Attempts to derive functional information from protein structures must take positional uncertainties in these structures into account. These uncertainties stem from several sources: the limits on the detail available from the diffraction pattern (quantified by the resolution); the extent to which the diffraction pattern one would expect based on the model corresponds to the true diffraction pattern (quantified by the R-factor); the protein motion and variations between unit cells in the crystal (modelled by the B-factor for each atom). There can also occasionally be large-scale errors in model fitting, which cannot meaningfully be quantified.

For analyses and predictive methods based on the fine detail of crystal structures (such as the methods for predicting enzyme function described below) it is useful to quantify this structural uncertainty in terms of an estimated standard deviation of the atom coordinates; this is known as the "standard uncertainty". A figure of this type can be compared with (for example) the extent of coordinate differences between two superposed relatives in order to determine whether a difference is significant. The Luzzati plot has long been used to obtain an standard uncertainty for a given structure (Luzzati, 1952). However, this is now regarded as providing a crude estimate which often merely gives an upper limit on the value (Laskowski, 2003). More recent methods for calculating standard uncertainty include the σ_A plot (Read, 1986) and another method proposed by

Cruickshank (1999), although this latter method ignores any improvement to precision which comes from the fact that bond lengths and angles have known values, and for this reason it overestimates the true error (Blow, 2002). Broadly speaking, for structures with a good R-factor, standard uncertainty tends to be within one-fifth to one-tenth of the resolution (Rhodes, 2000). The median of values quoted in PDB files is around 0.28 Å(Laskowski, 2003).

X-rays are scattered by the electron clouds of atoms; atoms with higher atomic numbers have a higher electron density, and produce more scattering. Hydrogen atoms have too little electron density for their positions to be determined by X-ray crystallography, except in structures with very high resolution. It is not possible to discriminate between atoms with similar atomic numbers using X-ray crystallography. Since the sequence of a protein is almost always known before its structure is determined, this is not generally a critical problem. However, this ambiguity concerning atom type means that the orientations of the amide groups of the residues asparagine and glutamine can be misassigned, since it is not possible to distinguish between the nitrogen and oxygen. Similarly, the orientation of the imidazole ring of histidine can be misassigned because it is not possible to distinguish between the nitrogen and carbon atoms in the ring. Furthermore, the identity of small molecule ligands (including metal ions) can be uncertain.

1.4 Enzyme evolution

It is possible to use bioinformatics methods (described in Section 1.5 below) to predict whether a protein is an enzyme, what its enzymatic function might be, which residues might be catalytic, and the chemical mechanism by which these residues operate. Many of these bioinformatics methods operate by using the sequence (see Section 1.5.1) or structure (see Section 1.5.2) of the protein of interest to identify its relatives, and then using knowledge about the function of these relatives to infer the function of the protein of interest. The ability of these methods to predict enzyme function depends upon the extent to which sequence and structure vary among enzymes of similar function.

1.4.1 How enzyme function changes as protein sequence diverges

Even very high sequence conservation between enzymes is not a completely reliable indicator of similar function; indeed, it is possible for the same protein to play radically different roles in different contexts (Whisstock & Lesk, 2003). The classic example of this is a protein that serves as a lactate dehydrogenase in some tissues, but also serves a wholly unrelated role as a structural "crystallin" in the eye, where it does not encounter its substrate (Wistow & Piatigorsky, 1987). Conversely, very distant relatives can retain similar functions. Some individual enzyme superfamilies conserve enzyme function across all their members; others are very functionally diverse (Todd *et al.*, 2001).

Several studies have investigated what proportion of homologous pairs of enzymes conserve the same EC classification at various levels of sequence identity. Wilson *et al.* (2000) found that function at the third level of the EC classification was fully conserved above 40% sequence identity, and that at 30% sequence identity, third-level EC function was still conserved in over 95% of cases, although conservation of function rapidly declines at levels of sequence identity below 30%. They also found that the first level of the EC classification (categories such as "oxidoreductase", "transferase", "hydrolase") was fully conserved above 25% sequence identity. Similar analyses by Devos & Valencia (2000) and Todd *et al.* (2001) produced very similar results. These two studies also analysed conservation of fourth-level EC function; Todd *et al.* found that this was conserved in over 85% of cases at 30% sequence identity, whereas Devos and Valencia found that fourth-level EC function was considerably less conserved even at higher levels of sequence identity, being only conserved in 60% of cases at 40% sequence identity.

Rost, however, performed a similar analysis and concluded that there was considerably less conservation of enzyme function (Rost, 2002). He found that fewer than 30% of pairs with sequence identity above 50% had conserved fourth-level EC numbers. The difference between this and the results in the previous paragraph stems from the different datasets employed. A study of this type will necessarily require a dataset that includes some proteins that are related to one another; this raises the question of what dataset to use in order to deal with possible questions of bias. The Wilson *et al.*, Devos and Valencia and Todd *et al.* studies described above employed datasets from the SCOP domain

classification, the FSSP (Holm *et al.*, 1992) structural alignment database, or the CATH domain classification respectively; in employing these datasets they made the assumption that the bias in terms of protein families in the datasets was representative of the bias in whole genomes. Rost argued that this was not the case, and he used a dataset which aimed to reduce this bias. Rost obtained a nonredundant set of protein sequences by grouping protein sequences into families on the basis of similarity as measured by HSSP (Sander & Schneider, 1991), and taking only one sequence from each such family. Each sequence in this nonredundant dataset was then compared with each sequence in a larger redundant dataset in order to obtain a set of pairwise sequence identities and EC number comparisons.

Tian & Skolnick (2003) took a different approach to dealing with the issue of dataset bias: they clustered their dataset into families on the basis of both pairwise sequence similarity and EC classification, measured levels of functional conservation at different levels of sequence similarity in sequence relatives of these families, and then averaged this functional conservation across all families. Their conclusions were intermediate between those of Rost and the earlier studies: they found that third-level EC function is conserved in 90% of cases above 40% sequence identity, whereas the fourth-level EC function is conserved in 90% of cases above 60% sequence identity.

Tian and Skolnick attribute the difference between their results and those of Rost to a number of sources. As noted above, Tian and Skolnick used a different method of grouping proteins into families from Rost. Furthermore, Tian and Skolnick used global sequence identity, whereas Rost used the level of sequence identity over local alignments from PSI-BLAST; Tian and Skolnick also tried using local sequence identity, found that this local sequence identity was less effective for assessing functional conservation (Tian & Skolnick, 2003).

1.4.2 Mechanisms of enzyme evolution

Enzymes can change function over the course of evolution. Most studies assume that enzymes change function from one specialised catalytic role to another specialised role, after being freed to change function, usually through an extra copy of the gene for the

enzyme being created through gene duplication (Zhang, 2003).

In some cases, the enzyme may begin by possessing multiple functions and then specialise over the course of evolution. Jensen proposed that very early in the evolution of life, most enzymes had a broad substrate specificity, and that modern enzymes have evolved by narrowing their substrate specificity (Jensen, 1976). Many modern enzymes have some degree of "catalytic promiscuity". This often takes the form of an enzyme that catalyses a similar reaction chemistry on a range of substrates; for example, chymotrypsin catalyses the hydrolysis not only of peptides, but also esters, thiol esters, acid chlorides and anhydrides (O'Brien & Herschlag, 1999). However, there are other cases where a single active site catalyses reactions that differ significantly; for example, the primary function of thymine hydroxylase is to oxidise a methyl group in thymine, but it also catalyses oxidation of thioethers, hydroxylation of unactivated C-H bonds, and epoxidation (Copley, 2003). There are several cases where an enzyme has a low level of an alternative activity which resembles the main activity of one its relatives. This suggests that the last common ancestor of the two enzymes may have been catalytically promiscuous, and that this aided the evolution of the alternative function (O'Brien & Herschlag, 1999). For example, *E. coli* alkaline phosphatase has a low level of sulphatase activity, and is related to arylsulphatases (O'Brien & Herschlag, 1998). However, the following discussion of theories of functional evolution will assume that in most cases the last common ancestor enzyme was not catalytically promiscuous.

When an enzyme alters its function over the course of evolution, there are several different properties which it might in principle conserve as others change. The catalytic mechanism might remain the same, the substrate specificity might remain the same, or the catalytic architecture might remain the same. These are not mutually exclusive: different properties may be conserved in different cases of enzyme evolution, and there is some overlap between the concepts of catalytic mechanism, substrate specificity, and catalytic architecture. The question of whether each model occurs and (if so) how frequently has been studied by examining the variation in reactions catalysed within homologous superfamilies of enzymes (Gerlt & Babbitt, 2001; Todd *et al.*, 2001).

It appears that most cases of enzyme evolution involve retaining some aspects of the catalytic mechanism while the substrate specificity alters. Todd *et al.* (2001) analysed

evolution within 31 enzyme superfamilies, each of which included enzymes with a range of functions. Details of catalysis were available for 27 of these superfamilies. The analysis found that catalytic mechanism was conserved in four of these 27 superfamilies, and that mechanism is "semi-conserved" (meaning that a common chemical strategy is used in the context of different overall transformations) in a further 18. One of the best-studied cases of conservation portions of catalytic mechanism is the enolase superfamily (Gerlt & Babbitt, 2001). There are at least 12 different reactions catalysed by enolase superfamily members. For all those enolase superfamily members where the catalytic mechanism has been determined, a residue serves as a base to abstract the α-proton of the carboxylate substrate in order to produce an enolate anion intermediate which is stabilised by a Mg^{2+} ion (Gerlt *et al.*, 2005).

It is more unusual for groups of related enzymes to share the same substrate specificity whilst catalysing different reactions. The survey of 31 superfamilies by Todd *et al.* (2001) found only one superfamily where the substrate was absolutely conserved (a superfamily of phosphoenolpyruvate-binding enzymes with a TIM barrel fold). There were a further six superfamilies which bound a common substrate type, such as DNA, sugars, or phosphorylated proteins. Todd *et al.* noted that conservation of catalytic mechanism and conservation of substrate specificity are not wholly unrelated: a common chemistry can sometimes require a common substrate moiety or cofactor.

There are a minority of cases where two homologous enzymes have no similarities in their substrate specificity or catalytic mechanism, but where some features of the active site are nonetheless conserved (Gerlt & Babbitt, 2001). Bartlett *et al.* (2003) analysed 24 pairs of homologous enzymes where the members of each pair differed in their functions. They found six cases where the pair of homologous enzymes shared no mechanistic steps, but nonetheless shared certain active site features, including catalytic residues, metals, and groups binding those metals.

Regardless of whether the enzyme retains substrate specificity, mechanism, or architecture, there are several mechanisms by which the physical changes to the enzyme that lead to a different activity can occur. The most common is likely to be incremental mutations around the active site. Other possibilities include post-translational modification of the protein, gene fusion adding a new domain, and a change in oligomerisation state. The

study by Todd *et al.* (2001) described above found changes in domain organisation within 27 of the 31 superfamilies it surveyed. When a change in enzyme function is caused by the addition of an extra domain to the catalytic domain of the enzyme, this is most frequently due to a change in substrate specificity. The extra domain may alter substrate specificity by providing a specific binding site for the substrate, or by reshaping the existing binding site in the catalytic domain.

The biological function of an enzyme can also be changed through a change in its biological context: change in its expression level, subcellular localisation or substrate concentration may alter its function even when the enzyme remains the same (Todd *et al.*, 2001).

1.4.3 Structural evolution of catalytic sites in enzymes of similar function

Catalytic sites can undergo structural change even when the function of the enzyme remains the same. The positioning of key elements in the catalytic site can alter slightly without affecting enzyme function, and relatives sometimes change the nature or sequence position of catalytic residues without altering enzyme function (Todd *et al.*, 2002). The extent to which catalytic sites can vary structurally without affecting function is relevant to the function prediction methods described in Section 1.5.3. This question has been addressed by studies looking at individual enzyme families.

Wallace *et al.* (1996) analysed Ser-His-Asp catalytic triads from several convergently evolved groups of hydrolases. They found that in the majority of these triads, the distance between the functional oxygens of the Ser and Asp residues was within 1.4 Å of the consensus distance over all triads. They also found that few non-catalytic Ser-His-Asp associations had this conformation.

Proteins from the enolase superfamily catalyse a wide range of reactions, but these reactions share common chemical features (Babbitt *et al.*, 1996). The example mechanisms of enolase and mandelate racemase are shown in Figure 1.8. Meng *et al.* (2004) looked at the structural variability of enolase superfamily catalytic sites, focusing on two catalytic residues and three residues involved in binding a catalytic metal. These residues are con-

served across enzymes with a large number of functions in the enolase superfamily. The study found that even in enzymes with different enzyme functions, these residues display an atom coordinate root mean square deviation (RMSD) of less than 3 Å in almost all cases; the majority of cases showed less than 2Å variation.

There are cases where a pair of homologous enzymes have a set of residues in common which perform the same mechanistic roles, but one or more of these residues occurs at a non-equivalent position in the protein sequence; sometimes even at a totally different location on the protein fold (Todd *et al.*, 2002). For example, in enolase a lysine residue located on the eighth β-strand of a $(\beta\alpha)_8$ barrel stabilises the charge on a carboxylate oxygen in the transition state (Figure 1.8a), whereas in its relative mandelate racemase the lysine that plays an equivalent mechanistic role (Figure 1.8b) is located on the second β-strand (Figure 1.9). Despite these different positions on the fold, the termini of the

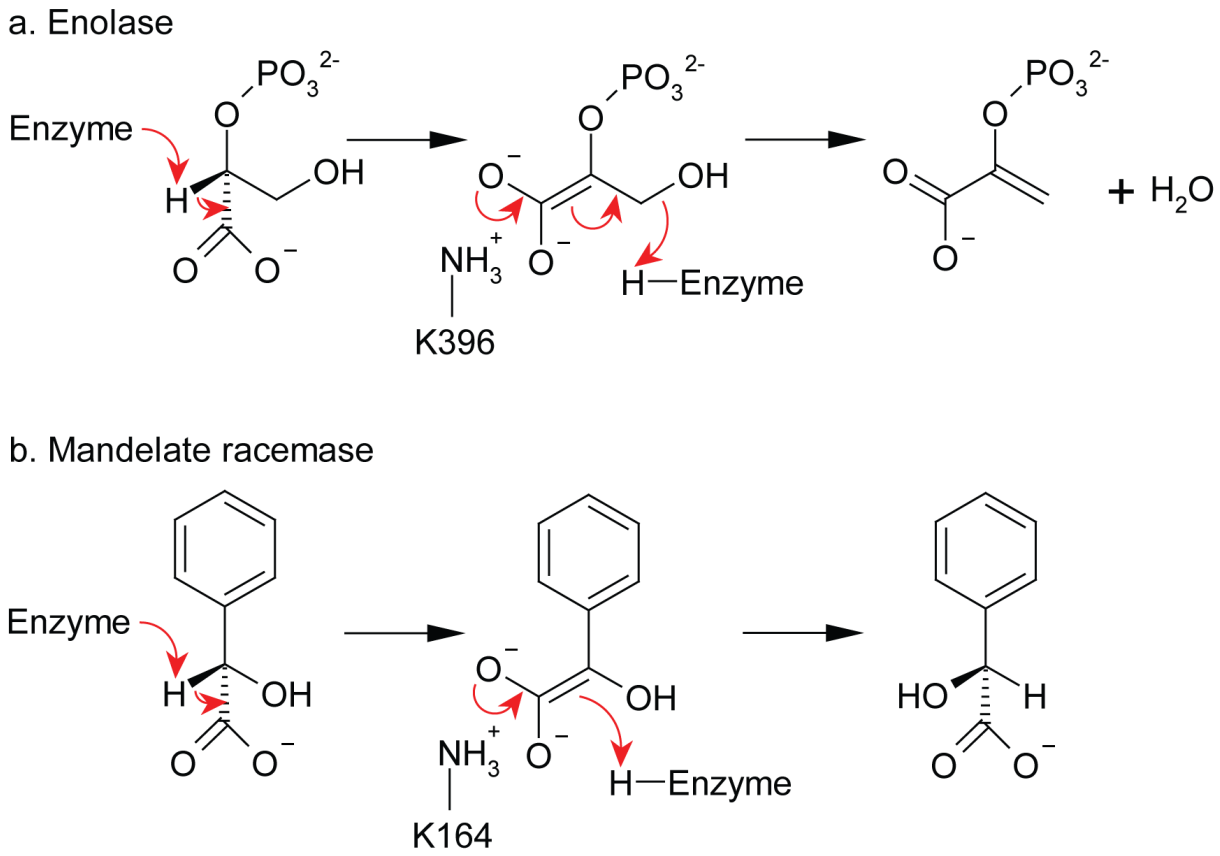


Figure 1.8: Mechanism of enolase and mandelate racemase.
(a) Enolase mechanism (Gerlt *et al.*, 2005). **(b)** Enolase mechanism (Gerlt *et al.*, 2005).

residues lie in similar locations in space relative to other catalytic residues (Hasson *et al.*, 1998). A study by Todd *et al.* (2002) discovered 17 examples of this type, spread over 16 superfamilies. Not all cases follow the pattern of the example just given: in almost half of these cases (eight instances), the residue identity is different in the two proteins; in more than two-thirds of the cases (12 instances) the catalytically relevant atoms of the residues lie in different spatial locations relative to the other catalytic residues. It is not clear why this phenomenon occurs, although several explanations have been proposed. It may be that the alternate catalytic residue location occurred randomly in the course of evolution, and happened to be more catalytically efficient than the original residue, which was subsequently lost. It is also possible that the common ancestor of the two enzymes may have had a third, less efficient, way of achieving this particular catalytic step (or the common ancestor may have had a completely different function and catalytic site), and the two enzymes separately evolved their current solutions. Finally, genetic rearrangements resulting in a circular permutation (where the N and C termini of the protein are fused and new termini are created at another location) may account for some cases.

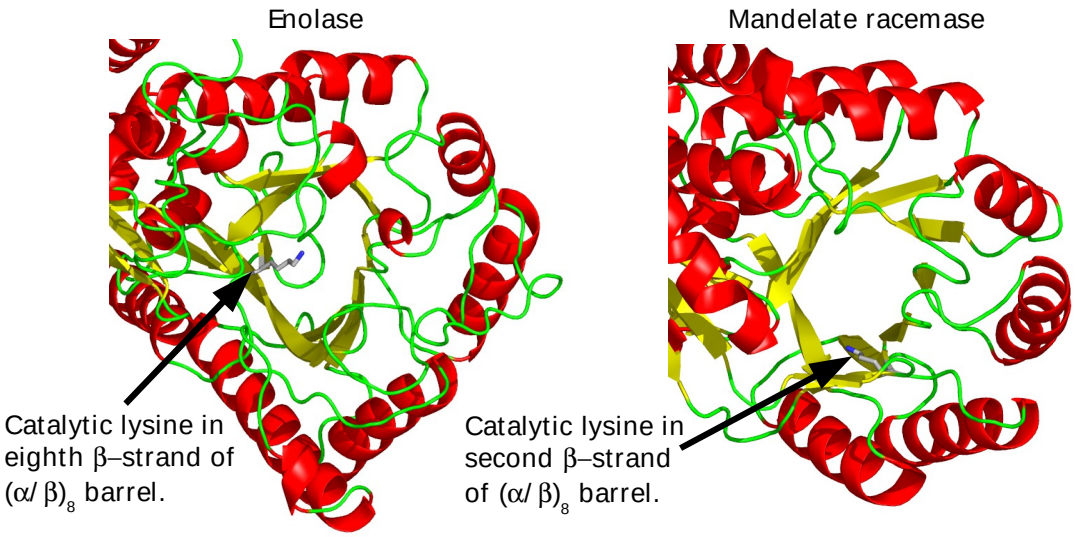


Figure 1.9: Catalytic residues playing equivalent roles that are found in non-equivalent positions in homologues.
The positions of catalytic lysines playing equivalent roles are shown in *Saccharomyces cerevisiae* (PDB entry 1els, Zhang *et al.* 1994) and *Pseudomonas putida* mandelate racemase (PDB entry 1mdr, Landro *et al.* 1994). This figure was prepared using Pymol (www.pymol.org).

1.5 Using bioinformatics to predict enzyme function and catalytic residues

Experimentally determining enzyme function is expensive, difficult, and hard to automate. This section of the current chapter describes those bioinformatics methods that have been developed to infer overall enzyme function and catalytic residues for enzymes where these details have not been established by experiment. The two tasks of predicting overall enzyme function and predicting catalytic residues are separate but often related: some methods predict enzyme function without predicting specific catalytic residues; others predict catalytic residues without predicting overall enzyme function; some predict catalytic residues as a means to predict the enzyme function.

Sequence-based methods can identify relatives, permitting transfer of annotation within the limits of certainty described above in the section on protein evolution (Section 1.4.1). This annotation can include both overall function and the identities of catalytic residues. Protein structure is generally more conserved than protein sequence (Chothia & Lesk, 1986), so methods based on protein structure comparison have the potential to be more powerful than those that only make use of protein sequence comparison. There are also methods which attempt the much more difficult task of predicting function or catalytic residues without inferring these from homologues, by looking for common features shared between unrelated enzymes.

A very large number of protein function prediction methods have been created; these have recently been reviewed by Pazos & Bang (2006), and also by Friedberg (2006). The following section focuses on those methods most applicable to enzyme function prediction, and particularly on approaches involving the matching of protein substructures.

1.5.1 Predicting function using sequence homology

Perhaps the most common use of bioinformatics in protein science is searching for homologues of a protein of interest. The hope is that these homologues are better characterised experimentally than the protein of interest, and that the functional features identified by experiment are conserved between the protein of interest and the homologue. As dis-

cussed above (Section 1.4.1), statistics are available for how frequently function changes with change in sequence, so it is possible to use homology to make informed speculations about overall enzyme function. Both protein sequence and protein structure can be used as a basis for identifying homologues, but sequence-based methods are applicable to a larger number of proteins than structure-based methods, simply because protein sequences are available in larger numbers (Berman *et al.*, 2007; UniProt Consortium, 2007). Where a sequence alignment is obtained, it may be possible to infer the locations of catalytic residues. However, most of the methods described in this section are concerned with prediction of enzyme function, and do not explicitly predict catalytic residues.

Once one or more homologues have been identified (by whatever means), it is useful to have a standardised vocabulary for describing protein functions, in order to be able to transfer the standardised description from the homologue to the protein of interest. For enzymes, the EC classification serves as a standardised vocabulary. However, most methods for function prediction aim to predict the function of non-enzymes as well, and for this reason they frequently make use of the Gene Ontology (GO), which is a hierarchical vocabulary of terms for describing the biochemical functions of gene products, and the biochemical processes and cellular components with which they are associated (Ashburner *et al.*, 2000).

The simplest way to identify a homologue is to search the protein sequence databases for one or more close relatives using a database search method based on pairwise sequence comparison. A number of such methods exist (Brenner *et al.*, 1998), but BLAST (Altschul *et al.*, 1990, 1997) is the most commonly used.

OntoBlast (Zehetner, 2003) and GOblet (Groth *et al.*, 2004) are both methods which employ BLAST to locate a set of homologues, and list the GO terms which these homologues share, using the BLAST E-value to score these matches. GOtcha (Martin *et al.*, 2004) takes a similar approach, but also combines the GO terms from individual matches into a single tree representing the GO hierarchy, with probabilities presented for each GO term.

PSI-BLAST (Altschul *et al.*, 1997) is a search method derived from BLAST which uses an alignment of high-scoring homologues from a BLAST search to create a sequence profile, which is used for further sequence searching. The homologues thus retrieved can

be used to create a new sequence profile and carry out another search; this process can be iterated as often as desired. This allows the detection of more distant relatives. PFP is a function prediction server, which searches for homologues using three iterations of PSI-BLAST, and obtains GO annotation from these homologues (Hawkins *et al.*, 2006). iCSA (George *et al.*, 2005) is a method which begins with an enzyme of known function, constructs a PSI-BLAST alignment of relatives of that enzyme, and predicts similar function for those relatives which have conserved catalytic residues. This method depends on the catalytic residue information in the Catalytic Site Atlas (CSA), which is described in the next chapter.

The above methods combine results from various homologues. However, it has been argued that function is more likely to be conserved for orthologues (homologues originating from speciation) than for paralogues (homologues originating from gene duplication), because a newly created pair of paralogues exist in the same genome, which frees one of them to adopt a different function while the other maintains the original function, as mentioned above (Sjolander, 2004). For this reason, several "phylogenomic" methods have been developed, which construct phylogenetic trees of homologous proteins in order to identify orthologues, and use these orthologues as a basis for predicting function. The methods RIO (Zmasek & Eddy, 2002), SIFTER (Engelhardt *et al.*, 2005), and Orthostrapper (Storm & Sonnhammer, 2002) search for homologues of a query protein, construct a phylogeny for this protein and the homologues, and use this to suggest the likelihood that each homologue is an orthologue, using bootstrap confidence values for the phylogeny to provide a measure of confidence in each case. SIFTER also uses GO to make specific functional predictions.

Homologues can also be identified by searching for sequence motifs expressed as regular expressions. There are various libraries of such motifs, including PROSITE (Hulo *et al.*, 2004), PRINTS (Attwood *et al.*, 2003), and BLOCKS (Henikoff *et al.*, 2000). Although these are not a function prediction method as such, they are normally associated with a specific protein function. Furthermore, these motifs often include key functional residues (including catalytic residues), so their presence may provide further evidence of functional similarity above and beyond homology with the protein from which the motif was derived.

Regular expressions describing sequence motifs are a relatively crude way to describe a

group of proteins. A more sophisticated approach is to use a multiple sequence alignment for a protein family as the basis for a Hidden Markov Model (HMM (Durbin *et al.*, 1999)), which captures the alignment in a probabilistic manner. (This can include an absolute requirement for key functional residues, like the regular-expression style motifs, though it is not mandatory.) This approach is used as a basis for the protein family databases PANTHER/LIB (Mi *et al.*, 2005), Pfam (Finn *et al.*, 2006), SMART (Letunic *et al.*, 2006), and TIGRFAMs (Haft *et al.*, 2003) These families are usually defined as having a specific function, so if the HMM matches a query protein, this implies that the query protein has the same function as the family. The InterPro resource integrates information from many of these motif and family databases (Mulder *et al.*, 2007).

Homologues can also be used to predict functional regions of a protein by identifying those regions which are most conserved in terms of sequence or structure. There are various methods which take this approach; they have been reviewed by Pazos & Bang (2006). However, these methods cannot distinguish catalytic residues from other functional residues, so they will not be described further here.

1.5.2 Predicting function using protein structure to identify homologues

Although far fewer protein structures have been determined than protein sequences, it is possible to discover more remote homologues by comparing the overall fold of a protein. Methods for searching databases of protein structures for those with a similar fold to a query structure include DALI (Holm & Sander, 1993), VAST (Madej *et al.*, 1995), CE (Shindyalov & Bourne, 1998), Matras (Kawabata, 2003), FATCAT (Ye & Godzik, 2004), FAST (Zhu & Weng, 2005), GRATH (Harrison *et al.*, 2003), CATHEDRAL (Pearl *et al.*, 2003) (which derives from GRATH), and AnnoLyze (Marti-Renom *et al.*, 2007). The details of these methods and their performance have been reviewed by Novotny *et al.* (2004) Most of these fold searching methods do not explicitly provide functional predictions, although AnnoLyze provides GO annotation and EC classification predictions, along with probability values scoring their degree of certainty.

For some enzyme homologous superfamilies, function is well conserved, and in these

cases identification of a protein as a member of the superfamily on the basis of structure can imply a shared function. However, many homologous superfamilies are functionally diverse, so if a protein has homologues of known function identified using structure comparison that cannot be recognised using sequence-based methods, generally only a very weak speculation can be made concerning its function. However, analyses of protein structure that focus on details around the putative active site can provide stronger evidence concerning protein function.

It is possible to use protein structure alignments as the basis for a version of the motif-based function predictions described in the previous section. PHUNCTIONER (Pazos & Sternberg, 2004) is a method that uses structural alignments (from the DALI-based resource FSSP (Holm *et al.*, 1992)) of proteins with a given GO function to construct sequence profiles of conserved residues. Protein sequences of unknown function can then be compared with these profiles to predict function.

1.5.3 Recognising distant homologues and cases of convergent evolution using template matching methods

The function of an enzyme depends upon the geometry of the catalytic residues and substrate binding residues. As noted in Section 1.4.3, these key residues can have well-conserved geometries in distantly-related proteins of similar function. If residues with a similar arrangement are found in a distant homologue, this suggests that it may have a similar function; conversely, if key catalytic residues are missing or have a different geometry, this suggests a change in function.

A number of methods exist which search protein structures for groups of residues that have a particular spatial arrangement (Najmanovich *et al.*, 2005). Such groups of residues are referred to in this work as "substructures". A substructure can correspond to the catalytic residues of an enzyme (although it can also correspond to any component of a protein structure). Substructure comparison methods can thus be used to identify catalytic residues, and thus predict enzyme function. The work described in this thesis makes use of substructure comparison methods, and for this reason these methods will be described in detail below.

Substructures can be identified by two types of method: template matching methods and pairwise comparison methods (Najmanovich *et al.*, 2005). Template matching methods search for the presence of a predefined substructure (a template) within a protein. Pairwise comparison methods search for similar substructures between a pair of proteins. These two types of method are obviously similar: template matching is effectively a pairwise comparison between two proteins, where one "protein" consists only of a few atoms and where (typically) all the atoms in this smaller "protein" must be matched. For the purposes of this chapter, the term "structure" will be used to encompass both templates and whole protein structures.

Pairwise comparison methods will not distinguish functional similarities from purely structural ones. However, template matching methods have the disadvantage of requiring a template library to be defined. A number of efforts have been made to create such libraries in a systematic or automated manner; these will be discussed later in this introduction.

This introduction focuses on methods that have the potential to detect cases of convergent evolution as well as distant relatives; that is, those methods which match similar patterns of atoms independently of:

- residue order in the protein sequence
- larger structural features
- bound ligands

Other methods will be briefly discussed later in this introduction.

1.5.3.1 Usefulness of templates for predicting function

The effectiveness of methods for identifying functional sites can be assessed by comparing the results of the method with some external classification of proteins and/or sites, such as the EC hierarchy. Most of the template matching methods described below were tested using a small number of templates, usually including one based on the Ser-His-Asp catalytic triad. However, there have not been any studies which employed a large library

of structural templates to look at how effective structural templates are, on average, at discriminating similar functional sites from random matches.

In order to be useful for function prediction, templates must not merely be capable of detecting functional similarity; they must (in at least a significant minority of cases) be more useful for detecting functional similarity than other existing methods. Template searching is a useful complement to methods based on sequence or overall structure, for the following reasons.

First, there are instances when proteins have independently evolved the same configuration of catalytic residues for carrying out similar reactions– convergent evolution. In these cases it may be possible to predict the common function on the basis of common catalytic residue conformation. This is discussed further in Section 1.5.4.

Second, catalytic residue conformation in homologous enzymes of similar function may remain conserved when the rest of the protein structure has diverged to the extent that it cannot be used to predict function.

Third, even when distant homologues can be identified using sequence methods, their correct sequence alignment may be ambiguous; a structural comparison of catalytic sites may resolve the ambiguity and thus suggest which residues are most likely to be involved in catalysis.

Fourth, even for homologues identifiable by sequence methods, identifying similar catalytic sites that are spread over multiple protein chains may be simpler using structural similarity of catalytic sites than by using sequence comparison.

Fifth, even if two enzymes are clearly recognisable as homologues on the basis of their sequence or overall structure, they may still have different functions (Gerlt & Babbitt, 2001). If they do have different functions, they will often not retain the same catalytic residue conformation– so a consideration of catalytic residue conformation will permit the hypothesis of similar function to be rejected.

1.5.3.2 Components of a substructure searching method

A method for substructure searching can be divided into four components. These are shown in Figure 1.10, and listed here.

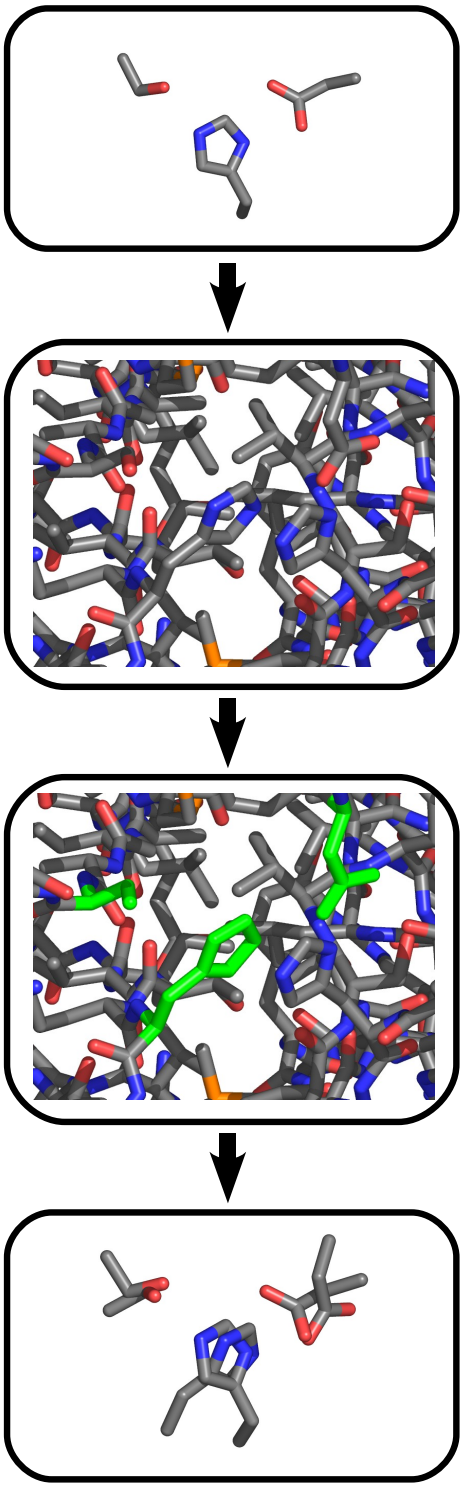
1. The method must specify the features that are to be used to represent the substructure. (Section 1.5.3.3.)

2. Most methods represent the geometry of the structures under comparison using a data structure that aids the search process. (Section 1.5.3.4.)

3. The method must have an algorithm for comparing the two data structures. (Section 1.5.3.4.)

4. The quality of the match between the two structures must then be scored. (Section 1.5.3.5.)

Relevant substructure matching methods are summarised in terms of these four components in Table 1.2, which also provides literature references. Each component is considered in more detail below.



1
The substructure that the program will search for is represented by a set of **features**.

2
The protein structure(s) which are to be compared are represented within the program using a **data structure** which facilitates searching.

3
An **algorithm** is used to locate similar substructures.

4
The **quality** of each match detected by the algorithm must be scored.

Figure 1.10: Components of a substructure matching method. This figure was prepared using Pymol (www.pymol.org).

Table 1.2: Methods for substructure searching. Methods are listed in order of the first associated publication. Where a data structure is not given, this means that the data structure was fairly simple/transparent and best considered together with the search method. The column T/P states whether a method is based on **T**emplate matching or **P**airwise comparison.

Method	Ref	T/P	Features used	Data structure	Search method	Scoring results
Fischer	(Fischer *et al.*, 1994)	P	Each residue represented by C_α position.	Geometric hash	Geometric hashing algorithm. This compares C_α positions within 12.5 radius spheres.	RMSD
ASSAM	(Artymiuk *et al.*, 1994; Spriggs *et al.*, 2003)	T	Each residue is represented by two pseudo-atoms, one near the start of the sidechain and one near its end. Each residue can be labelled with various properties, such as secondary structure type and solvent accessibility.	The pseudoatoms form the nodes in a graph; the edges are the interatomic distances.	Subgraph matching with Ullmann's subgraph isomorphism algorithm.	Maximum difference in distances between an atom pair in the template and an equivalent atom pair in the protein.
TESS	(Wallace *et al.*, 1997)	T	Three atoms from "reference residue" that define the frame of reference, and as many atoms as desired from nearby residues.	The grid positions of nearby atoms relative to the reference residue are stored in a hash table.	Geometric hashing algorithm. This uses a preprocessing step that defines the positions of all atoms in the protein relative to every possible reference residue, using a hash table.	RMSD
PINTS	(Russell, 1998; Stark *et al.*, 2003; Stark & Russell, 2003)	T/P	Each residue is represented by a pseudo-atom at the average position of a set of functional atoms.	N/A	Depth-first recursive search through all possible residue combinations.	E-value calculated using general a priori formula.
Pennec and Ayache	(Pennec & Ayache, 1998)	P	Each residue is represented by C_α and backbone C and N atoms	KD-tree	Resembles geometric hashing. For every possible pairing of residues between the two proteins, a transformation is calculated. Sets of residues with similar transformations indicate a matching substructure.	RMSD, information of the transformation involved (derived from mean Mahalanobis distance).
Fuzzy functional forms	(Fetrow & Skolnick, 1998)	T	No general definition: may include C_α distances, relative sequence positions, and residue position in relation to secondary structural elements.	N/A	Constraints are applied stepwise.	None

Method	Ref	T/P	Features used	Data structure	Search method	Scoring results
SPASM	(Kleywegt, 1999; Madsen & Kleywegt, 2002)	T	Each residue is represented by its C_α atom and a pseudo-atom located at the centre of gravity of the sidechain atoms.	N/A	All sets of residues in the structure that could match those in the motif are recursively generated. Any combination with one or more inter-atomic distances that differ by more than a set cutoff are discarded and the corresponding branch of the search tree is pruned.	RMSD
Functional-group 3D motifs	(Ye. Y. *et al.*, 2000)	T	A template is a set of "functional motifs", potentially overlapping functional groups.	N/A	Depth-first search of all possible functional motif combinations.	RMSD of atom pair distances.
SITEMINE	(Oldfield, 2002)	T	Each residue is represented by a single pseudo-atom at the average position of all atoms for that residue.	N/A	Difference-distance matrix between pseudo-atoms.	RMSD
Singh and Saha	(Singh & Saha, 2003)	T	A template is a set of atoms.	N/A	Iterative closest point algorithm converges on best site and best transformation simultaneously. The single best hit for a given protein is detected.	RMSD
DRESPAT	(Wangikar *et al.*, 2003)	P	Each residue is represented by a single functional atom.	Functional atoms form nodes in a graph; the distances between the atoms are edges.	Recurring patterns within a set of proteins are detected. All possible patterns are enumerated within each protein. All patterns with the same set of residues are compared. Patterns that recur multiple times in the set of proteins are recorded.	RMSD cutoff and significance measure for recurrence of pattern in a set of proteins.
Triads	(Hamelryck, 2003)	T	Each residue is represented by a set of functional atoms. A template consists of three residues.	Each template is a multidimensional vector with information on distances between atoms. Vectors are stored in a sphere/rectangle tree.	Nearest neighbour vectors extracted from a sphere/rectangle tree.	RMSD, E-values based on gamma distribution, Z-values.
NeedleHaystack	(Hoppe & Frommel, 2003)	T	A template is a set of atoms.	N/A	Resembles geometric hashing. A set of three widely separated atoms from the template forms an "anchor"; every possible match to that anchor in the protein is tried. Template and proteins are superposed using the anchor, and template atoms are paired with their nearest equivalents.	"Skip-penalised" RMSD, which increases if all template atoms are not matched.

Method	Ref	T/P	Features used	Data structure	Search method	Scoring results
SuMo	(Jambon *et al.*, 2003)	P	Each residue is represented by a group of functional atoms. These functional groups are gathered into sets of three, and various properties are derived including local atom density and centre of mass of nearby atoms.	Adjacent functional group triangles are connected to form a graph in which each triangle is a vertex.	Triangles from the two proteins match if they contain identical groups and resemble one another with regard to inter-atom distances, atom density and orientation. Matching triangles are stored as vertices in a comparison graph. An edge is set between triangles on this graph if they are adjacent. This comparison graph is used to identify independent regions of similarity.	RMSD
CSC	(Milik *et al.*, 2003)	P	Each residue is represented by one or two functional atoms.	Atoms form nodes in a graph; edges represent whether an atom pair is within a threshold distance or not.	Cliques are derived from the graph representation. Cliques are compared between protein pairs. Those where atom-pair distances are within 1 Å of one another are selected. Overlapping cliques are combined into templates.	RMSD of atom pair distances, ignoring certain distances to allow for motions between rigid subtemplates.
JESS	(Barker & Thornton, 2003)	T	A template is a set of constraints over a number of atoms.	KD-tree	Identifies potential matches in the protein by building on partial solutions.	E-value based on fitting RMSD data from a given template to a normal or bimodal distribution.
PDBSiteScan	(Ivanisenko *et al.*, 2004, 2005)	T	Each residue is represented by its protein backbone atoms.	Not described.	Triplets of residues are compared with template residues, and potential matches are expanded until they match the total number of residues in the template.	Largest single distance between any pair of equivalent atoms in superposed substructures.
Query3d	(Ausiello *et al.*, 2005a)	T	Each residue is represented by the C_α atom and a pseudo-atom at the average position of all sidechain atoms.	Not described.	All possible single-residue matches are considered; each match is extended to discover all possible two-residue matches, and so on until all template residues are matched.	RMSD

1.5.3.3 Features of the substructure

A substructure could be represented by describing the relative positions of all atoms in all relevant residues. However, it can be useful to abstract away some of the details, as this reduces the amount of storage and memory space required for representing the the protein structures to be searched. It also simplifies and thus speeds up the search process. Abstracting away sidechains avoids problems due to ambiguities in experimental identification of atoms (such as sidechain oxygen and nitrogen atoms in asparagine and glutamine residues) and atom nomenclature (such as the equivalent oxygen atoms in aspartate and glutamate residues). Abstraction also makes it easier for the template to match structures which have minor differences in sidechain conformation. The disadvantage of abstraction is the possibility of lowering template specificity.

Some template matching methods treat the structures as sets of atoms, independent of residues, and allow the user to decide which atoms to employ in their template definition (TESS, Singh and Saha, NeedleHaystack, JESS). The Functional-group 3D motif method treats a structure as a set of functional groups; there may be several potentially overlapping groups per residue.

Most methods, however, treat structures as sets of residues and define how each residue is to be represented. Some use C_α atoms (Fischer, fuzzy functional forms) or other backbone atoms (Pennec and Ayache, PdbSiteScan); this can be justified by the argument that backbone conformation is better conserved and more accurately experimentally determined than sidechain positions. However, others reason that the functional roles of residues are carried out by a subset of their sidechain atoms (for example, the carboxyl group of aspartate and glutamate residues), and therefore carry out searching using a set of functional atoms defined for each residue (DRESPAT, Triads, SuMo, CSC). Hydrophobic residues that lack any distinctive functional atoms are often ignored by these methods. Finally, a sidechain may be represented by one or more pseudo-atoms. These are points which do not fall at the position of any one atom, but at the average position of all atoms in the residue (SITEMINE), the average position of a set of functional atoms (PINTS, Query3d) or at some other position derived from the residue atom coordinates (ASSAM, SPASM). SuMo is unusual among the methods described here in representing

residue location by data beyond coordinates. SuMo makes use of a measure of residue burial, and a measure of residue orientation with regard to the centre of the protein.

Some methods only permit residues to match with residues with the same amino acid type. However, many methods allow similar residues to match one another (for example, Asp may match Glu).

1.5.3.4 Data structures and algorithms

Substructure comparison methods must use some internal representation of the structure of the protein(s) being searched (a "data structure"). The choice of data structure is closely tied to the nature of the algorithm used for substructure comparison.

Some methods employ a relatively direct approach of searching through all possible combinations of matching residues, without using any elaborate data structure (PINTS, SPASM, SITEMINE, Query3d, Functional-group 3D motifs). The relatively thorough approach taken by these methods is possible partly because they represent each residue using only one or two atoms. Additionally, PINTS only uses conserved hydrophilic residues. All these methods search through the possible combinations of matching residues, discarding any potential matches as soon as they are found to involve a single distance pair differing by more than a given cutoff. Both PINTS and SPASM use a depth-first recursive search pattern to go through the possible matches.

Several methods employ geometric hashing, a technique derived from the field of computer vision. This requires that the two structures under comparison be in a comparable orientation. Three atoms from one structure are used to define a geometric frame of reference, and three equivalent atoms from the other structure are used to define a frame of reference in an equivalent manner. This has the same effect as if the two structures were superposed using those three atoms, in the hope that these atoms form part of a region of similarity which would be brought into alignment by the superposition. The region of space around the reference atoms is divided into a grid, and the locations of atoms are recorded in a hash in terms of grid cells. (Recording locations in a hash greatly speeds up comparisons.) If the atoms surrounding the reference atoms fall into the same set of grid cells in both structures, this is a sign of a structural similarity. There are many possible equivalent sets of three reference atoms in the two structures; typically, a lengthy

preprocessing step is used to record atom positions in a hash relative to every possible set of reference atoms. Once this is done, comparisons can be carried out relatively swiftly using the hash structure. Geometric hashing was first used for substructure searching in the pairwise comparison method developed by Fischer. TESS applies geometric hashing to template matching. NeedleHaystack and the method of Pennec and Ayache make use of variations on the technique.

Several methods make use of graph theory. There are a number of well-studied algorithms in graph theory for characterising and comparing subgraphs. If protein structures can be represented as graphs, then the substructure searching problem can be reduced to an easily tractable problem in graph theory. A graph representation of a structure typically treats atoms or pseudo-atoms as vertices in the graph, with edges representing distances between pairs of atoms. Graph theory has been previously used in a similar fashion for comparison of secondary structural elements (Mitchell *et al.*, 1990). ASSAM was the first method to apply it to substructure searching, using Ullmann's subgraph isomorphism algorithm to carry out template matching. DRESPAT and CSC are pairwise comparison methods that identify cliques of interacting residues from the graph representation, and compare cliques from a given pair of proteins to look for matches. DRESPAT is the only method that compares substructures over sets of more than two proteins. It does this by carrying out all pairwise comparisons within the set of proteins, and then checking for any patterns identified by the pairwise search that recur multiple times in the set. The authors report that this is considerably more successful in detecting biologically meaningful patterns than pairwise comparison. SuMo uses an elaborate approach in which triangles of three residues are represented as vertices in a graph, and correspondences between triangles in a pair of proteins are represented as another graph.

Some template matching methods treat the template as a set of constraints that have to be met by a number of atoms, and apply these constraints stepwise to potential matches (Fuzzy functional forms, JESS). JESS builds up solutions atom by atom, finding all those atom combinations in the search proteins that satisfy all constraints on a subset of atoms from the template, then adding one more template atom to the subset. In order to increase the efficiency of the search, JESS stores atom data in a data structure known as a KD-tree. This is suited to handling the geometric aspect of the problem.

There are a few methods that don't fit into any of the categories above. The method of Singh and Saha uses the Iterative Closest Point Algorithm, which simultaneously selects the residues which match the template and determines the best transformation to fit the template to the match in the protein, improving both in an iterative manner. The Triads template matching method represents each template as a single multidimensional vector whose components are all the inter-residue distances between the template atoms. These vectors are then stored in a sphere/rectangle tree structure. This is an efficient data structure that can be used for performing nearest-neighbour searches in high-dimensional vector spaces. A set of triads that are nearest neighbours to the template in the tree structure are returned as the results.

1.5.3.5 Scoring results

Substructure searching methods typically return substantial numbers of results. A measure of the accuracy of the results must be used to distinguish close from distant matches. Many methods simply use coordinate root mean square deviation (RMSD). These include Fischer, TESS, SPASM, SITEMINE, Singh and Saha, Query3d, and SuMo. Needle-Haystack, Functional-group 3D motifs, and CSC use slight variations on RMSD. ASSAM uses the maximum difference in distances between a pseudoatom pair in the template and an equivalent pseudoatom pair in the protein. PDBSiteScan uses the largest single distance between any pair of equivalent atoms in superposed substructures. The method of Pennec and Ayache uses a measure derived from the mean of a distance metric (called the Mahalanobis distance) between matching residues.

Whilst RMSD is the standard measure of similarity between two molecular structures, it suffers from not being comparable between different substructures. The other methods detailed above have the same problem. Substructures with a different number of atoms are not comparable (since substructures with a larger number of atoms will tend to have a higher RMSD). Differences in the frequency of residues mean that the residue composition also affects the significance of a given RMSD. If one is considering functional similarity rather than structural similarity, then there is a further problem with RMSD: a given functional motif may resemble a motif that exists for structural reasons. If this is the case, then many hits can be expected from proteins with no functional relationship to the

template. This affects the interpretation of RMSD, or of any other measure that is not specifically calibrated for a given template.

The remaining methods attempt to assign a statistical significance and/or an E-value to a substructure match. (An E-value is the number of matches of the same quality which one would expect to occur at random in the database being searched.) PINTS uses perhaps the most sophisticated E-value scoring available. It converts RMSD values into E-values using a general method that can cope with templates of any size and residue type. The formula used is based on the geometry of a match with a given RMSD level. Several parameters are derived empirically from data. Because the same formula and parameters are used over all templates, the resulting score measure is valid for structural similarity, but not for functional similarity, for the reasons described in the previous paragraph.

Both JESS and Triads derive E-values from RMSD scores by fitting a theoretical statistical distribution to the observed distribution of RMSDs for a given template. JESS employs either a normal or a bimodal distribution; Triads uses a Gamma distribution. The distribution is fitted independently for each template, which means that the results are valid for functional comparisons.

DRESPAT is distinct from the methods described above in that it focuses on the significance of a set of proteins sharing a motif, rather than the significance of a pairwise match. DRESPAT uses RMSD to check for similarity between pairs of patterns. However, it assigns a statistical significance to sets of patterns detected over multiple proteins using an empirically derived formula that takes into account the number of times the pattern recurs, the number of residues in the pattern and the number of proteins searched.

1.5.3.6 Creating template libraries

The usefulness of a template matching method depends on the existence of a library of templates. Several attempts have been made to generate templates in an automated fashion.

Several of these template libraries have been built up using relatively simple criteria or external data sources for creating template patterns. The author of the Triads method extracted all triad templates lying within 4.5 Å of a heteroatom (other than water) from a database of superfamily representatives. SPASM uses a database of motifs that meet one

of three criteria: they consist of a sequential run of residues of the same type (such as four arginines), they are a spatial cluster of residues that share a property (all hydrophobic or all hydrophilic), or they all contact a ligand. PDBSite templates are based on the SITE records of PDB files; as described in Section 1.3.2.3, these SITE records are an optional section of the PDB file where its authors can record any notable features of the structure, and as such they can correspond to many different types of functional site. The PINTS server offers three template databases: one where templates are derived from the SITE records of PDB files, one that focuses on surface residues, and one containing templates based on residues that are within 3 Å of bound ligands. The Query3d method is part of the PDBfun resource (Ausiello *et al.*, 2005b); this permits users to construct templates from a range of functional sites in proteins, including clefts, ligand binding sites, and catalytic residues derived from the CATRES (Bartlett *et al.*, 2002) resource.

Karlin and Zhu developed a more complex method for identifying distinctive residue clusters (Karlin & Zhu, 1996). They converted 3D protein structures into arrays of sequences where residues are ordered by their distance from one another in the protein. They then used techniques for analysing sequence data to identify clusters of residues of a given type, such as clusters of histidines or cysteines, or acidic or hydrophobic patches. These clusters may correspond to functional sites, and could be used as the basis for a template library.

In principle, a pairwise matching program could be used to generate a set of templates based on regions of similarity between proteins of similar function. CSC appears to be the only pairwise matching program to have considered this possibility. CSC takes a specifically enzyme-oriented approach. This uses a pairwise enzyme comparison to extract EC number-specific templates. The authors selected sets of structures with the same EC number, and performed pairwise comparisons of proteins within these sets in order to discover common patterns. These common patterns were used as the basis for structural templates. The resulting templates need not have all distances defined: some distances are left undefined in order to allow for motion between rigid sub-templates. The study was a limited one, in that no algorithm was described for utilising the resulting templates, and in that the process was only carried out for two EC numbers.

A data mining approach was used by the author of SITEMINE to identify common

geometric combinations of residue atoms. This takes into account the symmetry of certain residues (such as aspartate and phenylalanine) and the fact that groups in some residues are equivalent (such as the amide groups in asparagine and glutamine). The hits from this data mining are significant in a geometric sense, but may not be biologically interesting. The data mining output includes information on ligand interactions in order to aid the identification of biologically meaningful hits. Each data mining hit is converted into a template representing the average positions of the equivalent atoms from the mining hit.

1.5.3.7 Other methods for substructure matching

Methods exist that focus not on atomic coordinates but on protein surface properties; these have been reviewed by Via *et al.* (2000). There are methods which require residues to have the same sequence order in the structures being compared, such as Conklin's machine learning approach (Conklin, 1995), SPratt2 by Jonassen *et al.* (2002), and TRILOGY by Bradley *et al.* (2002). Methods that match residues occurring in cavities on the protein surface have been developed by Schmitt *et al.* (2002) and Binkowski *et al.* (2003). A method has been described by Kobayashi & Go (1997) which searches for similar spatial arrangements of atoms surrounding a given ligand moiety. FEATURE (Wei & Altman, 2003) is a template matching program which represents functional sites in terms of the average physicochemical properties of a set of concentric spheres surrounding the site.

1.5.4 Function prediction using protein structure without identifying homologues

It is also possible to attempt to predict a function for an enzyme, or to attempt to identify its catalytic residues, without inferring these functional details from homologues of known function. This is a highly difficult task. Note that some of the methods discussed in this section do involve the identification of homologues, but these homologues are used purely for identifying conserved regions; they are not used as sources of functional annotation.

There are many cases of convergent evolution of enzyme function. In principle, it might be the case that the active sites of these enzymes resemble one another and that these similarities might form a basis for predicting function by comparing non-homologous

structures. A comparison of the SCOP structural classification and the EC classification by Galperin *et al.* (1998) found 34 EC numbers which occurred in more than one SCOP superfamily, implying that convergent evolution had occurred. However, because the EC classification only describes the substrates and products of a reaction, these enzymes may have entirely different mechanisms and active sites. There are cases where unrelated enzymes catalyse similar reactions using residues with similar geometries in similar dispositions relative to one another. The best studied example is the use of combinations of Ser, His, and Asp residues to catalyse hydrolysis reactions by a nucleophilic substitution mechanism. This has independently evolved on at least six occasions, and a number of chemically similar residue groups exist, such as Cys-His-Asp triads (Dodson & Wlodawer, 1998). Substructure comparison methods can be used to detect such similarities, and in fact several of the template matching methods described in the previous section have been shown to be capable of detecting the similar Ser-His-Asp groups in unrelated hydrolase enzymes. However, no general study exists that assesses how practical the detection of such cases of convergent evolution is for function prediction purposes.

It is possible to make an informed guess at the general location of the catalytic site on a protein structure without identifying homologues of known function. In single chain enzymes, the ligand is bound in the largest cleft in the protein in over 83% of cases (Laskowski *et al.*, 1996). PatchFinder (Nimrod *et al.*, 2005) is a method that identifies conserved patches of surface residues; the authors found that in 63% of cases, at least half the residues in PDB file SITE records are in the main patch identified by PatchFinder. These patches are relatively large, averaging 29 residues in size. Thus both cleft and surface conservation methods allow a rough identification of the general region of the active site; however, this does not provide any information about the enzyme function or the precise catalytic residues. EnSite searches for catalytic sites by identifying the largest single surface patch out of the 5% of the molecular surface that is closest to the centroid of the protein (the geometric average of all atom coordinates); this patch was found to overlap with the catalytic site in 74% of cases (Ben-Shimon & Eisenstein, 2005).

There have been a number of attempts to identify individual catalytic residues on the basis of their structural properties and residue conservation alone. Gutteridge *et al.* (2003) attempted to predict catalytic residues on the basis of their residue type, conservation,

depth within the protein, solvent accessibility, secondary structure type and the size of the cleft which they occur in. A neural network was trained with these parameters using a training set of known catalytic residues. This neural network can successfully predict 56% of catalytic residues (sensitivity); however, only one in seven predicted residues is catalytic (specificity). The non-catalytic predicted residues are, however, often close to the catalytic site. This suggests that residues around the active site can be distinguished on this basis, but that these parameters do not permit one to distinguish specific catalytic residues. Another study using machine learning to predict catalytic residues using a similar set of parameters (this time using a support vector machine approach) had a similar level of success (Petrova & Wu, 2006).

Various other approaches have been developed to predict catalytic residues from structure alone; these echo the limited levels of success achieved by Gutteridge *et al.*. SARIG (Amitai *et al.*, 2004) represents proteins as graphs, where residues are the nodes and edges represent interactions between them. Analysis of these graphs has shown that the "closeness" of a residue— the mean distance of its graph node to all other nodes— is significantly higher for catalytic and ligand-binding residues. Using an optimal closeness threshold, it was possible to predict catalytic residues with 46.5% sensitivity— but only 9.4% specificity. THEMATICS (Ondrechen *et al.*, 2001) estimates the pK_a of sidechains on the basis of the protein structure; residues where the predicted pK_a differs strongly from the normal pK_a value for that sidechain are predicted as being catalytic. THEMATICS does not appear to have been tested on a large enough dataset to permit an assessment of its accuracy.

1.5.5 Meta-servers for function prediction

There are a number of meta-servers which implement a range of pre-existing methods for function prediction and produce a summary report. These include InterProScan (Zdobnov & Apweiler, 2001), GeneQuiz (Hoersch *et al.*, 2000), ProFunc (Laskowski *et al.*, 2005), and JAFA (Friedberg *et al.*, 2006).

InterProScan is a search interface to InterPro, the compilation of motif and family data mentioned in Section 1.5.1. GeneQuiz includes sequence searching using BLAST,

motif searching using PROSITE and BLOCKS, as well as methods for predicting a variety of structural features from sequence.

JAFA carries out searches using various tools which express their results using Gene Ontology terms. These include GOFigure, GOblet, InterProScan, GOtcha and PhydBac2.

ProFunc includes a range of sequence based methods (searches for similar sequences in the PDB and in the UniProt sequence database(UniProt Consortium, 2007) using BLAST, searching against various motif and family databases) and structure based methods (fold comparison using SSM, surface cleft analysis, residue conservation analysis, searching for potentially DNA-binding helix-turn-helix motifs). It also makes use of the template matching program Jess to search against banks of templates representing enzyme active sites (taken from the CSA), ligand binding sites and DNA binding sites. Furthermore, each structure used as a query by ProFunc is used to create a set of "reverse templates", which are used to search a representative subset of the structures in the PDB. The template matches obtained by all of these methods are then ranked by comparing the protein environment of the match with that of the original template. In order to carry out this environment comparison, residues in equivalent positions in a 10 Å sphere around the template match are paired up. These paired residues are filtered to include only those where the residues are in the same relative sequence order. These remaining pairs are scored in a manner that takes into account the number of paired residues and the number of insertions that would be required in either protein sequence to bring these residues into alignment.

The GeneFun project has produced a web resource (www.scmbb.ulb.ac.be/GeneFun) that collates functional predictions for protein structures produced by structural genomics projects. These functional predictions include the output of the ProFunc and JAFA meta-servers.

1.6 The structure of this thesis

Chapter two introduces the Catalytic Site Atlas (CSA), a database of catalytic residues in proteins of known structure. This database underlies much of the work in Chapters three and five. Chapter two explains how the CSA has been expanded in its coverage and level

of detail by the present author. The catalytic residue information previously existing in the CSA has been augmented with information on the function of residues, information on the targets that catalytic residues act upon, information on the evidence that each residue is catalytic, and also details of cofactors. These new elements of the CSA are analysed and discussed. This chapter also examines what the domain/EC number combinations in the CSA imply about the divergent and convergent evolution of catalytic function in proteins.

Chapter three makes use of the catalytic residue information in the CSA to examine the structural variation that occurs between catalytic sites in related enzymes, and look at whether this structural variation between two enzymes is related to the evolutionary divergence that has occurred between them. This chapter also examines the utility of structural templates representing these catalytic sites. This analysis of structural template effectiveness includes a comparison of templates that use different sets of atom coordinates to represent the geometry of catalytic residues. The analysis also compares different means of scoring the quality of template matches.

The structural template approach described in Chapter three is capable of a range of applications. In Chapter four this approach is used to examine the evolution of a set of zinc and calcium binding sites serving structural roles. In addition, structural templates representing these zinc and calcium binding sites are used to search the PDB for cases where unrelated proteins have converged upon the same residue selection and geometry for metal binding. This analysis of the evolution of metal binding site structure is extended to examine how frequently metal binding sites are lost in the course of evolution, and what the structural basis of this loss is.

Chapter five returns to looking at the geometry of enzyme active sites. The process of deducing the catalytic mechanism of an enzyme from its structure is highly complex and requires extensive experimental work to validate a proposed mechanism. As one step towards improving the reliability of this process, Chapter five provides an analysis of the typical geometry of catalytic residues with regard to the substrate and one another. In order to analyse residue-substrate interactions, a dataset of structures of enzymes of known mechanism bound to substrate, product, or a substrate analogue is assembled. A separate dataset is produced comprising catalytic residues which act upon other catalytic

residues, based on the expanded information in the CSA. For both datasets, the distances between residues with a given catalytic function and their target moieties are extracted. The geometry of residues whose function involves the transfer or sharing of hydrogens (either with substrate or another residue) is compared with the geometry of a typical hydrogen bond.

Chapter six discusses the limitations imposed on this work by the data employed. It also compares the results from the preceding chapters, and examines possible avenues for future research regarding function prediction and the evolution of functional sites in proteins.

Chapter 2

The Catalytic Site Atlas

2.1 Introduction

There is a huge body of knowledge present in the scientific literature concerning enzymes. There were 4026 enzyme functions classified by the ENZYME database (Bairoch, 2000) using the EC system of classification as of the 2nd May 2007; on the same date there were 21,713 enzyme structures in the PDB. Knowledge that may exist concerning these enzymes includes their substrates, their products, their inhibitors, their kinetic behaviour, how they are regulated, their position in biochemical pathways, their sequence, their structure, their chemical mechanism, and the parts of the enzyme which are involved in substrate binding, catalysis, and regulation.

Some of this knowledge is collected in book compendiums which aim for comprehensive coverage of some aspect: pathways (Michal, 1999), inhibitors (Zollner, 1999), a specific group of enzymes (Barrett *et al.*, 1998), or a brief description of all enzymes (Purich & Allison, 2002). However, putting the information in the form of a relational database makes it swifter to access particular pieces of information, and easier to carry out analyses over large datasets. Various databases incorporate some of this knowledge. The BRENDA database (Schomburg *et al.*, 2002) is primarily concerned with enzyme kinetics and inhibition. MACiE (Holliday *et al.*, 2005) and EzCatDB (Nagano, 2005) are concerned with enzyme mechanism. The KEGG (Kanehisa *et al.*, 2006) and Reactome (Joshi-Tope *et al.*, 2005) databases include information on the position of enzymes within biochemical

pathways.

The Catalytic Site Atlas (CSA) is a database of catalytic residues within proteins of known structure, created by Gail Bartlett and Craig Porter (Porter *et al.*, 2004). The present author has been in control of expanding the depth and coverage of this database from version 2.1 onwards.

There are resources that were in existence prior to the CSA which provide some information on catalytic residues. PDB files themselves contain "SITE records", which list notable features in the protein. The features which are included in this section of the PDB file are at the discretion of the depositor of the structure; in practice, this means that SITE records often do not include details of catalytic residues. UniProt entries include ACT_SITE fields, which provide lists of catalytic residues. However, these tend to be less comprehensive than the lists of catalytic residues found in CSA entries (Porter *et al.*, 2004).

This chapter describes the structure of the Catalytic Site Atlas database. It looks at the growth of the CSA and its current coverage of enzyme structures, and analyses the types and functions of the residues and cofactors in the database.

2.2 The Catalytic Site Atlas

2.2.1 Types of entry in the CSA

Each CSA entry describes a specific catalytic site in a specific protein structure. The CSA includes two types of entry. A *literature entry* describes a catalytic site on the basis of accounts of that catalytic site in the scientific literature. Because these literature entries can only be produced by a human annotator examining the relevant literature, they are time-consuming to produce, and can only cover a fraction of the enzyme structures in the PDB. For this reason, the sequence comparison program PSI-BLAST is used to identify relatives of the literature entries, and the literature annotation is transferred to these relatives, in a manner described below. These entries identified on the basis of homology are referred to as *homologous entries*.

2.2.2 Outline history of the CSA

The level of detail included in each CSA entry has increased over time. Before explaining the nature of individual CSA entries, it is therefore useful to provide an outline history of the CSA. This outline also explains which contributions were made by the present author.

The information in version 1.0 of the CSA was largely compiled by Gail Bartlett, and included 177 entries. Version 2.0 was the work of a larger team of annotators, and included 514 entries. Each entry in version 2.0 contained the same types of information as the entries in version 1.0. The MySQL database that contains this data, the website through which it can be accessed, and the process by which homologous entries are identified, were designed by Craig Porter (Porter *et al.*, 2004).

The further literature entries added in version 2.1 and the current version 2.2 have included more information in each entry concerning the function of individual residues and the evidence that those residues were catalytic. The expanded format of these entries was designed by the present author in collaboration with Alex Gutteridge, and the necessary changes to the database were implemented solely by the present author. These literature entries were annotated by teams of summer students under the supervision of the present author along with Alex Gutteridge (for the entries added in version 2.1) and Gemma Holliday (for the entries added in version 2.2).

2.2.3 CSA annotation

Each entry in the CSA describes the residues which are involved in the chemistry of catalysis. Residues are defined as catalytic if they play one or more of the following roles:

1. Forming or breaking a covalent bond as part of the catalytic mechanism.
2. Gaining or losing an electron, or acting as a medium for electron tunnelling.
3. Altering the pK_a of a residue or water molecule directly involved in the catalytic mechanism.
4. Stabilising a transition state or intermediate, thereby lowering the activation energy for a reaction.

5. Activating the substrate in some way, such as by polarising a bond to be broken, or exerting steric strain.

6. Sterically preventing nonproductive chemical reactions.

Residues which only serve a ligand-binding or regulatory purpose are not included. Residues which are covalently bound to a cofactor may be included if that covalent bond to the cofactor is broken in the course of the reaction, as occurs in many enzymes using pyridoxal phosphate as a cofactor.

A CSA literature entry from version 2.0 or earlier (referred to for convenience below as "low-annotation entries") provides a list of residues meeting these definitions, and specifies whether they act via their sidechain, backbone amide, backbone carbonyl, or some combination of these. N-terminal amino groups are labelled as backbone amides, and C-terminal carboxyl groups are labelled as backbone carbonyls. Whilst these terminal groups have different chemical properties to the backbone groups, they are rarely involved in catalysis, and hence are not given a unique classification in the database. Those few residues which act via their C_α atoms are labelled as acting via their sidechains.

A low-annotation entry does not include cofactors such as NAD or metal ions. A low-annotation entry may include a list of publications relevant to the entry, some introductory information describing the biological significance of the enzyme, and a free text description of the mechanism.

Table 2.1 shows an example low-annotation entry. This describes a serine protease. As related in the "Mechanism" section of the entry, Ser 146 acts as a nucleophile to attack the carbonyl carbon of the peptide bond that is cleaved by the enzyme. This thus meets the first criterion given above for catalytic residues. Similarly, His 397 accepts a proton from Ser 146, and therefore is included as a catalytic residue by the same criterion. Asp 338 alters the pK_a of His 397 to make it a more effective base; it is therefore included as a catalytic residue on the basis of the third criterion given above. Both Gly 53 and Tyr 147 stabilise the negatively charged tetrahedral intermediate which results from the nucleophilic attack of Ser 146 on the substrate. They therefore meet the fourth criterion given above, and so are also included as catalytic residues. These last two residues act via their backbone amides; this is noted in the column titled "Functional part".

Introduction	Carboxypeptidase D is a serine protease which specifically removes basic or hydrophobic residues from the C-terminus of the substrate protein. Carboxypeptidase D is a member of the alpha beta hydrolase family and contains a Ser-His-Asp catalytic triad typical of the family. Carboxypeptidase D from yeast and wheat have had their structures determined, The wheat catalytic triad is made up of residues from both subunits of the homodimer whilst yeast carboxypeptidase D is a monomer, however, both have similar active site geometries.
Mechanism	Carboxypeptidase D uses a catalytic triad to activate serine 146 as a nucleophile to attack the scissile peptide bond. Histidine 397 and aspartate 338 from the neighbouring subunit complete the triad. The backbone amides of glycine 53 and tyrosine 147 make up the oxyanion hole to stabilise the tetrahedral intermediate.

Residue	**Chain**	**Number**	**UniProt number**	**Functional part**
GLY	A	53	62	Backbone amide
SER	A	146	158	Sidechain
TYR	A	147	159	Backbone amide
ASP	B	338	340	Sidechain
HIS	B	397	392	Sidechain

References	PubMed ID 7727364

Table 2.1: Example of a low-annotation CSA entry.
This is the entry for PDB entry 1bcr.

CSA literature entries from versions 2.1 or 2.2 (referred to for convenience below as "high-annotation entries") include more information. For each residue, there is the following information:

- A free text description of its chemical function within the reaction.
- A description of the chemical function chosen from a controlled vocabulary of terms, such as "Electrostatic", "Nucleophile", "Electron tunnelling medium".
- A description of the target that the residue acts upon: substrate, residue, cofactor, water. It is made clear which of the functions chosen from the controlled vocabulary apply to which target. (In the database entries, residues which electrostatically stabilise the transition state are assigned the target description "Transition state"; however, this is largely redundant with the function description "Electrostatic", so

for the purposes of the analysis in this chapter, these cases are treated as having the target description "Substrate".)

- A list of the evidence which suggests that the residue is catalytic. Each piece of evidence describes three things:
  - The publication from which the evidence comes.
  - Whether the evidence relates to the current protein, or to one of its relatives. If the latter, this relative is identified by its UniProt identifier.
  - The nature of the evidence, described by a phrase from a controlled vocabulary, such as "Mutagenesis of residue", "pH dependence of reaction", "Residue is covalently bound to intermediate, based on structural data".

Literature entries from version 2.1 onwards include cofactors as well as residues. "Cofactors" are defined as including groups that are permanently associated with the protein, whether by covalent or noncovalent bonds. This includes all metal ions, iron/sulphur groups, nicotinamide adenine dinucleotide (NAD), nicotinamide adenine dinucleotide phosphate (NADP), flavin adenine dinucleotide (FAD), flavin mononucleotide (FMN), and pyridoxal phosphate (PLP), amongst others. (Note that in this chapter, all metals are referred to by their element abbreviations. These abbreviations are intended to encompass all oxidation states of the metal- for example, a reference to "Fe" should be taken as encompassing Fe^{2+} and Fe^{3+}.)

Table 2.2 shows an example high-annotation entry. This describes a metallo-β-lactamase, which employs a tyrosine and two zinc ions. The entry for the first zinc ion describes its function in free text: "Zn1 polarises the substrate amide to make it more electrophilic. Zn1 is part of the oxyanion hole, stabilising the tetrahedral transition state. Zn1 acidifies a water molecule, protonating the nitroanion intermediate.". The controlled vocabulary description of the function expresses this by stating that the zinc has a function of "Electrostatic" acting on the targets "Substrate", and "Water". The evidence that this is a catalytic zinc all originates from the publication with PubMed ID 9811546, which describes the protein currently under consideration. The evidence in this case is that the presence of the zinc is essential for catalysis, and that structural data shows that it is

appropriately positioned to serve this catalytic role.

2.2.4 Homologous entries

In order to increase the coverage of the CSA, catalytic residue annotation is transferred from literature entries to homologous protein structures. These homologues are identified by searching all protein sequences in the CSA database with the sequence comparison program PSI-BLAST (Altschul *et al.*, 1997). PSI-BLAST carries out an initial BLAST search, and uses all matches with an E-value quality better than 5×10^{-4} to construct a profile that is used for further searches. This process continues for five iterations, after which PSI-BLAST outputs a multiple sequence alignment including all sequences with an E-value better than the 5×10^{-4} threshold value.

Generally speaking, these homologous entries are only included in the CSA if the residues which align with the catalytic residues in the parent literature entry are identical in residue type. In other words, there must be no mutations at the catalytic residue positions. There are, however, a few exceptions to this rule:

1. In order to allow for the many active site mutants in the PDB, one (and only one) catalytic residue per site can be different in type from the equivalent in the parent literature entry. This is only permissible if all residue spacing is identical to that in the parent literature entry, and there are at least two catalytic residues.

2. Sites with only one catalytic residue are permitted to be mutant provided that the residue number is identical to that in the parent entry.

3. Fuzzy matching of residues is permitted within the following groups: [V,L,I], [F,W,Y], [S,T], [D,E], [K,R], [D,N], [E,Q], [N,Q]. This fuzzy matching cannot be used in combination with rules (1) or (2) above.

It is always possible that the catalytic residue assignments in a homologous entry are incorrect. This is a particular danger when the catalytic function of the homologous entry differs from that of the original literature entry. For this reason, homologous entries which have a different EC number from the original literature entry are clearly marked on the CSA website.

Introduction	The L1 metallo-beta-lactamase from Stenotrophomonas maltophilia is unique among beta-lactamases in that it is tetrameric. S. maltophilia has emerged as a significant hospital-derived pathogen of immunocompromised hosts such as cancer, cystic fibrosis and transplant patients. L1 is localised to the periplasm and hydrolyses carbapenem drugs, conferring antibiotic resistance. L1 is of the class 3a metallo-beta-lactamases and binds two Zn(II) ions for the hydrolytic reaction.
Mechanism	1) Zn1 polarises the substrate carbonyl to activate the group as an electrophile. 2) A hydroxide ion bridges the zinc ions. This is nucleophilic and attacks the substrate carbonyl. 3) A tetrahedral transition state is stabilised by Zn1, a helix dipole and Tyr 191. 4) The substrate ring amide is cleaved, with the nitrogen leaving as an anion, stabilised by Zn2 acting as a superacid. The other end of the amide is a carboxylic acid. 5) A water molecule, acidified by the zinc ions, protonates the nitroanion. The resulting hydroxide can be nucleophilic in the next catalytic cycle.

Residue	**Chain**	**Number**	**UniProt number**	**Functional part**	**Function**	**Target**	**Description**
TYR	A	191	212	Sidechain	Electrostatic	Substrate	Tyr 191 is part of the oxyanion hole, stabilising the tetrahedral transition state. This may be directly or via a water molecule.

Evidence from paper	**Evidence concerns**	**Evidence type**
PubMed ID 9811546	Current protein	Residue is positioned appropriately (ligand position known)

Residue	**Chain**	**Number**	**UniProt number**	**Functional part**	**Function**	**Target**	**Description**
ZN	A	269	-	-	Electrostatic	Substrate	Zn1 polarises the substrate amide to make it more electrophilic. Zn1 is part of the oxyanion hole, stabilising the tetrahedral transition state. Zn1 acidifies a water molecule, protonating the nitroanion intermediate.
					Electrostatic	Water	

Evidence from paper	**Evidence concerns**	**Evidence type**
PubMed ID 9811546	Current protein	Ligand is essential for catalysis
PubMed ID 9811546	Current protein	Residue is positioned appropriately (ligand position known)

Residue	**Chain**	**Number**	**UniProt number**	**Functional part**	**Function**	**Target**	**Description**
ZN	A	268	-	-	Electrostatic	Water	Zn2 stabilises the nitroanion intermediate. Zn2 also acidifies a water molecule, to protonate the nitroanion.
					Electrostatic	Substrate	

Evidence from paper	**Evidence concerns**	**Evidence type**
PubMed ID 9811546	Current protein	Ligand is essential for catalysis
PubMed ID 10433708	Related protein: UniProt P25910	Kinetic studies
PubMed ID 9811546	Current protein	Residue is positioned appropriately (ligand position known)

Table 2.2: Example of a high-annotation CSA entry.
This is the entry for PDB entry 1sml.

2.3 Analysis of the contents of the CSA

2.3.1 Coverage growth

There are a number of ways in which one can measure the extent to which the CSA covers the enzymes in the PDB. This coverage has grown with successive versions of the CSA.

The number of literature entries in the PDB has grown from 177 in version 1.0 to 880 in version 2.2 (Figure 2.1). The total number of PDB codes covered (including homologous entries) is currently 18784, which comes close to the 22019 enzymes in the PDB (Figure 2.2).

The number of catalytic domains covered as defined by the CATH classification (Pearl *et al.*, 2005) is increasing, as shown in Figure 2.3. The number of domains covered is only 47% of the 936 catalytic CATH homologous superfamilies in the PDB. However, this figure of 936 catalytic homologous superfamilies is based on all CATH homologous superfamilies featured in PDB entries with non-zero EC numbers; this includes non-catalytic domains, because it is not possible to distinguish which domains are catalytic without knowing

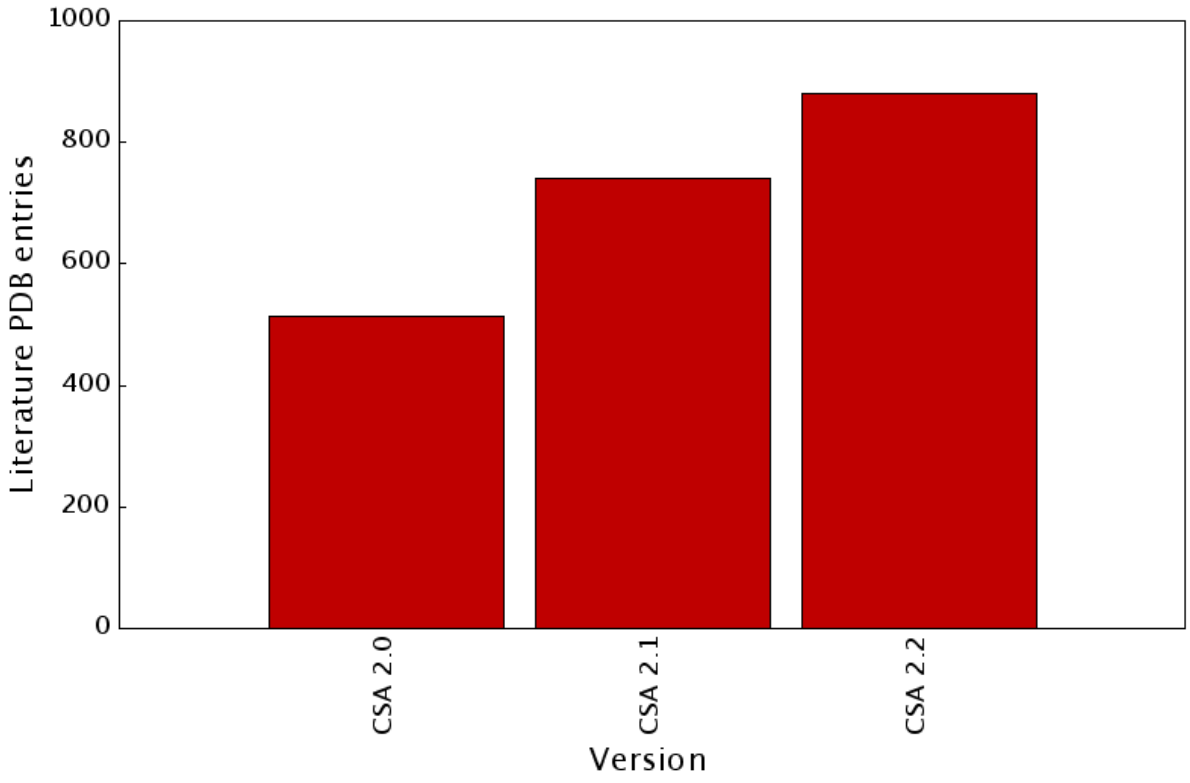


Figure 2.1: Literature PDB entries in the CSA.

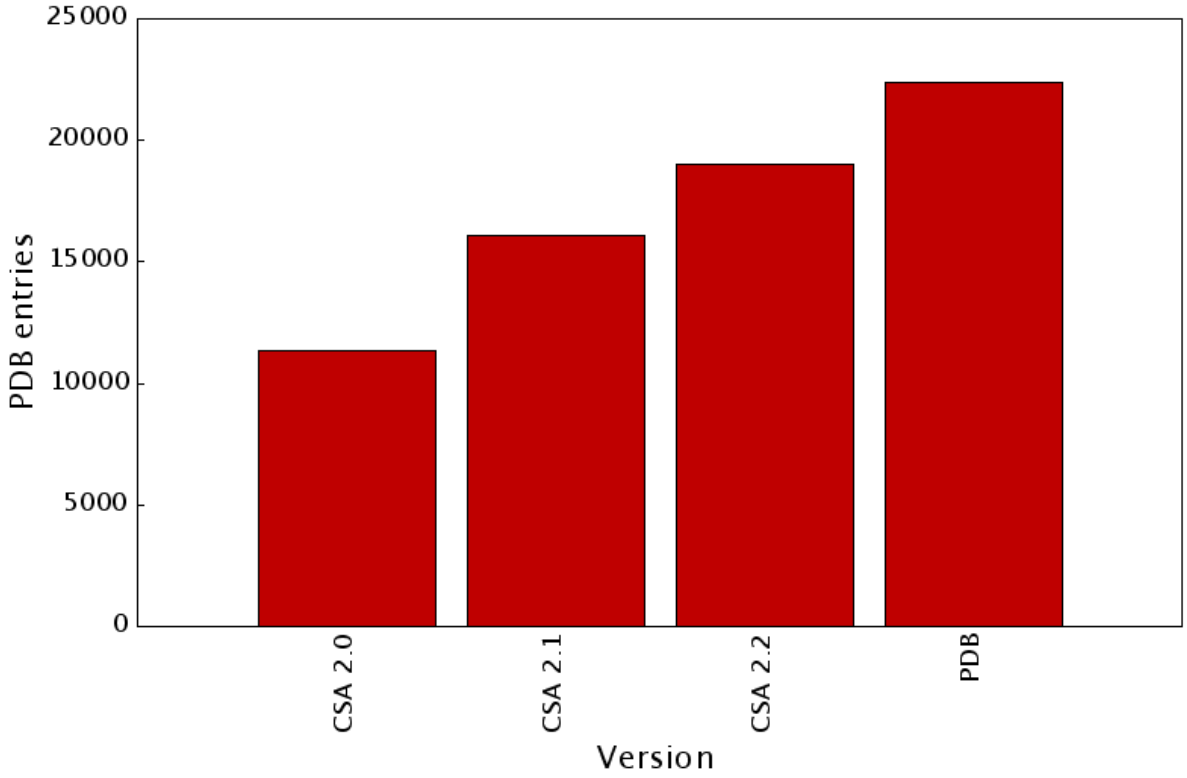


Figure 2.2: All PDB entries in the CSA.
Includes both literature entries and homologous entries.

which are the catalytic residues. For this reason, this figure for all catalytic domains in the PDB is likely to be a considerable overestimate, and the true percentage coverage of catalytic domains by the CSA is likely to be considerably more than 47%.

The range of enzyme functions covered can be measured using the EC classification. As mentioned in Chapter one, the fourth level of the EC classification is usually quite fine-grained, often differentiating between subtly different substrates that undergo the same reaction. For this reason, it is useful to use the range of third-level EC numbers as a measure of coverage, ignoring the last part of the EC number. The range of third-level EC numbers covered by the literature entries is growing (Figure 2.4); 171 numbers are currently covered, which constitutes 75% of the 227 third-level EC numbers in the whole PDB.

It is possible to derive a nonredundant subset of literature entries in the CSA which has no duplication of third-level-EC/CATH-homologous-superfamily combinations. This should represent a count of the number of independent evolutions of third-level EC func-

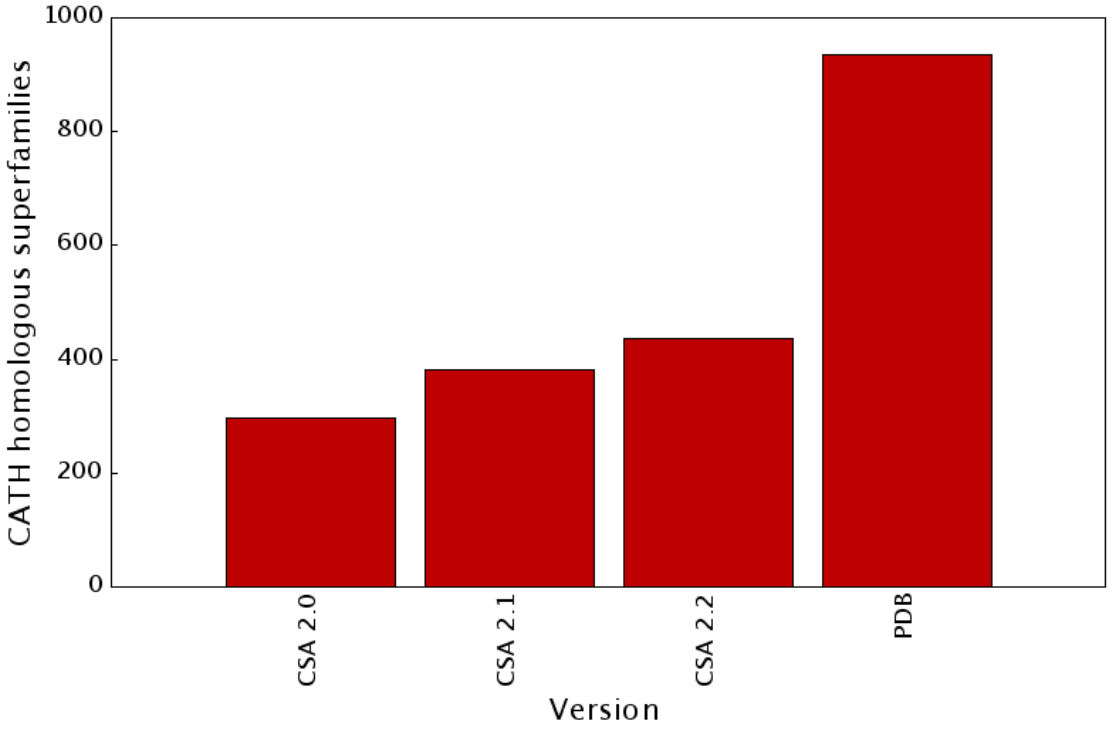


Figure 2.3: Catalytic CATH domains represented in the CSA.

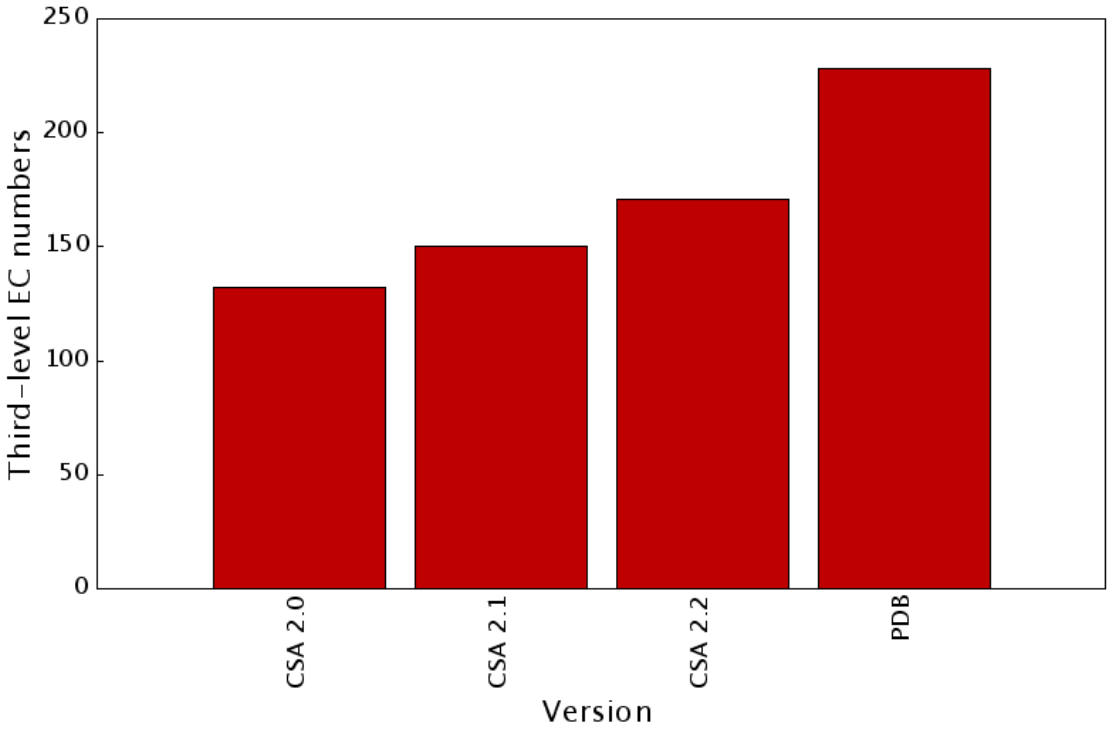


Figure 2.4: Third-level EC numbers represented in the CSA.

tions. The growth in the size of this nonredundant subset is charted in Figure 2.5. This has grown roughly in proportion to the total number of CSA literature entries; there are now 563 nonredundant entries.

2.3.2 Independent evolution of function

Certain third-level EC numbers recur in enzymes with unrelated catalytic domains (as classified by CATH). This suggests an independent evolution of this catalytic function.

The number of independent evolutions of third-level EC numbers were counted by examining all the third-level EC numbers and catalytic CATH homologous domains pertaining to each literature CSA entry. Only domains which contained catalytic residues were included. Where there were two or more literature CSA entries which shared a third-level-EC/CATH-domain combination, all but one of these CSA entries was discarded, because they may not represent independent evolutions of that third-level EC function. All the remaining literature CSA entries were examined, and the number of times that each third-level EC number recurs was recorded. Note that where catalytic residues are spread

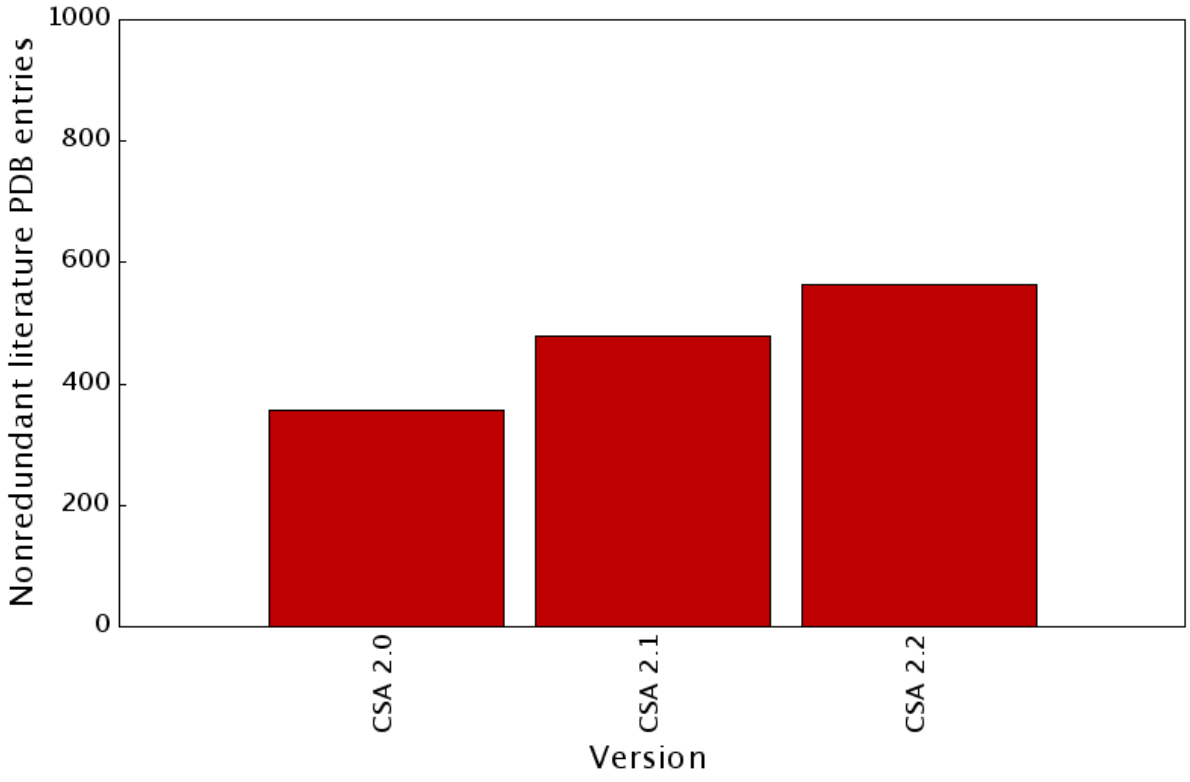


Figure 2.5: Size of nonredundant subset of literature PDB entries in the CSA.

over multiple domains in a single enzyme, this is only counted once (because independent evolutions rather than domains are being counted). Figure 2.6 shows how many functions have apparently evolved only once, and how many have apparently evolved multiple times.

Although the majority of third-level EC functions have evolved only once (66 of the 170 third-level EC numbers for which CATH classifications were available) or twice (29 third-level EC numbers), a considerable minority have evolved more than twice (75 third-level EC numbers). Those functions which have evolved more than 10 times are listed in Table 2.3.

2.3.3 Versatile catalytic domains

Some CATH homologous superfamilies have developed a range of different enzymatic functions over the course of evolution, whilst others have conserved the same enzymatic function across all members.

EC number	Times evolved	Reaction name
1.11.1	10	Peroxidases
3.1.3	10	Phosphoric monoester hydrolases
3.1.1	10	Carboxylic ester hydrolases
3.5.1	11	Hydrolases acting on carbon-nitrogen bonds (other than peptide bonds) in linear amides
2.4.2	12	Pentosyltransferases
1.1.1	12	Oxidoreductases acting on the CH-OH group of donors, with NAD+ or NADP+ as acceptor
2.7.1	13	Phosphotransferases with an alcohol group as acceptor
2.3.1	13	Acyltransferases transferring groups other than amino-acyl groups
4.1.1	14	Carboxy-lyases
2.5.1	15	Transferases transferring alkyl or aryl groups, other than methyl groups
3.2.1	16	Glycosidases
4.2.1	20	Hydro-lyases

Table 2.3: Cases of independent evolution of third-level EC numbers.
This table lists all EC numbers in the CSA which have evolved 10 or more times (as judged by CATH homologous superfamily assignment).

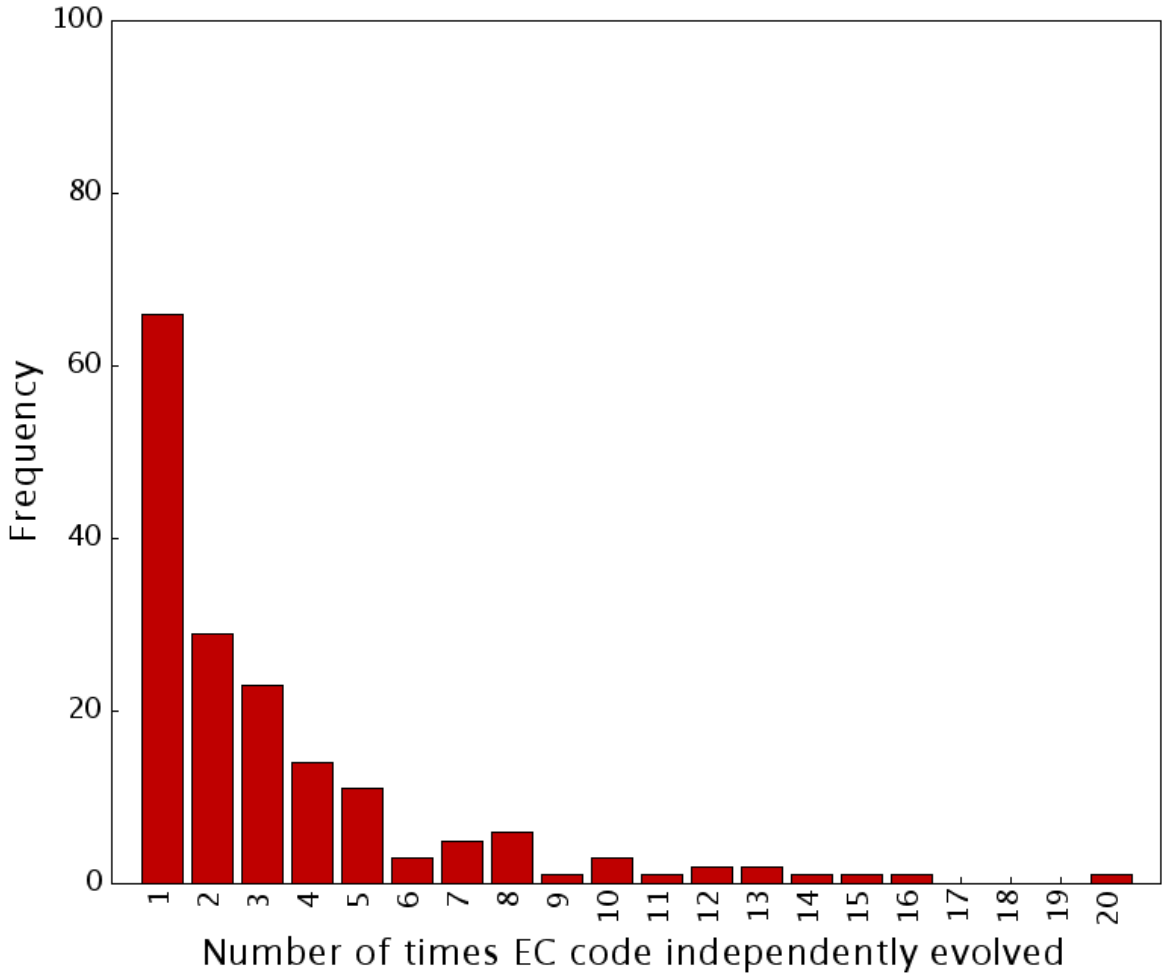


Figure 2.6: Cases of independent evolution of enzymatic functions.
Each y-axis value shows how many different third-level EC numbers have independently evolved the number of times specified on the x-axis.

The number of functions for each domain was counted by recording all the third-level EC numbers that were associated with a given CATH domain in any CSA literature entry. (Note that this differs significantly from the approach taken to count the number of independent evolutions of each third-level EC number described in the previous section.) Figure 2.7 shows how many domains in the CSA have only a single third-level EC number, and how many are associated with multiple EC numbers.

The majority of domains in the CSA (321 out of 437) conserve the same third-level EC number across all their CSA PDB entries. A significant minority have two to four different third-level EC numbers (102 domains), and there are a very small number of highly versatile domains with many third-level EC numbers (14 domains). These highly

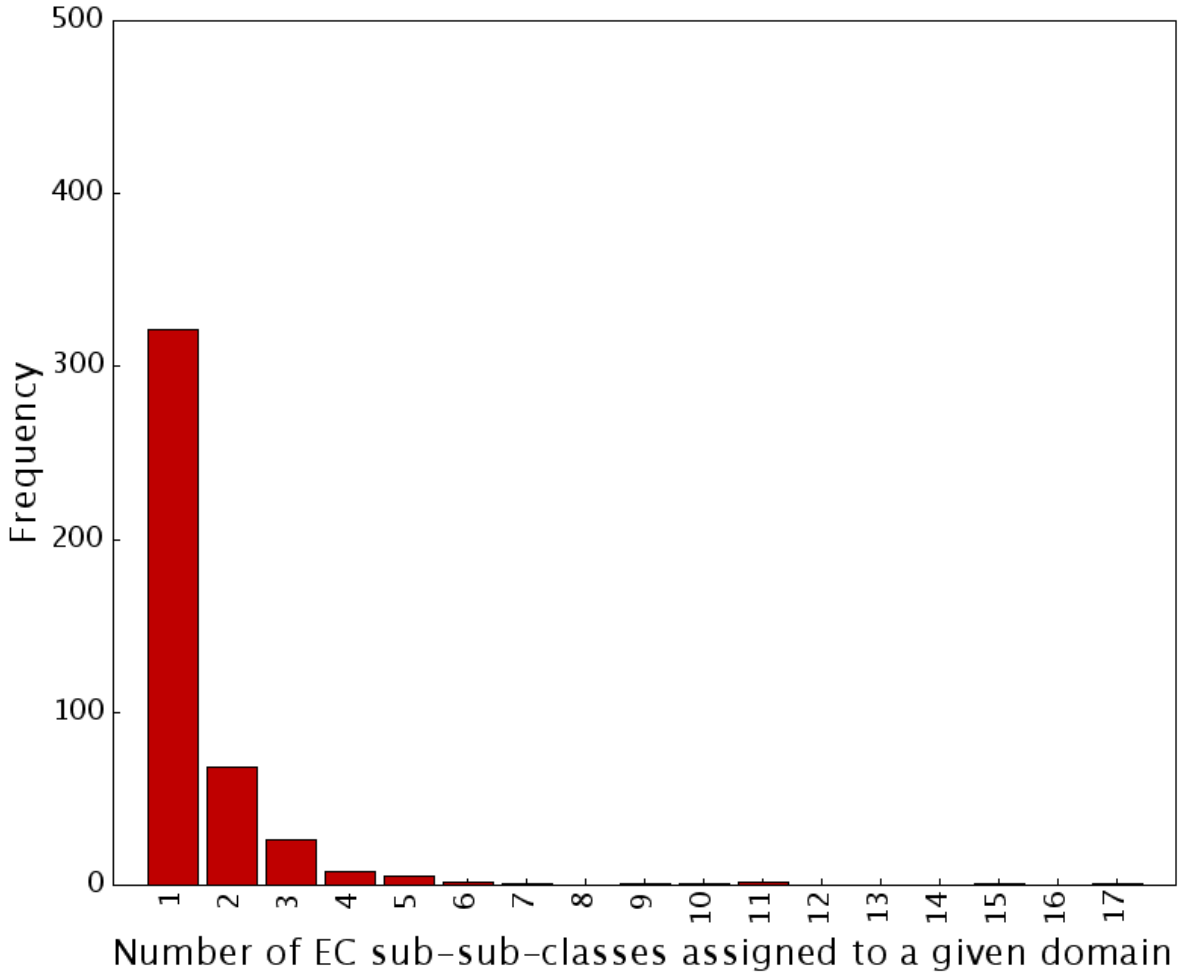


Figure 2.7: Cases of domains with multiple functions.
Each y-axis value shows how many different CATH domains have have the number of different third-level EC numbers specified on the x-axis.

versatile domains are listed in Table 2.4.

2.3.4 Nonredundant dataset

The following analyses are based on a nonredundant subset of the literature entries in the CSA (with no duplication of third-level-EC/CATH-homologous-superfamily combinations). This has 563 members.

2.3.5 Total number of residues

Figure 2.8 shows the distribution of the number of residues (ignoring cofactors) employed by each catalytic site in the nonredundant set of literature entries. Although there is

CATH code	Functions	Domain name
2.40.10.10	5	Trypsin-like serine proteases
3.20.20.60	5	Phosphoenolpyruvate-binding domains
3.30.360.10	5	Dihydrodipicolinate reductase, domain 2
3.40.50.970	5	Lyase
3.60.20.10	5	Glutamine phosphoribosylpyrophosphate, subunit 1, domain 1
3.20.20.140	6	Metal-dependent hydrolases
3.40.50.1000	6	None
3.90.226.10	7	2-enoyl-CoA Hydratase, chain A, domain 1
3.50.50.60	9	Hydrolase inhibitor
3.40.640.10	10	Type I PLP-dependent aspartate aminotransferase-like
3.40.50.1820	11	α/β hydrolase
3.40.50.300	11	P-loop containing nucleotide triphosphate hydrolases
3.40.50.720	15	NAD(P)-binding Rossmann-like domain
3.20.20.70	17	Aldolase class I

Table 2.4: Versatile domains.
This table lists all CATH domains in the CSA which have 5 or more third-level EC numbers.

considerable variation, 85% of the enzymes employ between one and four residues.

Note that it is possible for a catalytic site to have no catalytic residues (for example, in an enzyme that achieved catalysis simply by binding two substrates and bringing them into an appropriate orientation to react with one another). However, because the CSA is essentially a database of catalytic residues, such sites would not be recorded in the CSA. For this reason, there are no catalytic sites with zero residues listed here.

In 17% of cases only a single residue is used. At the other extreme, there are individual sites with ten, fifteen, and twenty-four residues. The case with ten catalytic residues is *Nicotiana tabacum* 5-epi-aristolochene synthase (PDB entry 5eat), and the case with fifteen residues is *Alicyclobacillus acidocaldarius* squalene-hopene cyclase (PDB entry 2sqc); both of these catalyse complex cyclisation reactions on relatively large substrates. The case with twenty-four residues is *Escherichia coli* ubiquinol oxidase (PDB entry 1fft), which has complex electron and proton transfer chains.

The roles played by residues in enzymes with such a large number of catalytic residues

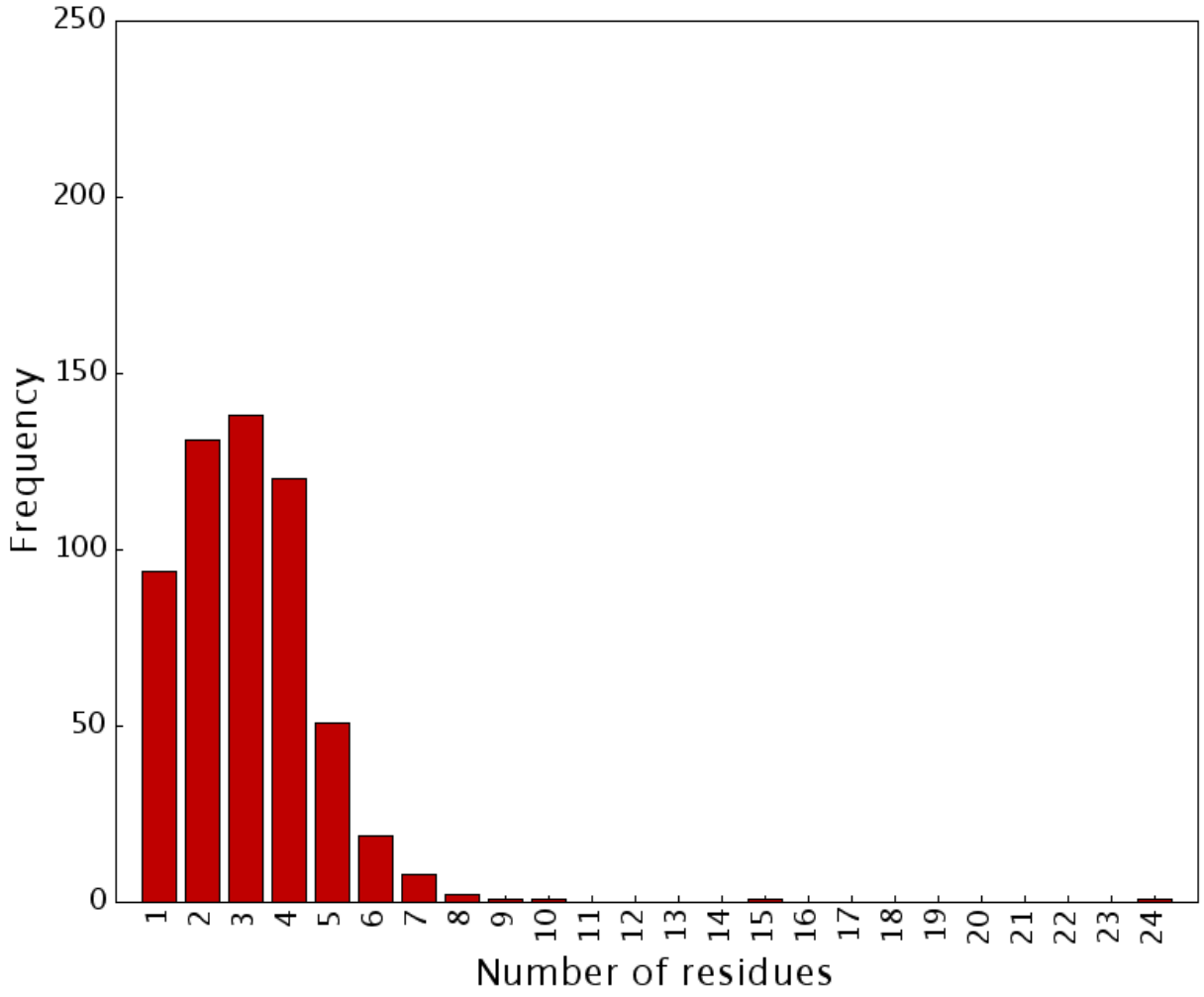


Figure 2.8: Distribution of number of catalytic residues per enzyme.
This analysis applies to residues (other than cofactors) from the nonredundant set of literature CSA entries. Each y-axis value shows how many different literature entries have the number of catalytic residues specified on the x-axis.

can be illustrated by considering the case of 5-epi-aristolochene synthase in more detail. This enzyme converts farnesyl diphosphate into the bicyclic sesquiterpene 5-epi-aristolochene. The roles played by residues in the mechanism of this enzyme (Figure 2.9) were proposed by Starks *et al.* (1997) on the basis of the protein structure, mutagenesis evidence, and previous studies of the chemical changes undergone by the substrate in the course of the reaction (Cane, 1990).

The reaction begins with the separation of the pyrophosphate group. The extra negative charge on this pyrophosphate group would be stabilised by the positively charged residues Arg264 and Arg441, and also by three magnesium ions (Figure 2.9a). Constrain-

ing the pyrophosphate in this area of positive charge serves to keep it separate from the reactive cation intermediates formed by the remainder of the substrate in subsequent stages of the reaction.

The backbone carbonyls of residues 401 and 402 are not involved in hydrogen bonding in the α-helix in which they occur; instead they are directed towards the positive charge delocalised over C1, C2, and C3 at this stage (Figure 2.9b). The hydroxyl of the sidechain of Thr403 is similarly oriented to stabilise this charge.

The subsequent cyclisation of the substrate is proposed by Starks *et al.* to be aided by stabilisation of the charge that develops on C11 via a cation-π interaction with the phenyl ring of Tyr527, and by abstraction of a proton from C13 by Asp525 (Figure 2.9c). The proton acquired by Asp525 is transported away by a proton transfer chain including Tyr520 and Asp444, which lie within hydrogen bonding distance of one another (Figure 2.9d). Tyr520 then donates its proton to the double bond at C6 in the intermediate, and this residue could accept a proton from Asp444 in a concerted fashion (Figure 2.9e). This results in a bicyclic cationic intermediate with a charge on C3 which would again be stabilised by the backbone carbonyls of residues 401 and 402 and the hydroxyl group of the sidechain of Thr403.

Subsequent rearrangement results in an intermediate with a positive charge on C7, which would be stabilised by the aromatic indole group of Trp273 (Figure 2.9f). Ultimately, Trp273 abstracts a proton from C8 in a final intermediate, producing the product (Figure 2.9g and 2.9h). Trp273 is the only residue positioned to accept a proton from this atom without substantial reorientation of the cyclic intermediate; furthermore, mutation of this residue to nonaromatic or aromatic residues prevents synthesis of 5-epi-aristolochene (Starks *et al.*, 1997).

The locations of the above residues relative to the substrate are shown in Figure 2.10.

2.3.6 Catalytic residue frequency

Figure 2.11 shows the number of each residue type occurring in the nonredundant set of literature entries, broken down into residues acting via their sidechains, residues acting via their backbone amide groups, residues acting via their backbone carbonyl groups, and

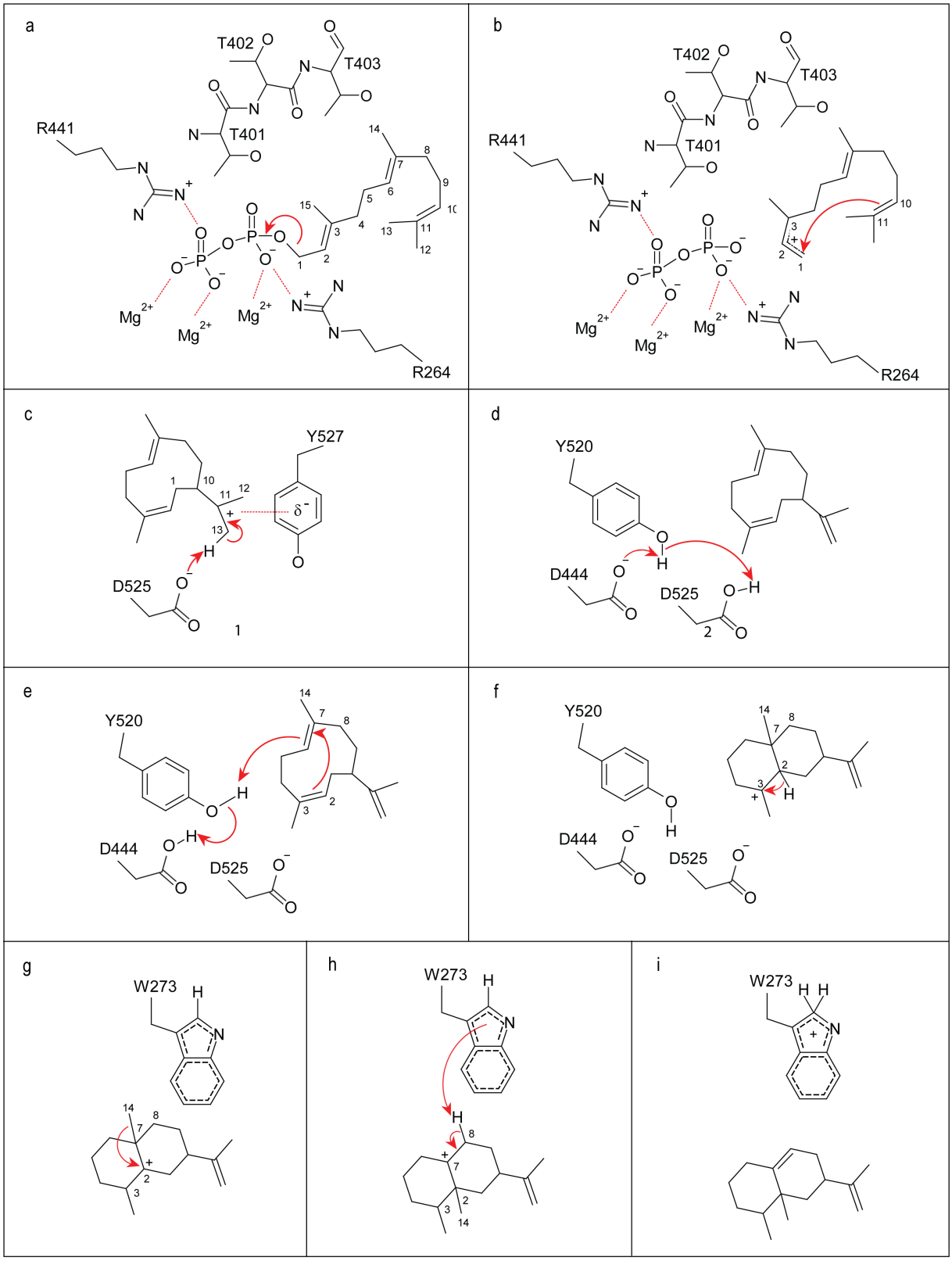


Figure 2.9: 5-epi-aristolochene synthase mechanism.
The mechanism shown is as proposed by Starks *et al.* (1997).

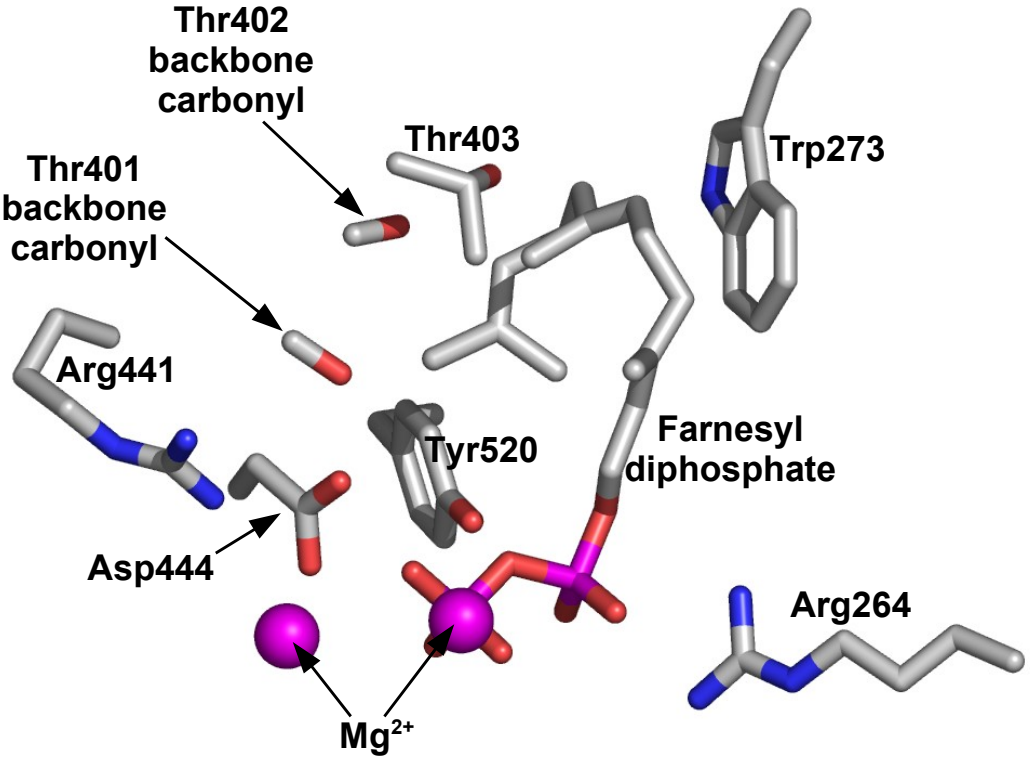


Figure 2.10: Catalytic residues in 5-epi-aristolochene synthase.
These are the catalytic residues according to the mechanism proposed by Starks *et al.* (1997). This mechanism is described in Figure 2.9. This figure is based on the structure with PDB identifier 5eau (Starks *et al.*, 1997). This structure features a bound substrate analogue (trifluoro-farnesyl diphosphate) which differs from the natural substrate (farnesyl diphosphate) only in that there are three fluorine atoms covalently bound to carbon 15; these fluorine atoms are not shown in this figure. Asp525, Tyr527 and one of the catalytic magnesium ions were not located in this structure, although their location was determined in other structures which featured different substrate analogues. This figure was created using Pymol (www.pymol.org).

groups that are not standard residues. This last group includes everything that is marked as a "hetero" group in PDB files; it mostly consists of cofactors, and is referred to as "cofactors" for convenience below, but it also includes modified residue sidechains. Because these cofactors are only listed in the high-annotation literature entries, the cofactor data is drawn from the nonredundant set of high annotation literature entries only.

Sidechains account for 90% of all catalytic residue functions. The catalytic sidechains (Figure 2.11a) are dominated by the charged residues His, Asp, Glu, Arg, and Lys, in that order; these five residues account for 66% of all catalytic sidechains. The charged residues are followed by the uncharged polar residues: Tyr, Cys, Ser, Asn, Thr, Gln, Trp, in that order, accounting for 30% of catalytic sidechains. The remaining nonpolar residues account for a mere 4% of all catalytic sidechains; Phe accounts for more than half of these nonpolar sidechains. Possible reasons for these residue preferences are discussed below in the context of the functions performed by each of these residues.

Catalytically acting backbone amide groups are relatively rare, accounting for 8% of catalytic residue actions (Figure 2.11b). There are fewer catalytic backbone amides than there are of any of the individual charged catalytic sidechains. The backbone amides are dominated by Gly, which accounts for 34% of all catalytic backbone amides. This is followed by Cys, Thr, and Ser, which between them account for another 24% of catalytic backbone amides. Because Gly has only a hydrogen atom as a sidechain, its backbone amide is more accessible than those of other residue types. The greater backbone flexibility provided by Gly might also be important to permitting backbone amide activity.

Backbone carbonyl groups rarely act in a catalytic capacity; they account for just 1% of catalytic residue actions (Figure 2.11c). There is no obvious pattern to the residue types of the handful of cases present here.

Metal ions account for 63% of all cofactors (Figure 2.11d). The most common metal ions are Mg, Zn, Mn, Ca, Fe, and Cu, in that order. The other dominant groups of cofactors is nucleotide derivatives used in redox reactions (NAD, NADP, FAD, FMN) which make up 16% of all cofactors. There are smaller numbers of iron-containing groups (haem and various iron-sulphur groups) which make up 5% of cofactors.

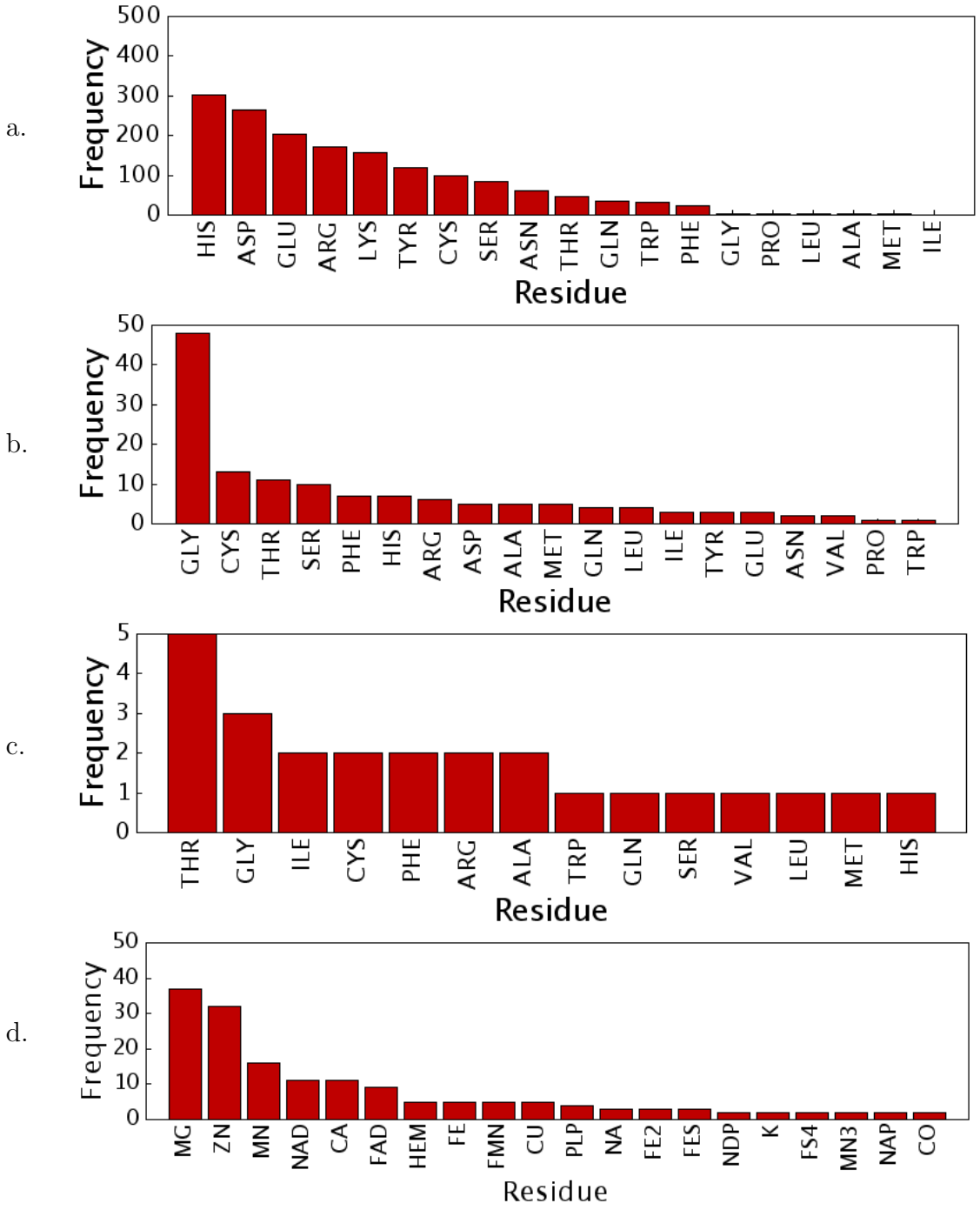


Figure 2.11: Catalytic residue frequencies.
Number of each residue type occurring in the nonredundant set of CSA literature entries. (a) Residues acting via sidechain. (b) Residues acting via their backbone amide groups. (c) Residues acting via their backbone carbonyl groups. (d) Cofactors. Data for cofactors comes from the nonredundant set of high-annotation literature entries. Only the 20 most common cofactors are shown. The cofactors are described using their standard PDB three-letter abbreviations. Many of these correspond to standard abbreviations; those that do not are the following: HEM is haem, FE2 is Fe^{2+}, FES is a 2Fe — 2S iron-sulphur cluster, NDP is NADPH, FS4 is a 4Fe — 4S iron-sulphur cluster, MN3 is Mn^{3+}, NAP is NADP. Note that ions other than FE2 and MN3 have an unspecified oxidation state.

2.3.7 Catalytic residue propensity

The figures for catalytic residue frequency can be normalised according to the frequency with which the residues are found in proteins, to give a value for the propensity of each residue to be catalytic. The propensity ratio for each residue was calculated by dividing the catalytic residue frequency (obtained as described above) by the frequency with which that residue occurred at all positions (including noncatalytic residues) in proteins in the nonredundant set. Each ratio was then expressed as a percentage of the sum of the ratios for all residues. These percentage propensities are shown in Figure 2.12.

The effect of this normalisation on the rank order of catalytic sidechains is mostly slight (Figure 2.12a). Most notably, the dominance of His is emphasised considerably more in the context of its relative rarity, and the rare residue Cys moves up to second place. Trp is also seen as one of the most catalytically active of the uncharged polar residues (exceeded only by Tyr).

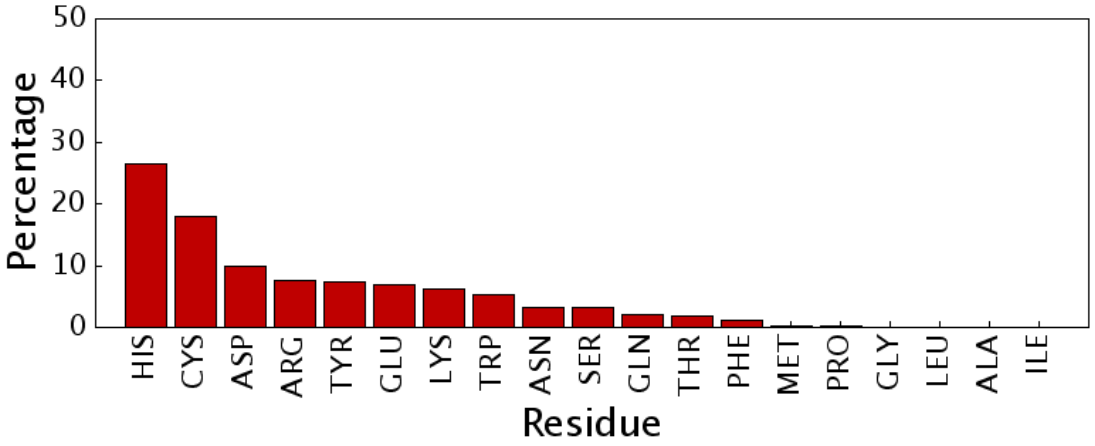
The rarity of Cys results in its having the highest catalytic propensity for backbone amide activity (Figure 2.12b). In second place is Gly, which has a considerably higher backbone amide catalytic propensity than any of the remaining residues, although its dominance is less notable than that observed for the absolute numbers of catalytic backbone amides described above. Thus, whilst the relatively high frequency of occurrence of Gly makes some contribution to its high frequency among catalytic backbone amides, the greater access to the backbone amide allowed by its lack of a sidechain is evidently the main explanation.

There are so few catalytic backbone carbonyls occurring in the nonredundant set of literature entries that the residue propensity figures for them are probably not meaningful (Figure 2.12c).

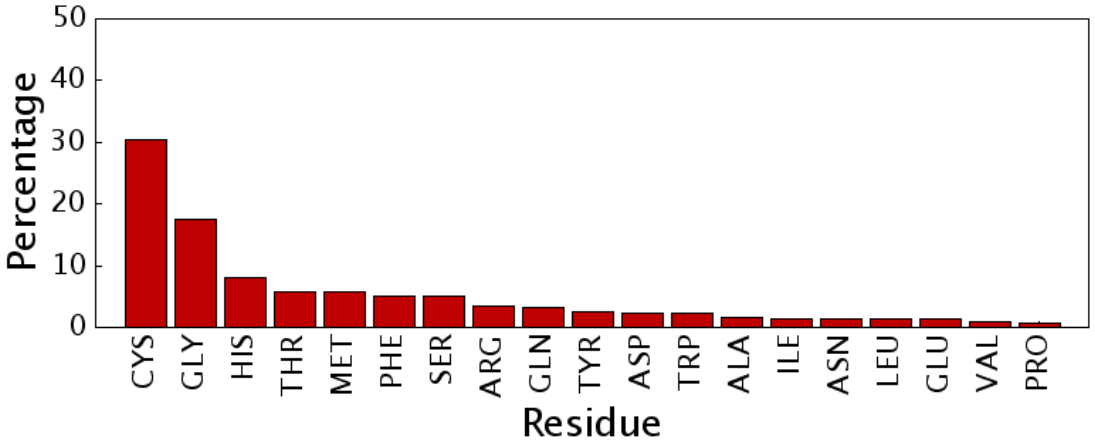
2.3.8 Nonredundant subset of high-annotation entries

A nonredundant subset of the high-annotation entries was obtained. This dataset has 309 members. These entries have considerable functional annotation, and were used to analyse the function and targets of the residues they contain, and the evidence that they are catalytic.

a.



b.



c.

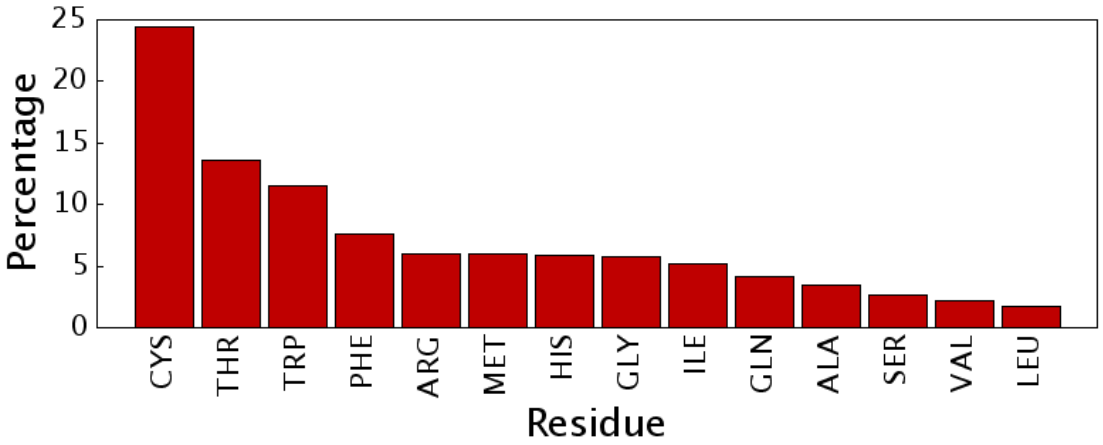


Figure 2.12: Catalytic residue propensities.
The propensity ratio for each residue was calculated by dividing the frequency with which the residue occurred as a catalytic residue (in the nonredundant set of CSA literature entries) by the frequency with which that residue occurred at all positions (including noncatalytic residues) in proteins in the nonredundant set. Each ratio was then expressed as a percentage of the sum of the ratios for all residues. (a) Propensities for residues acting via sidechain. (b) Propensities for residues acting via their backbone amide groups. (c) Propensities for residues acting via their backbone carbonyl groups.

2.3.9 Residue functions

Figure 2.13 shows the frequencies of functions for residues acting via their sidechains; Figure 2.14 shows the frequencies of functions for groups that are not standard residues. These frequencies came from the nonredundant subset of high-annotation entries.

The most frequent residue function for sidechains (Figure 2.13) is electrostatic (49% of all sidechain functions), followed by acid/base (37%). Nucleophilic residues are considerably rarer (9%), and other functional roles only occur in very small numbers, accounting for only 6% of sidechain functions.

The most common function for cofactors (Figure 2.14) is also electrostatic (52% of all cofactor functions). Electron donor/acceptor, hydride transfer and the associated nucleophile and electrophile functions (see below for discussion of these three) occur in relatively greater numbers than for sidechains, because of the role of non-residue groups in the electron and hydride transfer aspects of redox reactions. Relatively few non-residue groups (5%) play an acid/base role.

Residues acting via their backbone groups (either amide or carbonyl) all have an electrostatic function, and for that reason they do not feature in Figure 2.13. There are two cases which are labelled as backbone groups in the database which have different functions, but both are terminal groups rather than peptide bond components: the N-terminal amino group from Thr 2 in PDB entry 1vas acts as a nucleophile, and the C-terminal carboxyl group of Trp 270 in PDB entry 1gxs acts as an acid and base.

2.3.9.1 Residue-function combinations for residues acting via their sidechains

An analysis was also carried out of the combinations of residue and function which occur within the nonredundant subset of high-annotation literature entries. As mentioned above, almost all of the backbone groups served an electrostatic role, so these residue-function combinations were only analysed for residues acting via their sidechains and groups other than standard residues. Table 2.5 shows these combinations for residues acting via their sidechains.

Residues with an electrostatic function are dominated by the charged residues Arg, Asp, Lys, His, and Glu (in that order), with a small number of others, especially Asn.

	Electro-static	Acid/ base	Nucleo-phile	Steric strain	Electron donor/ acceptor	Modified	Electro-phile	Steric hin-drance	Radical forma-tion
HIS	52	104	3	0	4	3	0	0	0
ASP	62	70	9	0	0	0	1	0	0
GLU	42	61	5	0	0	0	0	1	0
LYS	55	30	6	1	0	5	3	0	0
ARG	81	7	0	4	2	0	0	1	0
TYR	25	34	3	4	3	0	0	0	1
CYS	8	10	35	0	0	0	1	0	0
SER	21	11	17	0	0	0	0	1	0
ASN	35	5	1	0	0	0	0	0	0
THR	14	6	1	0	0	0	0	0	0
TRP	16	1	0	2	1	0	0	0	0
GLN	18	0	0	0	0	0	1	0	0
PHE	12	0	0	2	5	0	0	0	0
ILE	3	0	0	0	0	0	0	0	0
PRO	0	0	1	1	0	0	0	0	0
GLY	0	0	0	1	0	0	0	0	1
MET	1	0	0	0	1	0	0	0	0
LEU	1	0	0	1	0	0	0	0	0
VAL	1	0	0	0	0	0	0	0	0
ALA	0	0	0	0	0	0	0	1	0

Table 2.5: Residue-function combinations for sidechain-acting residues.
All values are absolute counts of numbers of residue-function combinations within the nonredundant subset of high-annotation literature entries. Note that if a given residue in a given enzyme has more than one function, then it will be counted more than once within this table. The colour scale is a quick guide to the relative frequency of different combinations. The most common combinations are red, the least common are white.

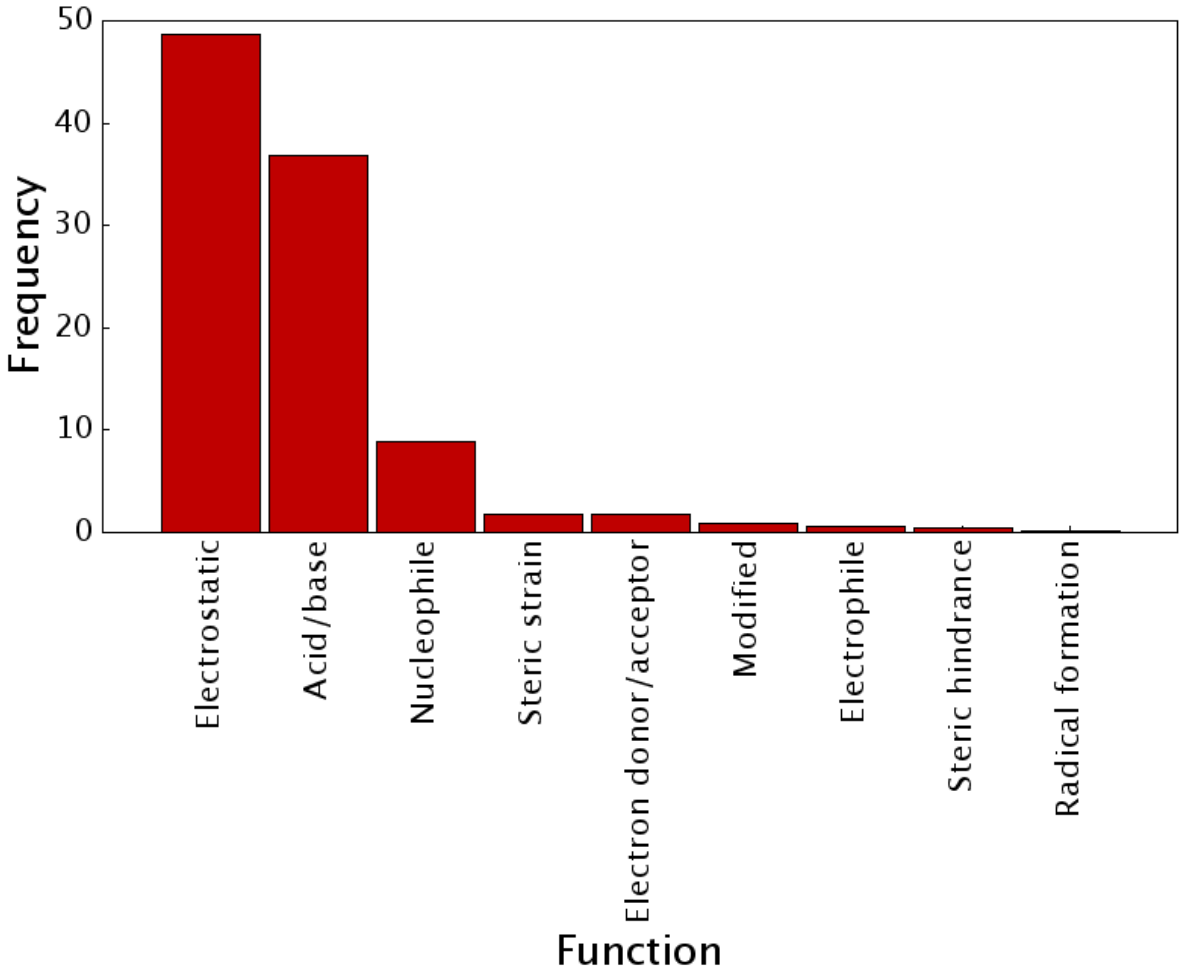


Figure 2.13: Function frequencies for residues acting via their sidechains. These frequencies came from the nonredundant subset of high-annotation entries.

"Electrostatic" activity as defined here may involve hydrogen bonding, or it may be purely electrostatic. Arg is the residue which most commonly serves an electrostatic role; its ability to form a large number of hydrogen bonds may be important here.

Residues acting as acids or bases are dominated by those with pK_a values near to 7: His (pK_a 6.1), then Asp (pK_a 3.9), Glu (pK_a 4.1), Tyr (pK_a 10.1), Lys (pK_a 10.5). Few other residues act in this manner. Despite a pK_a of 8.0, Cys seldom acts in this capacity, perhaps because it is a relatively infrequently occurring residue.

Nucleophiles are dominated by Cys, followed by Ser. Cys predominates because sulphur is a reactive element, and because the low pK_a of Cys means that it can lose its proton more easily than Ser. Asp and Glu do occur as nucleophiles; however, their negative charge is delocalised over the whole of the carbonyl, making them less potent

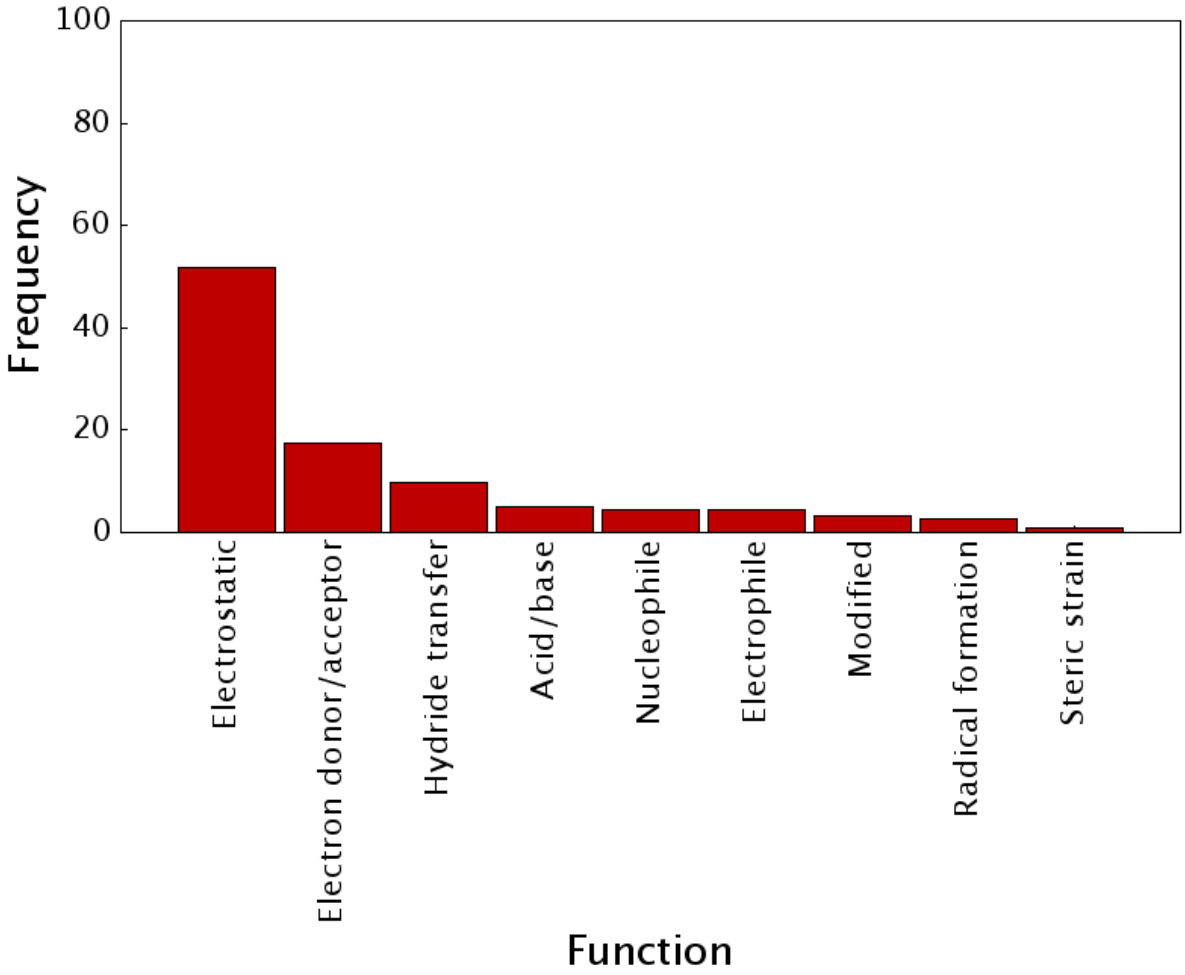


Figure 2.14: Function frequencies for non-residues.
These frequencies came from the nonredundant subset of high-annotation entries.

nucleophiles (and thus less frequently employed) than Ser and Cys, where the charge is concentrated on a single atom. The hydroxyl group of Tyr occasionally serves as a nucleophile, but the charge on the deprotonated hydroxyl oxygen becomes delocalised over the phenyl ring, making it a less potent nucleophile than any of the above residues. Thr is seldom employed as a nucleophile, despite its chemical similarity to Ser. This is probably due to the steric bulk of the extra methyl group that differentiates Thr from Ser.

The main residues which act via steric strain are Arg and Tyr, which are both relatively large. The bulky residues Trp and Phe also sometimes serve this function.

Phe, His and Tyr account for the majority of the electron donor/acceptor residues; presumably this is due to their delocalised ring structures. The distinction between residues acting as an electron tunneling medium and those which are electron donor/acceptors is

not always clear from the experimental evidence and/or the description in the primary literature, so it is possible that some of these cases are misannotated.

The function name "modified" refers to residues which are permanently covalently bound to another group which is catalytically active. The most commonly modified residue is Lys, which is bound to the cofactor pyridoxal phosphate (PLP) in a range of enzymes. The chemistry of these enzymes involves Lys carrying out a nucleophilic attack on PLP and then accepting an electron pair from PLP to form a double bond. Where Lys accepts an electron pair to form a double bond, this Lys is annotated in the CSA as acting as an electrophile, although it is debatable whether this constitutes acting as an electrophile in a strict sense. For this reason Lys is also the most common electrophile residue.

There is no clear pattern to the small number of residues that act via steric hindrance or radical formation.

2.3.9.2 Residue-function combinations for cofactors

Table 2.6 shows the residue-function combinations for cofactors. Almost all electrostatically acting groups are metals, with Mg and Zn being by far the most common, followed at some distance by Mn and Ca. Possible reasons for this are considered in the discussion section below.

The groups which accept and donate electrons are dominated by the classic components of electron transfer chains: Cu, FMN and Fe, followed by FAD and NAD and other Fe-based groups.

Other functions are relatively infrequent. The main groups which transfer hydrides as part of redox reactions are NAD, FAD, FMN and NADP. Acid/base and nucleophile functions are carried out by a smattering of groups, most of which are modified amino acid residues.

Fe and PLP are the main cofactors that act as electrophiles. Fe acts as an electrophile by accepting a ligand. PLP is normally found covalently bound to lysine; its usual role involves it acting as an electrophile to accept a nucleophilic attack from the substrate, which displaces it from the lysine. The residue function "modified" signifies that a cofactor is covalently joined to another group; the main cofactor which this is applied to is PLP.

	Electro-static	Electron donor/ acceptor	Hydride transfer	Acid/ base	Nucleo-phile	Electro-phile	Modified	Radical forma-tion	Steric strain
MG	37	0	0	0	0	0	0	0	0
ZN	32	0	0	0	0	0	0	0	0
MN	16	0	0	0	0	0	0	0	0
FAD	0	3	4	1	1	1	1	1	0
NAD	0	1	8	0	1	1	0	0	0
CA	11	0	0	0	0	0	0	0	0
FE	1	5	0	0	1	3	0	0	0
PLP	1	1	1	0	1	3	2	1	0
FMN	0	3	3	0	0	0	0	0	0
CU	1	5	0	0	0	0	0	0	0
HEM	0	5	0	0	0	0	0	0	0
NA	3	0	0	0	0	0	0	0	1
CSE	0	0	0	1	1	0	1	0	0
PCD	1	0	1	0	0	1	0	0	0
FGL	0	0	0	1	1	0	1	0	0
FES	0	2	0	0	0	0	0	1	0
FE2	0	2	0	0	0	0	0	1	0
CSO	1	0	0	1	1	0	0	0	0
NDP	0	0	2	0	0	0	0	0	0
K	2	0	0	0	0	0	0	0	0

Table 2.6: Residue-function combinations for cofactors.
All values are absolute counts of numbers of cofactor-function combinations within the nonredundant subset of high-annotation literature entries. Note that if a given cofactor in a given enzyme has more than one function, then it will be counted more than once within this table. The colour scale is a quick guide to the relative frequency of different combinations. The most common combinations are red, the least common are white. The names given for cofactors are the three-letter codes used by the PDB database. Only the 20 most common cofactors are shown; these differ slightly from the most common cofactors in Figure 2.11d because residues with multiple functions are over-represented in this table. The cofactors are described using their standard PDB three-letter abbreviations. Many of these correspond to standard abbreviations; those that do not are the following: HEM is haem, FE2 is Fe^{2+}, FES is a 2Fe — 2S iron-sulphur cluster, NDP is NADPH, CSE is selenocysteine, PCD is molybdopterin cytosine dinucleotide, FGL is formylglycine, CSO is hydroxycysteine. Note that ions other than FE2 have an unspecified oxidation state.

Radical formation is too rare to meaningfully discuss the residue types involved.

2.3.10 Residue targets

For high-annotation entries, each catalytic function of a residue has one or more corresponding targets recorded: whether the residue acts upon the substrate, another residue, water, a cofactor, or whether the residue is modified (this generally means that it is covalently bound to a moiety such as PLP which carries out the reaction). The target description "substrate" includes any interaction with the substrate (or any intermediate based on the substrate). The target description "cofactor" includes any interaction with a cofactor (or any intermediate based on a cofactor).

Figure 2.15 shows the number of residue operations acting upon each target type within the nonredundant subset of high-annotation literature entries. Note that if a given residue in a given enzyme has more than one target or more than one function, each target-function combination is treated as a distinct "residue operation" and counted separately. By far the most common target is the substrate (64% of all residue operations), followed by other residues (19% of residue operations).

These results can be better interpreted in the context of Table 2.7, which shows the frequency of different target-function combinations. The most common combination is electrostatic stabilisation of the "substrate". (Note that since these are catalytic residues, this is actually preferential electrostatic stabilisation of a portion of a transition state derived from the substrate; the label "substrate" serves to distinguish this from the other target categories above.) The next most common category is acid/base interaction with the substrate, followed by electrostatic stabilisation of one residue by another. Nucleophilic attacks are almost invariably on the substrate. Interactions between pairs of catalytic residues tend to be electrostatic, with some acid/base activity. Water tends be the target of acid/base activity, although electrostatic interactions are nearly as frequent. Most interactions with cofactors are electrostatic, although electron donor/acceptor functions are also common, reflecting the role frequently played by cofactors in electron transfer chains. The target description "modified" is, by definition, only applied to residues whose function is labelled as "modified".

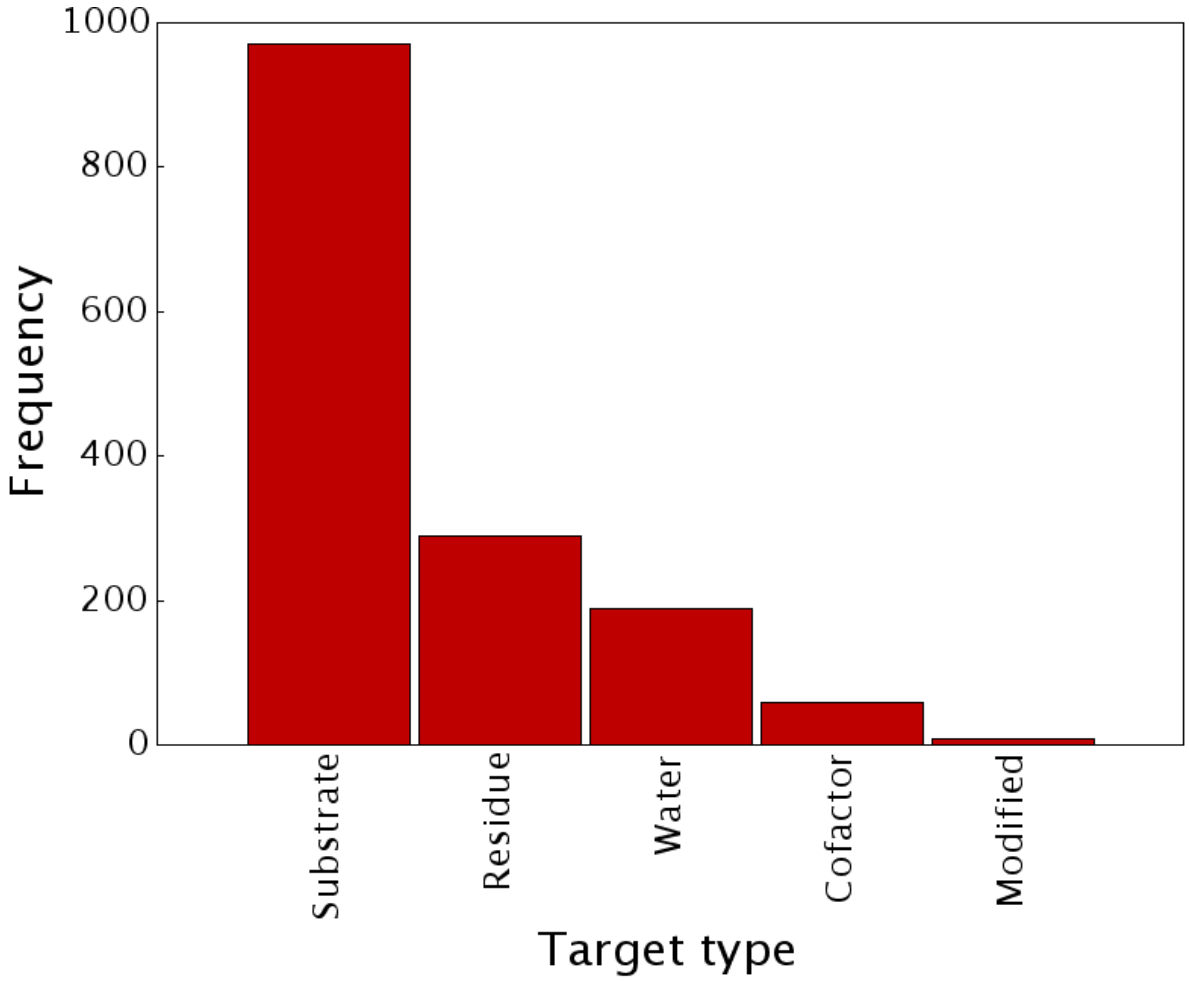


Figure 2.15: Target frequencies.
These frequencies came from the nonredundant subset of high-annotation entries. The target description "substrate" includes any interaction with the substrate (or any intermediate based on the substrate) which involves the formation or breaking of covalent bonds. The target description "cofactor" includes any interaction with a cofactor (or any intermediate based on a cofactor).

2.3.11 Evidence that residues are catalytic

As described above, the high-annotation CSA literature entries include a description of the evidence that a given entry is catalytic. Table 2.8 lists all the possible descriptions that could be applied to a piece of evidence, and provides abbreviations for these descriptions which will be used in the graph, table, and discussion which follow.

Figure 2.16 shows the frequency with which different pieces of evidence are used in the nonredundant subset of high-annotation entries. It's most common for residues to be tagged as catalytic on account of being in the correct position in the structure to

	Electrostatic	Acid/base	Nucleophile	Electron donor/acceptor	Hydride transfer	Steric strain	Electrophile	Modified	Radical formation	Steric hindrance	Electron tunneling medium
Substrate	517	256	101	34	23	14	14	1	7	3	0
Residue	175	82	9	14	0	4	2	3	0	1	0
Water	87	102	0	0	0	0	0	0	0	0	0
Cofactor	24	7	3	18	0	2	1	2	1	0	1
Modified	0	0	0	0	0	0	0	9	0	0	0

Table 2.7: Target-function combinations.
All values are absolute counts of numbers of target-function combinations within the nonredundant subset of high-annotation literature entries. Note that if a given residue in a given enzyme has more than one target or function, then it will be counted more than once within this table. The colour scale is a quick guide to the relative frequency of different combinations. The most common combinations are red, the least common are white.

be catalytic. The other main evidence sources are residue conservation, inference from homologues, and mutagenesis. Apart from mutagenesis, these are relatively weak sources of evidence. However, Table 2.9 (which shows the frequency with which particular combinations of evidence and function) reveals that residue position is used as a source of evidence disproportionately by residues serving an electrostatic role. Residues whose role in the reaction is more direct, involving the breaking and formation of covalent bonds, tend to have stronger evidence: acid/base residues have proportionately more emphasis on mutagenesis evidence, and nucleophiles have proportionately even more emphasis on mutagenesis, as well as some support from structures of covalently bound intermediates. Electron donor/acceptors and electrophiles tend to be cofactors, and as such much of their evidence falls into the category "Ligand is essential for catalysis", which signifies that when the cofactor is absent, catalysis does not occur. Other functions broadly follow the pattern set by electrostatically acting residues.

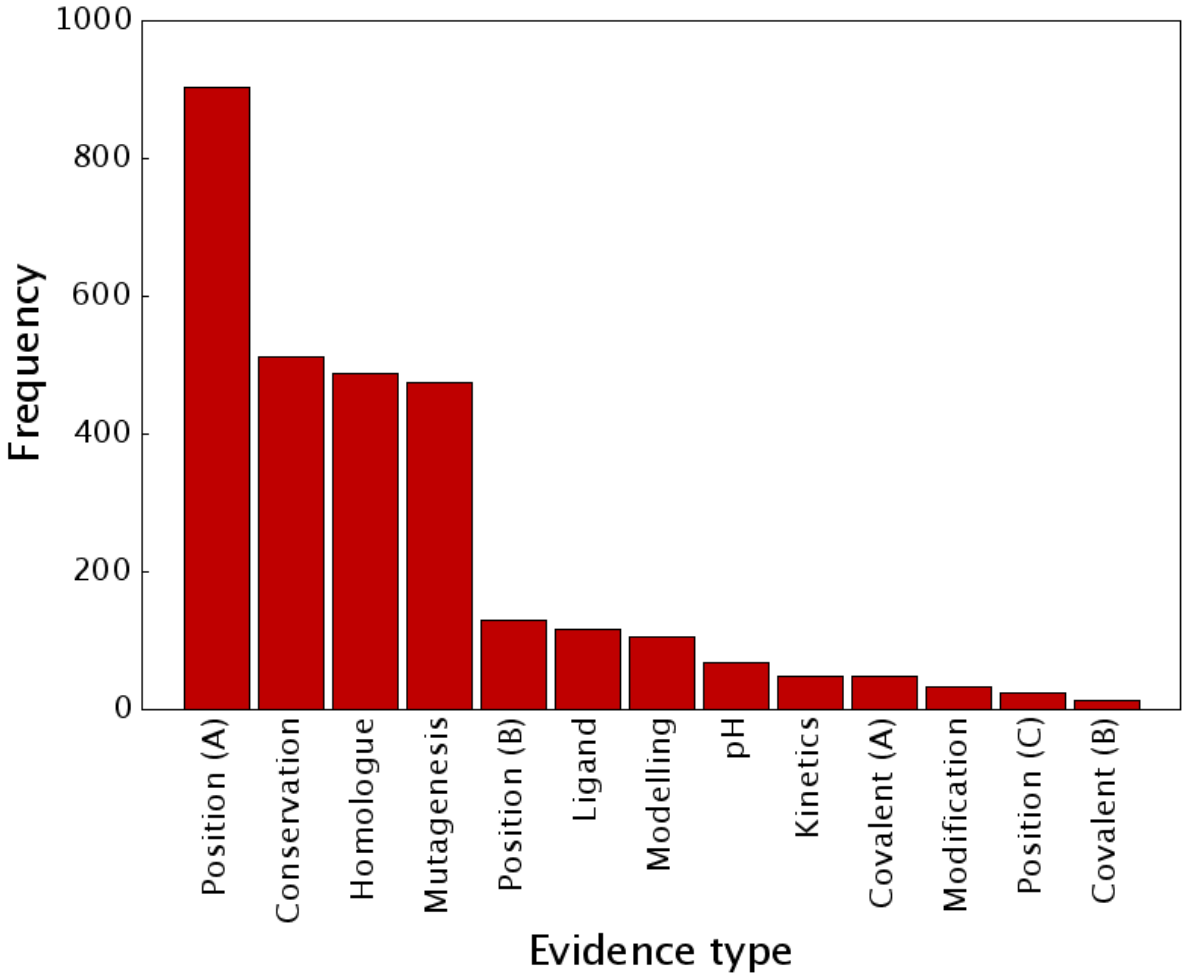


Figure 2.16: Evidence type frequencies.
See Table 2.8 for abbreviations. These frequencies came from the nonredundant subset of high-annotation entries.

Abbreviation	Evidence description
Conservation	Conservation of residue
Covalent (A)	Residue is covalently bound to intermediate, based on structural data
Covalent (B)	Residue is covalently bound to intermediate, based on non-structural data
Homologue	Structural similarity to homologue of known mechanism
Kinetics	Kinetic studies
Ligand	Ligand is essential for catalysis
Modelling	Computer modelling
Modification	Chemical modification of residue
Mutagenesis	Mutagenesis of residue
Position (A)	Residue is positioned appropriately (ligand position known)
Position (B)	Residue is positioned appropriately (ligand position hypothetical)
Position (C)	Residue is positioned appropriately (presence or absence of ligand not recorded)
pH	pH dependence of reaction

Table 2.8: Evidence descriptions and their abbreviations.
A list of all the descriptions that can be applied to individual pieces of evidence in literature CSA entries, and abbreviations for these descriptions.

2.4 Discussion

2.4.1 Growth of the CSA

The total number of entries in the CSA has expanded considerably from version 2.0 to version 2.2, to the point where 85% of enzyme structures in the PDB are covered. The remaining non-annotated enzymes in the PDB include those cases which it was not possible to annotate, because there was insufficient information available in the literature. The number given here for non-annotated enzymes in the PDB also includes some PDB structures which have been assigned an EC code but where the catalytic residues are not present, such as structures of non-catalytic domains from multidomain enzymes.

PDB entries which have many relatives with similar EC numbers have been prioritised for annotation as CSA literature entries; this means that the remaining non-annotated

	Electrostatic	Acid/base	Nucleophile	Electron donor/acceptor	Hydride transfer	Steric strain	Electrophile	Modified	Radical formation	Steric hindrance	Electron tunneling medium
Position (A)	520	267	58	42	16	17	12	6	5	4	1
Conservation	252	179	49	19	1	2	2	3	2	0	0
Mutagenesis	217	190	69	2	0	6	4	2	1	2	1
Homologue	242	152	55	19	4	5	6	3	5	0	0
Ligand	67	3	3	27	14	1	4	2	1	0	0
Modelling	57	37	8	8	0	5	1	2	1	1	0
Position (B)	58	41	8	1	1	2	0	1	0	0	0
pH	31	43	6	0	0	0	0	1	0	0	0
Covalent (A)	7	5	33	4	1	1	5	6	1	1	0
Kinetics	22	20	5	3	1	0	2	1	1	0	0
Modification	15	11	12	1	0	0	1	2	0	0	0
Position (C)	17	6	0	0	2	0	0	0	0	0	0
Covalent (B)	1	0	10	3	0	0	2	1	1	0	0

Table 2.9: Evidence-function combinations.
All values are absolute counts of numbers of evidence-function combinations within the nonredundant subset of high-annotation literature entries. See Table 2.8 for abbreviations. Note that if a given residue in a given enzyme has more than one function, then it will be counted more than once within this table. The colour scale is a quick guide to the relative frequency of different combinations. The most common combinations are red, the least common are white.

PDB entries are disproportionately those with few relatives of similar function. When these cases are added as literature entries, their smaller number of relatives means that they will contribute fewer homologous entries to the CSA. This means that it is possible that the rate of growth of the coverage of the PDB by the CSA will slow down with future updates.

The number of entries in the PDB continues to expand exponentially (Berman *et al.*, 2007); some of the new entries added to the PDB will be automatically annotated as literature entries by the CSA, but others will not. For this reason, there will continue to be scope for expansion of the CSA.

2.4.2 Independent evolution of function, and versatile domains

The majority of third-level EC functions appear to have evolved on more than one occasion. As noted in Chapter one, this convergent evolution of overall function may not correspond to convergent evolution of catalytic sites. The distribution of numbers of independent evolutions of third-level EC numbers is broadly similar to those distributions reported by earlier analyses of smaller datasets, as is the distribution of numbers of third-level EC numbers per domain (Galperin *et al.*, 1998; Thornton *et al.*, 2000; George *et al.*, 2004).

The EC classification contains some third-level functions that are less closely defined than others in terms of their chemistry. Many of the most commonly evolved third-level EC functions are those where the definition of the function is fairly broad, such as EC 2.5.1, representing "Transferases transferring alkyl or aryl groups, other than methyl groups".

Some of the highly versatile domains identified in Table 2.4 are also artefacts of the structure of the EC classification. For example, three of the five third-level EC functions performed by enzymes with catalytic CATH domain 3.40.50.970 correspond to oxidoreductases with pyruvate as a substrate. The α/β hydrolase fold (CATH code 3.40.50.1820) is a case where one fold uses a common set of catalytic residues performing conserved chemical roles (a Ser-His-Asp triad which operates mechanistically in the same way as the α-chymotrypsin example given in Chapter one) to carry out a wide range of overall reactions; this fold includes proteases, lipases, esterases, dehalogenases, peroxidases and epoxide hydrolases (Nardini & Dijkstra, 1999). Other versatile CATH domains are associated with binding a particular cofactor, and use this cofactor to carry out a range of reactions: domain 3.50.50.60 is associated with FAD, domain 3.40.640.10 is associated with PLP, domain 3.40.50.720 is associated with NAD and NADP. Similarly, the P-loop containing homologous superfamily 3.40.50.300 catalyses various reactions involving the hydrolysis of nucleotide triphosphates, and most enzymes with domain 3.90.226.10 are associated with coenzyme A. The single most versatile CATH domain is the aldolase superfamily, 3.20.20.70. This is an instance of the commonly occurring $(\alpha/\beta)_8$ barrel fold. This particular subset of $(\alpha/\beta)_8$ proteins includes a range of enzymes with some com-

mon features to their chemistry, including the formation of Schiff bases with lysine and the stabilisation of enediolate intermediates. It remains uncertain whether this group of functionally diverse proteins genuinely has a common ancestor (Gerlt & Raushel, 2003).

2.4.3 Roles of residues and cofactors

The catalytic residue propensity and residue-function distribution data are similar to those previously reported by Bartlett *et al.* (2002) in their analysis of the 178 enzyme dataset which would later become version 1 of the CSA. The catalytic residue propensities and function frequencies are also similar to those reported by Holliday *et al.* (2007b) in their analysis of the 202 entries in MACiE, a database of catalytic mechanisms.

There has been less study of the frequencies with which different cofactors occur, and how often they play specific roles. Metal ions account for the majority of cofactors recorded in the CSA. The main role played by metals is electrostatic; metals can provide a higher charge density than any of the charged residue types. The chief electrostatically acting metals are (in declining order of frequency) Mg, Zn, Mn, and Ca. These positively charged ions often help stabilise the negative charge that accumulates on intermediates in the many enzymatic reactions that proceed via a nucleophilic substitution mechanism. These four most common metals are all relatively small divalent cations, and thus provide a higher charge density than easily available monovalent cations such as Na^+ or K^+. (Trivalent cations such as Al^{3+} tend to stabilise unproductive reaction intermediates, and exchange ligands in their first coordination shell far more slowly than divalent cations (Frausto da Silva & Williams, 2001).) Mg, Zn, Mn and Ca are also all metals with a single most stable oxidation state. Ca^{2+} and Mg^{2+} both have full outer shells; Zn^{2+} has a full d-orbital, and Mn^{2+} has a relatively stable half-full d-orbital. This makes them less able to change oxidation state than other bioavailable metals, like the transition metals Fe and Cu. This stability may be the reason that these metals are favoured for playing electrostatic roles: they are unlikely to engage in unwanted redox reactions with the substrate (Lippard & Berg, 1994).

Mg^{2+} is more frequently employed than other divalent cations. This is due to its high abundance in the cell (Wolf & Cittadini, 2003).

Zn^{2+} has a number of properties that account for the fact that it frequently plays a catalytic role in enzymes (Frausto da Silva & Williams, 2001). Zn^{2+} is a more powerful Lewis acid than any divalent cation other than Cu^{2+}: it has a strong ability to accept electrons, suiting it for stabilising negative charges. Zn^{2+} also exchanges ligands relatively swiftly.

It is possible that the relatively high numbers for catalytically active Mn^{2+} ions are misleading. For those enzymes recorded in the CSA as having catalytic active Mn^{2+}, it is often unclear whether the biologically relevant ion is Mn^{2+} or another divalent cation. Mn also occurs in Mn^{4+} and Mn^{3+} forms. Mn plays redox roles in lignin peroxidases, superoxide dismutases and photosystem II (Frausto da Silva & Williams, 2001). However, all the Mn ions occurring in the set of enzymes analysed in this chapter played electrostatic roles.

As described above, Cu and Fe are stable in a number of oxidation states, which makes them able to gain and lose electrons easily. For this reason, they tend to serve electron transfer and electrophile roles.

Chapter 3

Using structural templates to recognise catalytic sites and explore their evolution

3.1 Introduction

Three-dimensional patterns of catalytic residues can be strikingly conserved between distantly related enzymes. They can also be very similar in unrelated enzymes of similar function. The classic example of both these phenomena is the Ser-His-Asp catalytic triads found in a variety of hydrolases (Wallace *et al.*, 1996). Understanding the extent of this structural conservation can provide a basis for using data concerning catalytic sites for enzyme function prediction.

There is an increased need for structure-based function prediction methods now that structural genomics projects are determining the structures of many proteins whose function is unknown (Brenner, 2001). One method for predicting the functions of these structures is to search these structures for three-dimensional residue patterns resembling known catalytic sites. How well does this method of function prediction work? This *technical* question is dependent on the *scientific* questions of how much structural variation exists between the catalytic sites of enzymes of similar function, and what the reasons are for the variation that is present. As described in Chapter one, this is a question that has

previously only been examined for a few enzyme families.

Having a *library* of structural templates representing catalytic sites helps address the scientific question of the structural variability of catalytic sites; it helps address the technical problem of how useful catalytic site templates are; and it provides a resource for function prediction. As described in Chapter one, various libraries of structural templates exist. However, the only publicly available bank of catalytic site templates yet constructed using information taken directly from the literature is the limited set in the PROCAT database (Wallace *et al.*, 1997). In the work described in this chapter, information in the CSA was used to create automatically a library of structural templates for the template matching program Jess (Barker & Thornton, 2003).

Each literature entry in the CSA has an associated homologous family (**CSA family**) of structures in the PDB, identified using PSI-BLAST (Altschul *et al.*, 1997). The structural template library was used to analyse how catalytic site structure varies within CSA families. The ability of these templates to predict protein function was assessed using two types of analysis: a **family analysis** (to examine how well a template representing a single CSA family discriminates family matches from random ones); and a **library analysis** (examining whether the right template comes out as the top match when one structure is run against the whole template library).

These analyses show that structural templates can discriminate well between family members and random matches, largely because there is a high level of structural conservation of catalytic residues in most CSA families. This library should prove a useful resource for functional site recognition in structures produced by structural genomics initiatives. Additionally, it will aid analyses of convergent evolution of active sites and the relationship between active site geometry and catalytic ability.

The methodology of the family analysis is derived from an analysis of structural templates carried out by Gail Bartlett in the Thornton research group and described in her PhD thesis (Bartlett, 2004). The family analysis described here was carried out independently by the present author using a new set of computer programmes written by himself, differs from the earlier analysis in a number of details, and uses a more extensive dataset. The analysis of statistical measures, and the library analysis, are wholly original to this thesis.

3.2 Results

3.2.1 Dataset

Templates were created from 147 non-redundant CSA families derived from version 2.0 of the CSA (Table 3.1). These families had a total of 2392 members. This set of families is considerably smaller than the 514 families found in the whole of version 2.0 of the CSA. This is because the families used to create templates had to meet several requirements which were not met by all families in the CSA: they had to be non-redundant with other families; they had to include least three catalytic residues; they have to have at least two member structures based on X-ray crystallography with a resolution better than 2.5 Å; they also had to be suitable for the family-specific calculation of template match significance described below. These requirements are fully described in the Methods section of this chapter.

How well this dataset covers enzyme functional space can be quantified by the proportion of EC numbers it covers. The family members covered 31% of all EC numbers at the

12as	1a05	1a4l	1a50	1a65	1afr	1afw	1aj0	1aj8	1ald
1apx	1apy	1aq0	1aq2	1asy	1at1	1auo	1ay4	1azy	1b2m
1b6b	1b6g	1b73	1b8f	1bbs	1bcr	1bd3	1bhg	1bol	1brm
1bs0	1btl	1bwp	1bzy	1c4x	1cb8	1cbg	1cd5	1cde	1cel
1chd	1cns	1coy	1cqq	1ctn	1cwy	1czf	1d8c	1d8h	1daa
1dae	1dco	1de6	1dhp	1dio	1diz	1djl	1dl2	1dnk	1dnp
1do8	1ecf	1ecl	1eh5	1eh6	1eix	1elq	1els	1euy	1exp
1eyi	1eyp	1f6d	1f75	1f8m	1f8x	1fnb	1foh	1fr8	1ga8
1gal	1gim	1gpa	1gpr	1h7o	1hrk	1iph	1kas	1kim	1kra
1lam	1ldm	1luc	1mas	1mbb	1mhl	1mpy	1mrq	1ni4	1nsp
1og1	1onr	1opm	1oya	1pfk	1pfq	1pj5	1pjb	1pnt	1pow
1pvd	1qe3	1qfe	1qfm	1qgx	1qh9	1qk2	1qpr	1qrz	1rbn
1rpx	1sca	1ses	1smn	1tah	1uae	1uag	1ula	1xtc	2ayh
2dln	2gsa	2hdh	2nac	2npx	2pda	2pfl	2pgd	2pth	2tmd
2tps	2xis	3csm	3nos	5cpa	5fit	7odc			

Table 3.1: PDB entries in dataset.
PDB codes for the literature entries of the 147 non-redundant CSA families used in the study.

third level, and 5% of all EC numbers at the fourth, most detailed, level of classification. They covered 38% of all third-level EC numbers for which a protein structure is available (18% of all fourth-level EC numbers).

A Jess template consists of a description of a series of atoms, in terms of constraints on those atoms. A match to a template is a group of atoms in the query file which correspond to the atom descriptions in the template; in other words, a set of atoms which satisfy the constraints.

For the templates used in the analysis described in this chapter, the constraints used to describe each atom were:

- Atom name as described in the PDB file.
- Residue name as specified in the PDB file.
- Whether the atom must be in the same residue as other specific atoms described in the template. Atoms are required by the template to come from the same residue as one another if they are in the same residue as one another in the structure from which the template is derived.
- Distances to other atoms specified in the template. These inter-atom distances must be the same as those between the atoms in the structure from which the template is derived, to within a margin of 6 Å. For the first atom in the template, no such inter-atom distance constraints are specified. For the second atom, the distance to the first atom is specified. For the third atom, a distance is specified to each of the first two atoms, and so on, such that every atom has an inter-atom distance specified for each atom which has previously been described in the template.

Templates have an XML-based format, described in Figure 3.1. A "template" element describes a single template; an "atom" element describes a single atom; a "select" element contains the constraints for that atom. The constraints which are described within this element do not themselves follow an XML format, but are composed of a set of functions. For example, the "isAtomNamed(XXXX)" function requires the atom name to be the argument XXXX. Further functions are described in Figure 3.1.

In most cases the atom names and residue names required by the template are those which occur in the PDB file. Exceptions are described in the Methods section of this chapter. The sequence position and ordering of the residues in the template is not specified; the residues matched by the template can have a different sequence ordering from the equivalent residues in the structure which formed the basis of the template.

Two different types of template were created for each CSA family member. These two template types used different atom subsets to represent the position of each residue. The first type of template represented residues in terms of the positions of their C_α and C_β atoms, and was therefore a reflection of backbone orientation. The second type represented each residue using three functional atoms, and was a reflection of the orientation of the ends of the residue sidechains. This allowed an investigation into which atom subset resulted in the most effective templates. An example is shown in Figure 3.2.

3.2.2 Structural variation of catalytic sites

The template library was used to analyse the degree to which catalytic sites vary within CSA families. For each family, a representative template was selected. The degree of difference in the geometry of catalytic sites for each family member was quantified as the RMSD between their template and the representative template for the family. One potential cause of difference between catalytic sites is evolutionary divergence. The overall evolutionary divergence of each family member was quantified as their percentage sequence identity to the representative member for that family. In order to assess the contribution of evolutionary divergence to differences between catalytic sites, an examination was carried out of the relationship between the RMSD of a family member and its percentage sequence identity. Because the size of a template can affect the significance of matches to it with a given level of RMSD, analysis was carried out separately for templates consisting of three, four, and five residues. The results are shown in Table 3.2.

3.2.2.1 Functional atom templates against C_α/C_β templates

The mean RMSD for functional atom templates was higher than the mean RMSD for C_α/C_β templates at any given template size. This may simply be a consequence of the

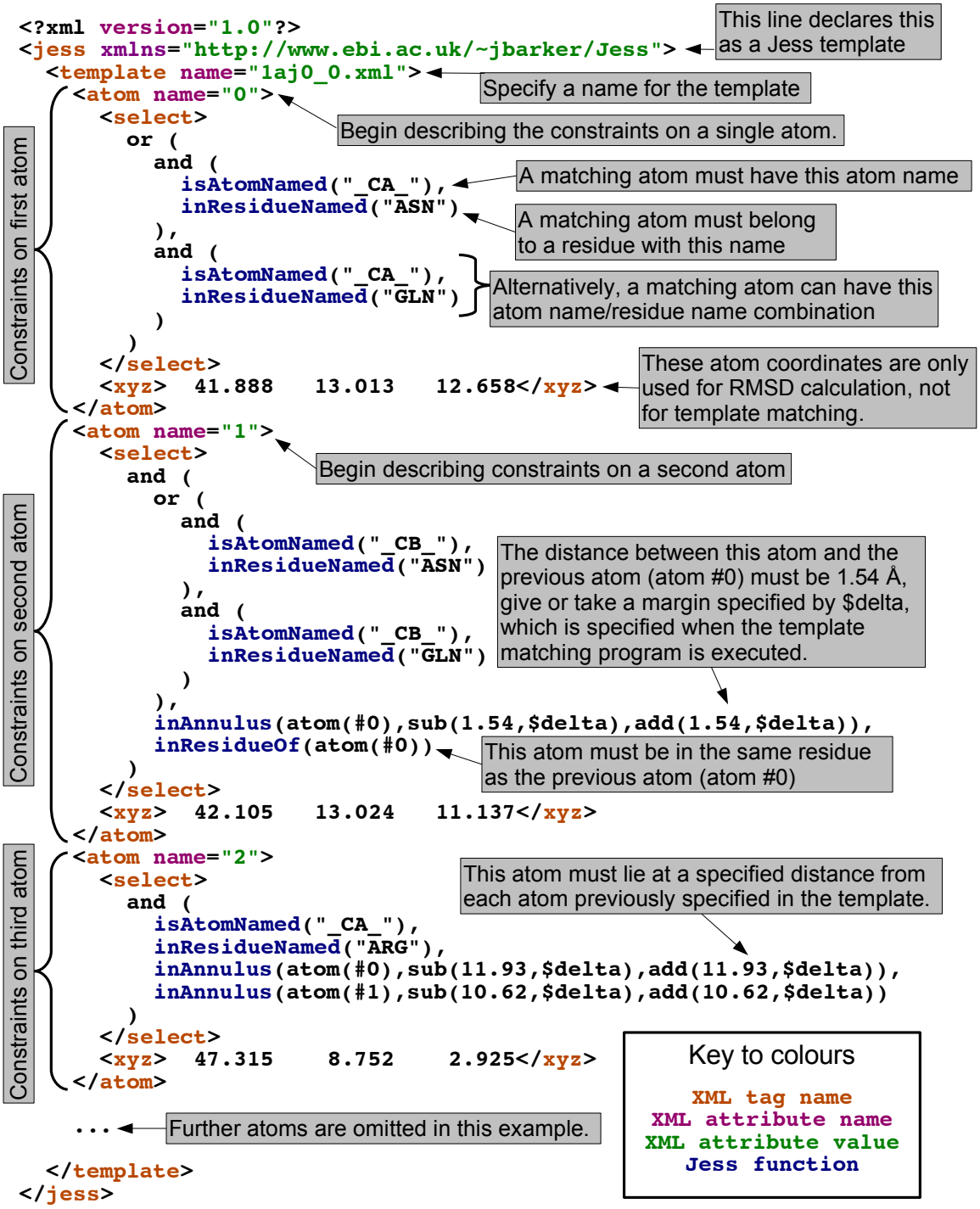


Figure 3.1: Structural template format.
Note that the "name" attribute of the atom tag is an identifier to be used elsewhere in the template, and should not be confused with the atom name given in the PDB file (e.g. " CA "). Note also that this is a different template to the one used in Figure 3.2.

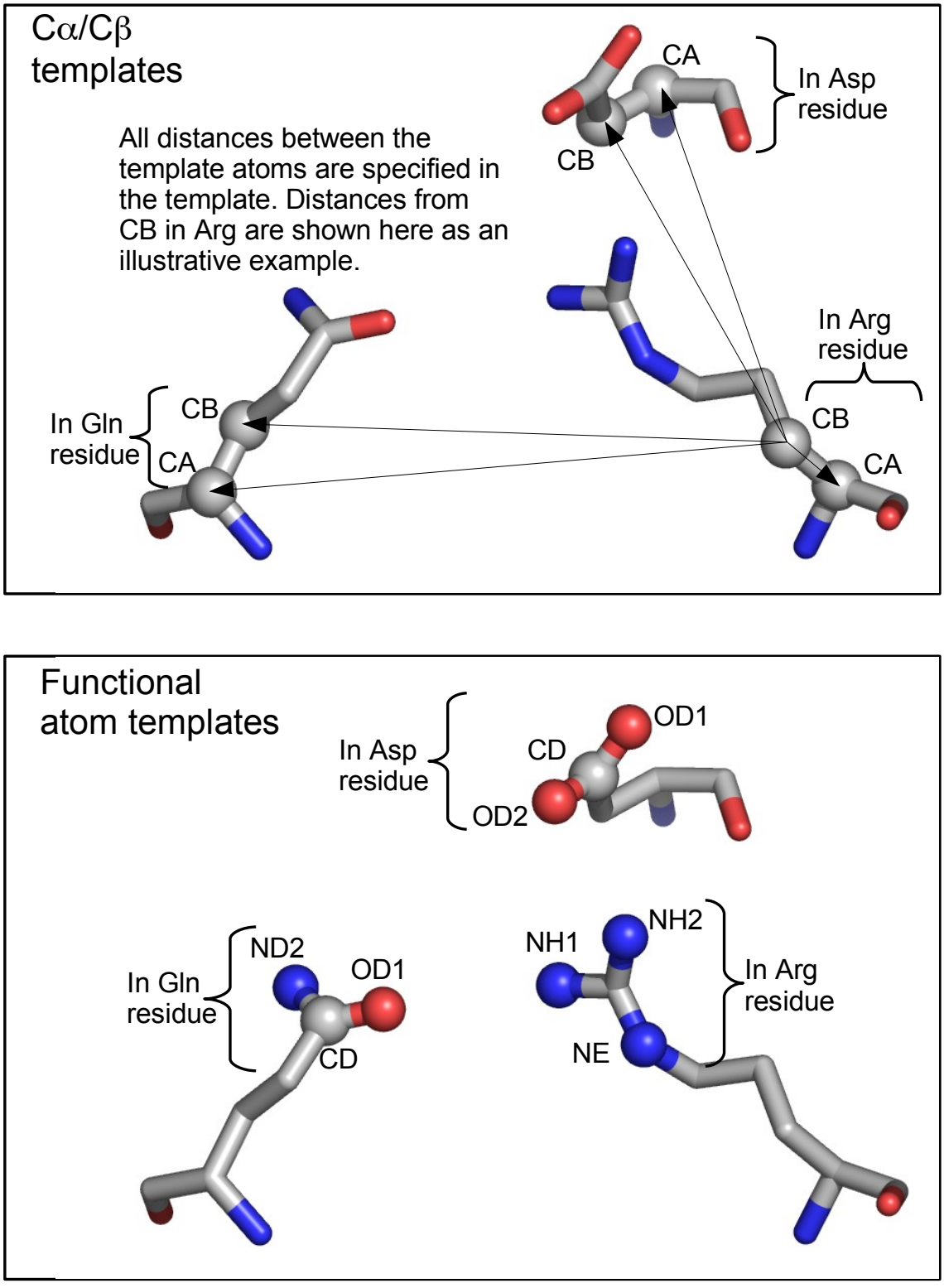


Figure 3.2: Structural template depiction.
The atom and residue names specified for each residue are shown. Note that this is a different template to the one used in Figure 3.1. The example shown was created from PDB entry 12as ((Nakatsu *et al.*, 1998)). This figure was created using Pymol (www.pymol.org).

	Template type	3-residue templates	4-residue templates	5-residue templates
Number of families	-	95	49	3
Number of templates	Functional atoms	1455	463	11
	C_α/C_β	1455	468	14
Mean RMSD (standard deviation)	Functional atoms	0.53 (0.78)	0.56 (0.71)	2.19 (0.48)
	C_α/C_β	0.30 (0.61)	0.29 (0.53)	1.02 (0.56)
Correlation of RMSD and sequence ID (significance)	Functional atoms	-0.31 (1.62×10^{-12})	-0.04 (0.18)	0.66 (0.01)
	C_α/C_β	-0.44 (1.62×10^{-12})	-0.15 (5.69×10^{-4})	-0.19 (0.26)

Table 3.2: Catalytic site structural similarity and its relationship to sequence similarity. The number of templates does not include the representative templates. See Figures 3.6, 3.7 and 3.8 for corresponding distributions. Correlation was calculated using Spearman's rank correlation.

larger size of the functional atom templates, or it may in part reflect greater structural variability in catalytic residue functional atoms than in C_α and C_β atoms; this is discussed further below.

3.2.2.2 RMSD variation between family members

For three residue templates (Figures 3.3 and 3.6) and four residue templates (Figures 3.4 and 3.7), catalytic site structure was highly conserved. RMSD was below 1 Å for 95% of C_α/C_β templates for these template sizes. Even when there was very little sequence similarity, most structures were still conserved: RMSD was below 1 Å for 80% of three and four residue C_α/C_β templates where the sequence identity was 20% or less. The remaining minority of templates had RMSDs between 1 Å and 5 Å. Such divergent templates were relatively abundant below 40% sequence identity. However, they were also observed in

large numbers above 95% sequence identity: high structural divergence can occur with little or no change in sequence. This was the result of cases where the catalytic site of an enzyme can change in conformation, generally in response to binding of a substrate, inhibitor, cofactor or allosteric effector. There were few templates with sequence identity between 40% and 95% that had RMSD > 1 Å; this is probably just a reflection of the fact that there were few templates with sequence identity between 40% and 95% overall. Only a small number of outliers had RMSDs above 5 Å. There were too few five-residue templates (Figures 3.5 and 3.8) to comment meaningfully on the distribution of their RMSDs.

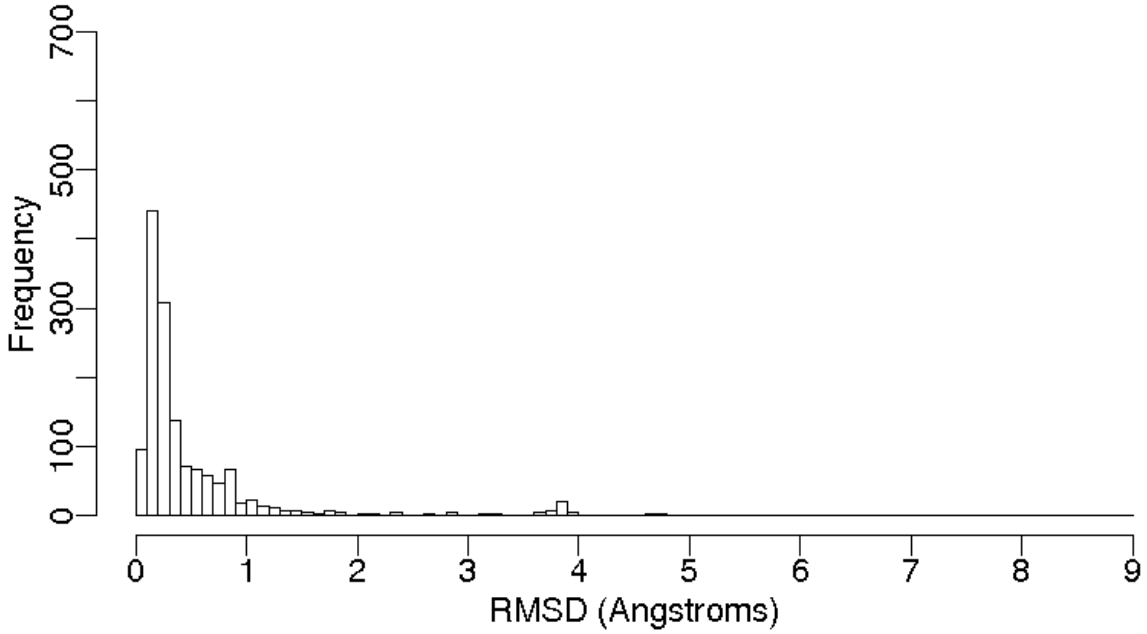
3.2.2.3 Correlation between RMSD and sequence similarity

There was a statistically significant correlation at the 0.05 level between sequence identity and RMSD of protein pairs for the three-residue templates, and some other cases, but this correlation was always very weak. It is possible that aggregating together all data for templates of a given size might have obscured patterns visible in individual families. Therefore, the correlation between sequence identity and RMSD was calculated for each individual CSA family. There was a statistically significant correlation at the 0.05 level for only a minority of those CSA families for which correlation significance could be calculated: 39% of families for C_α/C_β templates and 35% for functional atom templates. Evolutionary divergence appears not to be a major cause of structural differences between catalytic sites.

These statistics can be set in context by considering a few individual CSA families. Figure 3.9 shows nine example families, three from each template size. C_α/C_β templates are used in these examples because of the greater structural conservation of C_α and C_β atoms compared to functional atoms. The three-residue examples will be discussed in more detail here, and their structures are shown in Figure 3.10.

The aldolase CSA family (reaction shown in Figure 3.11) is based on the literature entry human aldolase A, with PDB code 1ald (Gamblin *et al.*, 1991). Members of this family have three residues that act catalytically: an aspartate, a glutamate and a lysine. All catalytic sites in the family are very similar to one another (Figure 3.10a); the differences between them may well simply represent experimental uncertainty in atom

a. Three residue functional atom templates.



b. Three residue C_α/C_β templates.

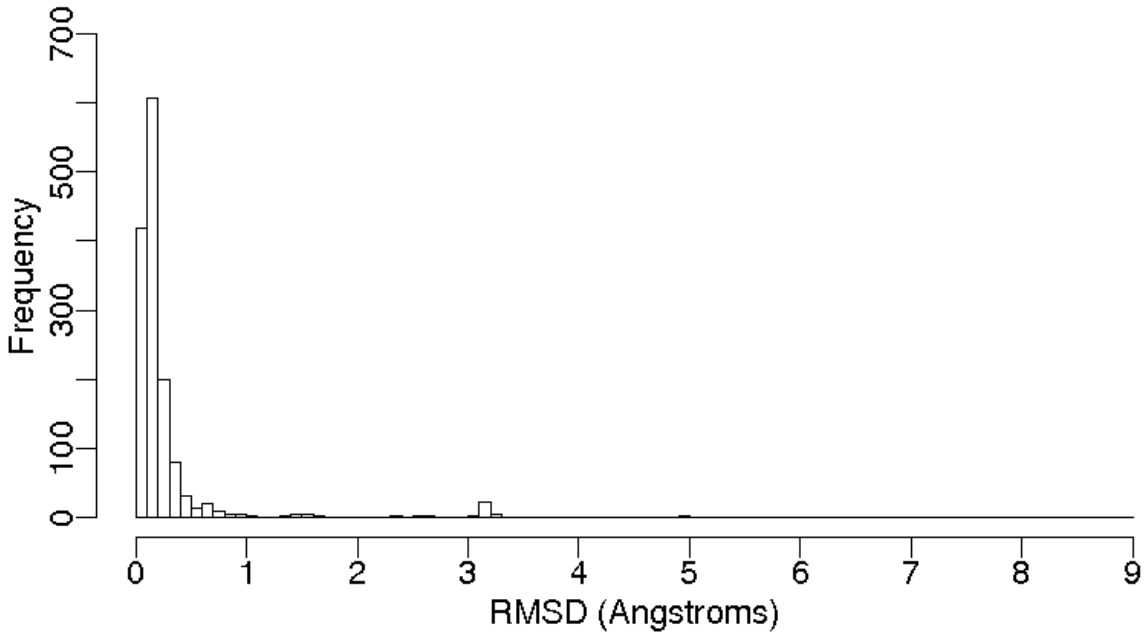
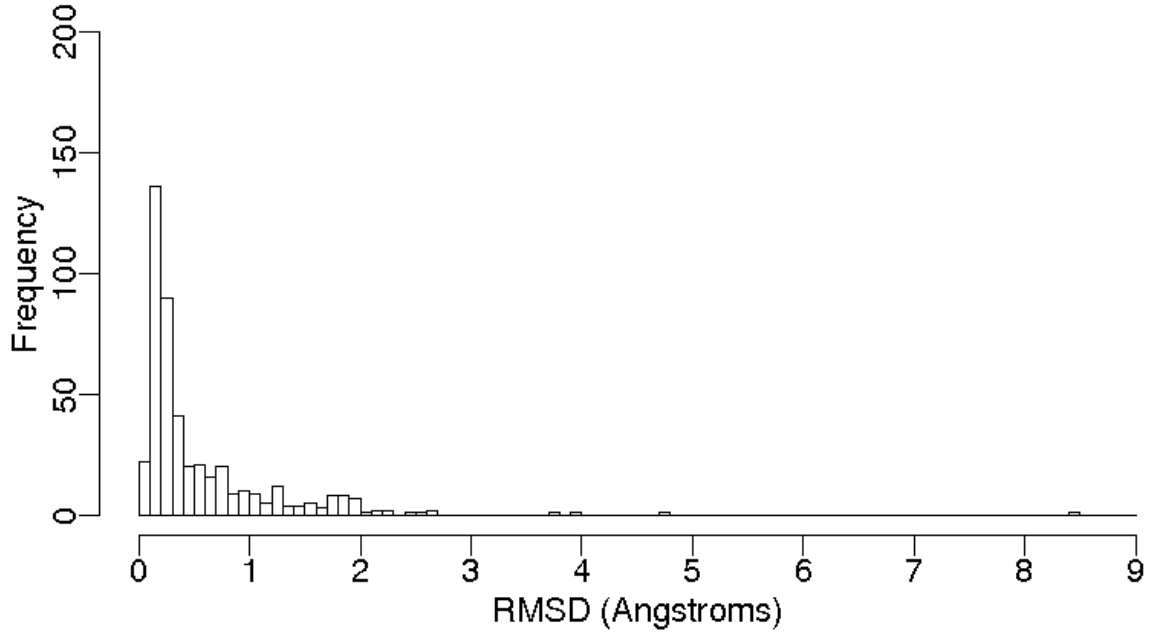


Figure 3.3: Structural variation of catalytic sites consisting of three residues. Each data point counted is a comparison between the representative entry for a given CSA family and another one of that CSA family's members. These graphs include all such comparisons within all CSA families in this analysis where the templates consisted of three residues.

a. Four residue functional atom templates.



b. Four residue C_α/C_β templates.

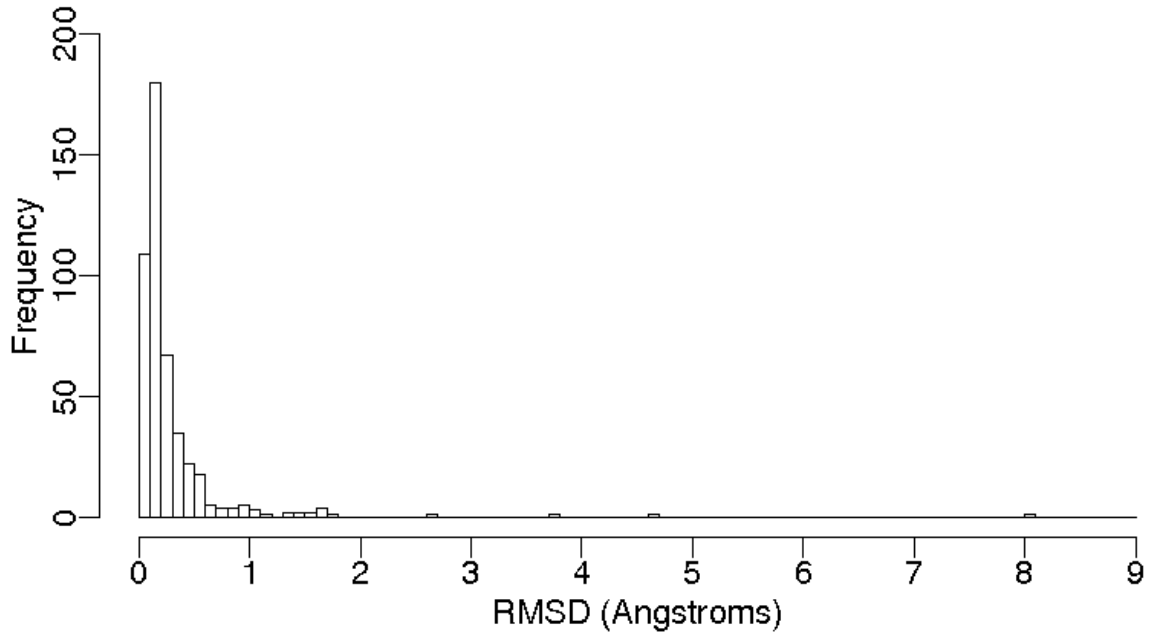
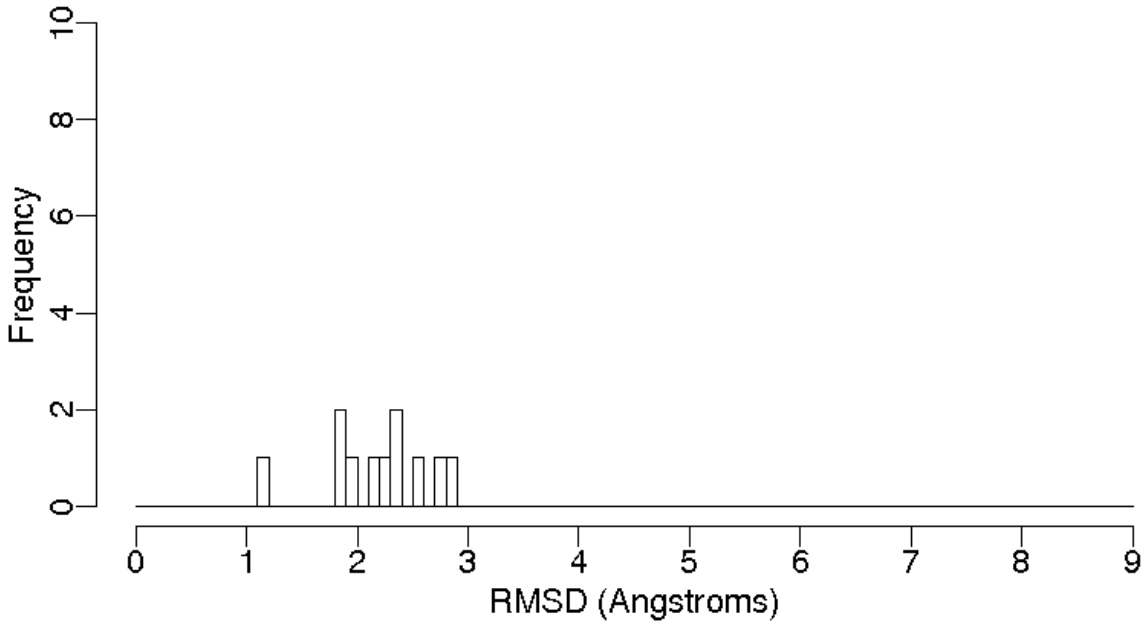


Figure 3.4: Structural variation of catalytic sites consisting of four residues. Each data point counted is a comparison between the representative entry for a given CSA family and another one of that CSA family's members. These graphs include all such comparisons within all CSA families in this analysis where the templates consisted of four residues.

a. Five residue functional atom templates.



b. Five residue C_α/C_β templates.

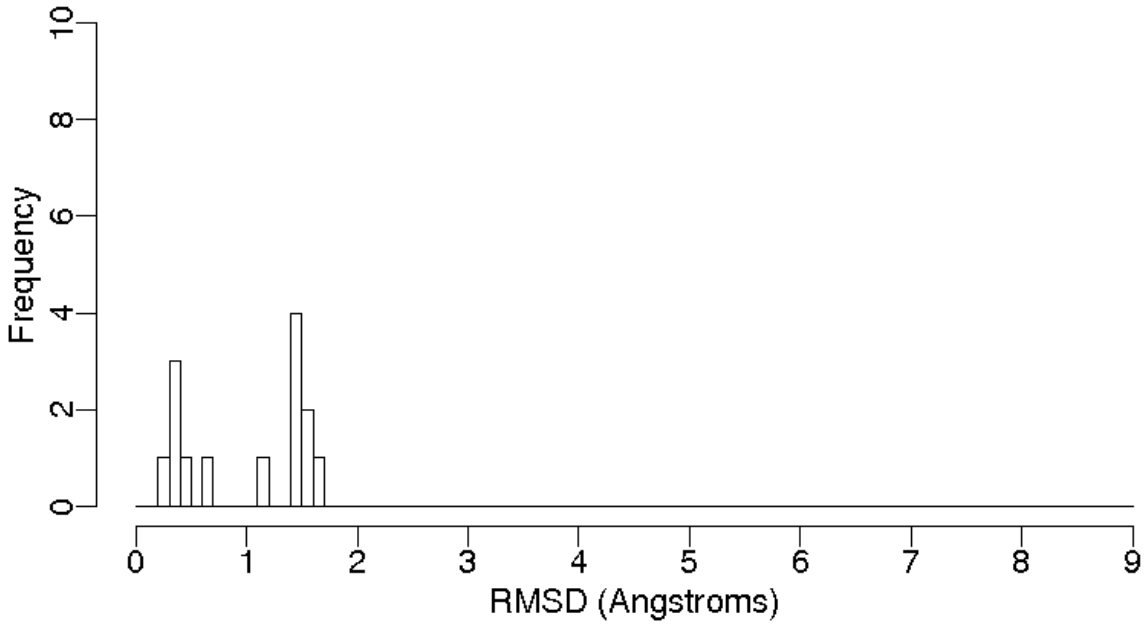


Figure 3.5: Structural variation of catalytic sites consisting of five residues. Each data point counted is a comparison between the representative entry for a given CSA family and another one of that CSA family's members. These graphs include all such comparisons within all CSA families in this analysis where the templates consisted of five residues.

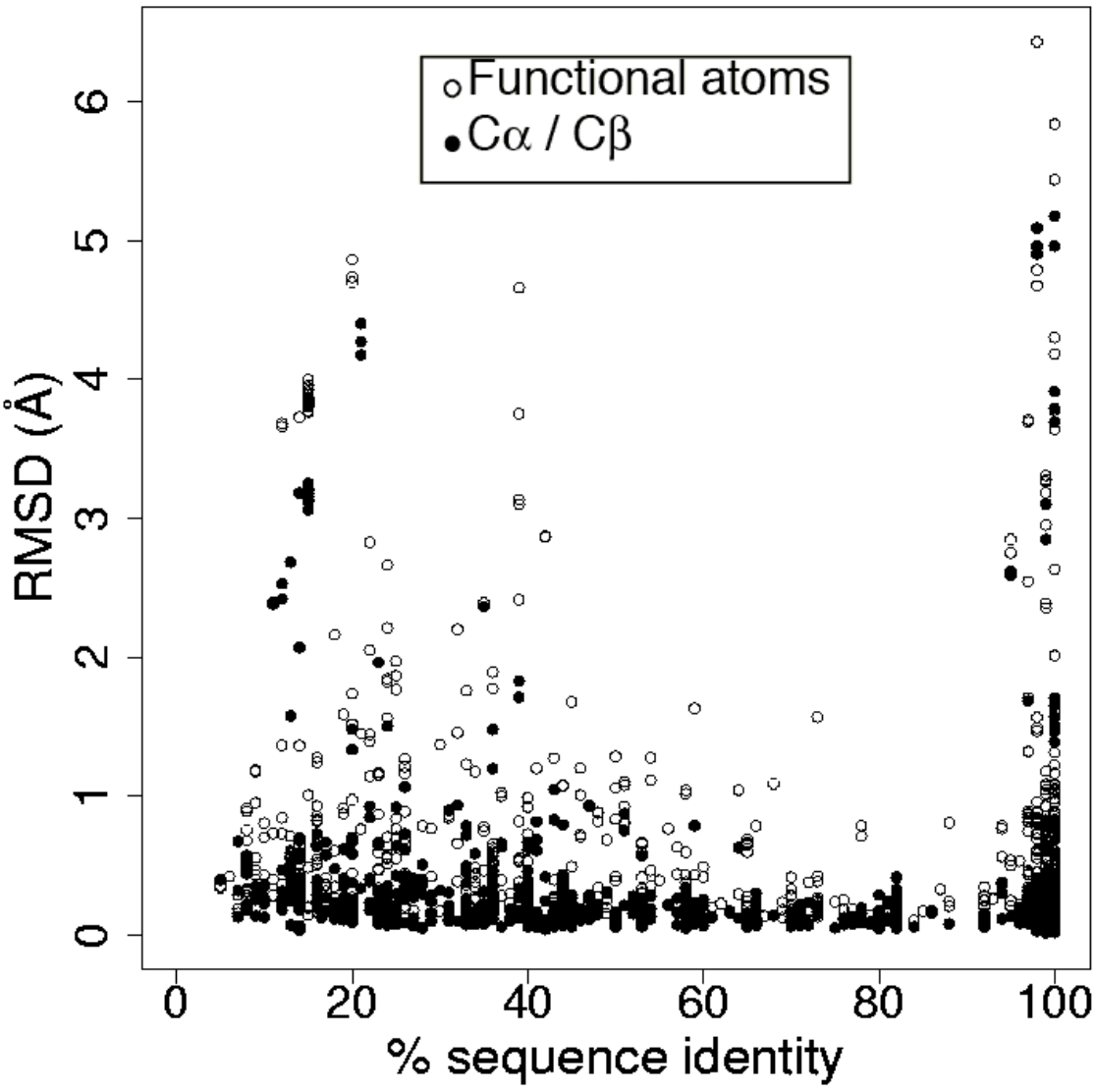


Figure 3.6: Relationship between sequence similarity and catalytic site structure similarity for three-residue templates.
Each point on the scatter plot corresponds to a single comparison between the representative entry for a given CSA family and another one of that CSA family's members.

coordinates. This high similarity exists in spite of some variation in sequence: all but one of the sequence identities are 70% or less, the lowest being 44%. Because the differences in structure are trivial, there is no relationship between sequence identity and RMSD for this family (Figure 3.9a).

The catechol 2,3-dioxygenase family (reactions shown in Figure 3.12) is based on the literature entry *Pseudomonas putida* catechol 2,3-dioxygenase, with PDB code 1mpy (Kita

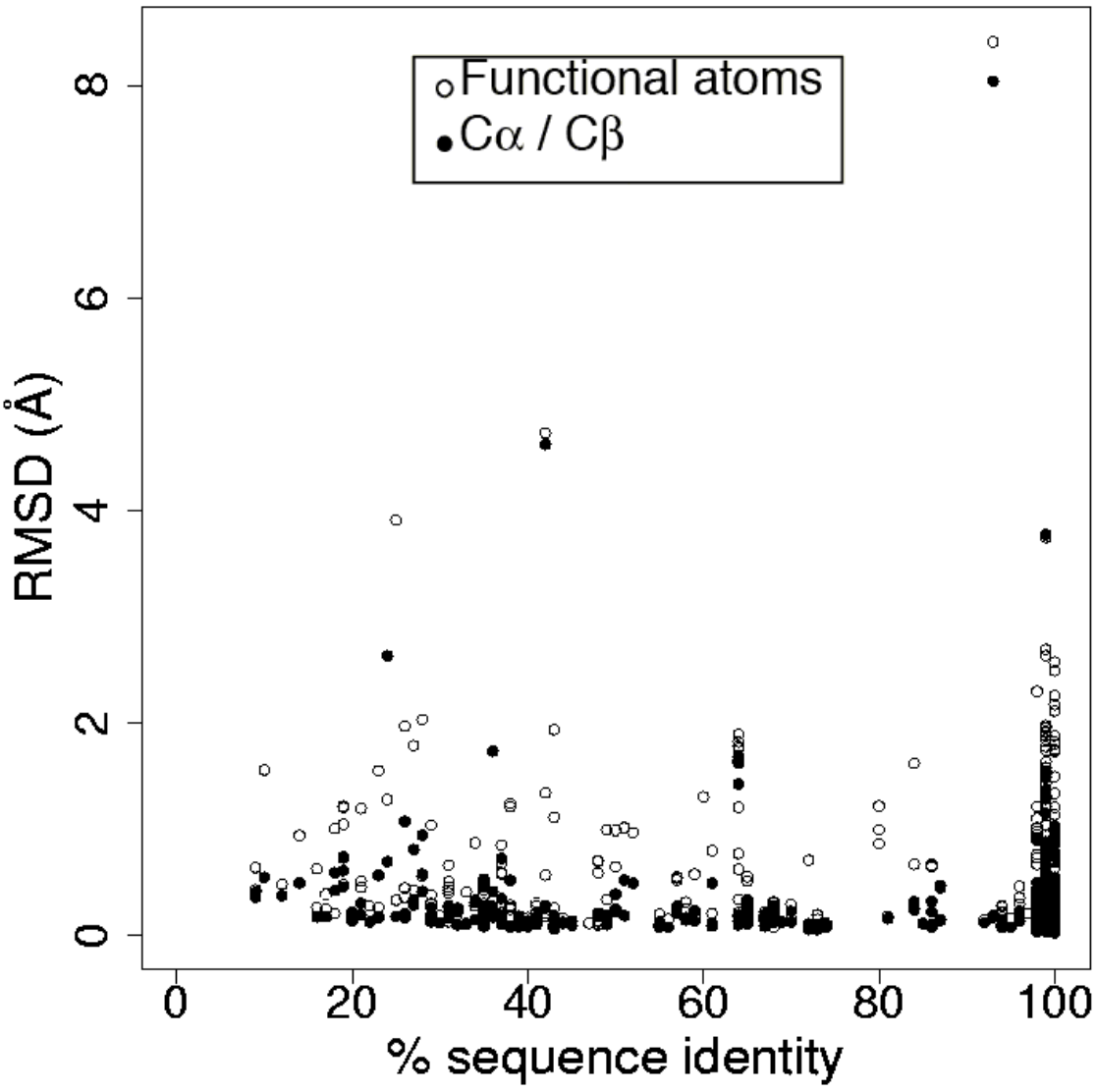


Figure 3.7: Relationship between sequence similarity and catalytic site structure similarity for four-residue templates.
Each point on the scatter plot corresponds to a single comparison between the representative entry for a given CSA family and another one of that CSA family's members.

et al., 1999). This family shows a correlation between sequence identity and structure (Figure 3.9b). Each catalytic site is composed of two histidines and a tyrosine. The catalytic sites fall into three groups, shown in Figure 3.10b. There is a group of structures for the protein corresponding to the representative template, a biphenyl-2,3-diol 1,2-dioxygenase (EC 1.13.11.39) from the bacterium *Burkholderia cepacia*; these all have 100% sequence identity and negligible RMSD. There is a group of biphenyl-2,3-diol 1,2-

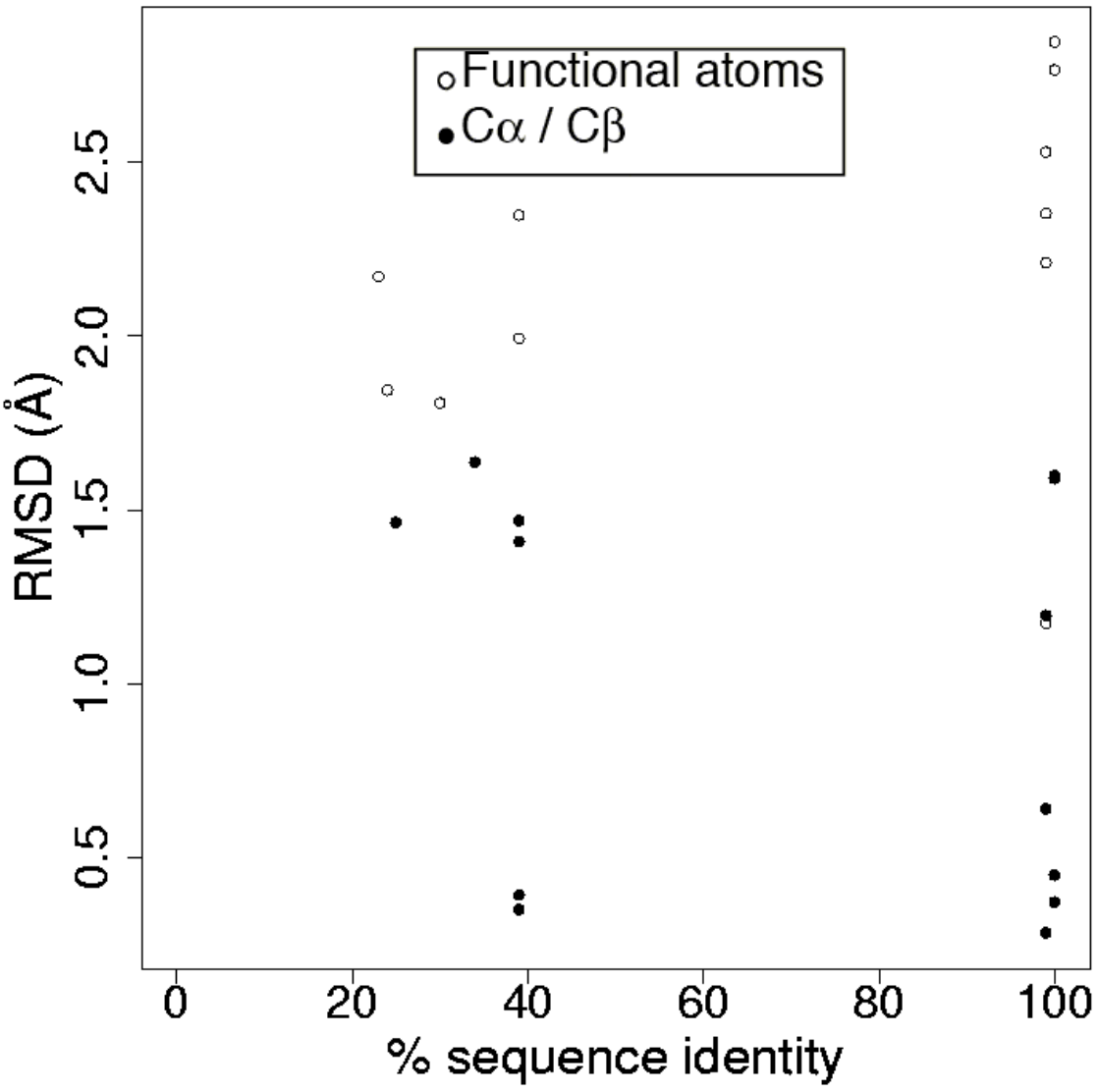


Figure 3.8: Relationship between sequence similarity and catalytic site structure similarity for five-residue templates.
Each point on the scatter plot corresponds to a single comparison between the representative entry for a given CSA family and another one of that CSA family's members.

dioxygenases from bacteria of the genus *Pseudomonas*, with sequence identity around 65% and RMSDs slightly higher than those for the *Burkholderia* enzymes. Finally, there is a group of enzymes with a slightly different activity: homoprotocatechuate 2,3-dioxygenase (EC 1.13.11.15). These have 20% sequence identity and the highest RMSDs. These homoprotocatechuate 2,3-dioxygenases have a slightly larger separation between the two catalytic histidines than other members of this CSA family, leading to the increased

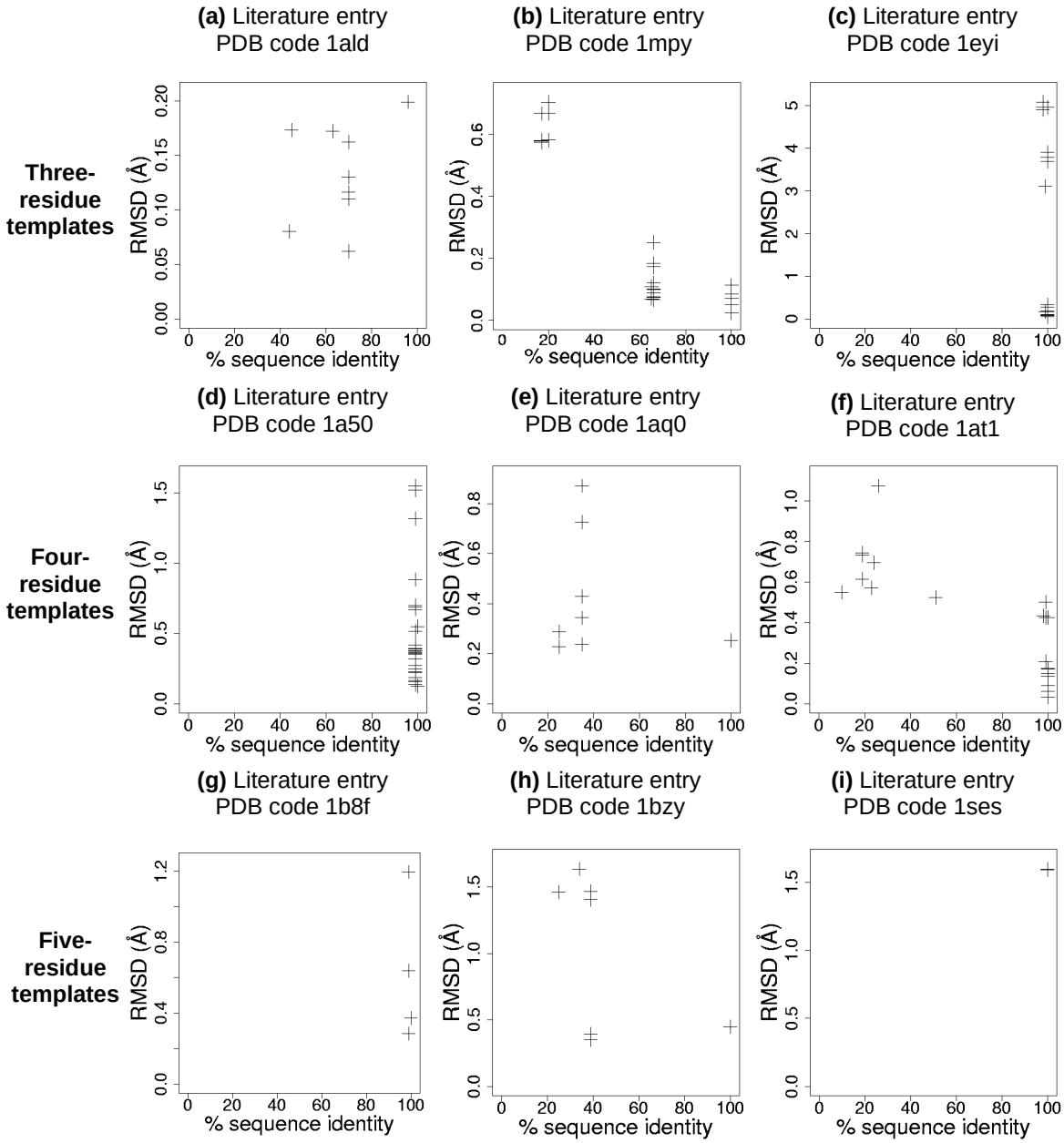


Figure 3.9: Relationship between sequence similarity and catalytic site similarity for example families.

Each point on each scatter plot corresponds to a single comparison between the representative member for a given CSA family and another one of that CSA family's members. All RMSDs are derived from C_α/C_β templates. **(a)** Literature entry 1ald, human aldolase A (Gamblin *et al.*, 1991) **(b)** Literature entry 1mpy, *Pseudomonas putida* catechol 2,3-dioxygenase (Kita *et al.*, 1999) **(c)** Literature entry 1eyi, *Sus scrofa* fructose 1,6-bisphosphatase (Choe *et al.*, 2000) **(d)** Literature entry 1a50, *Salmonella typhimurium* tryptophan synthase (Schneider *et al.*, 1998) **(e)** Literature entry 1aq0, *Hordeum vulgare* 1,3-1,4-β-glucanase (Keitel *et al.*, 1993) **(f)** Literature entry 1at1, *Escherichia coli* aspartate carbamoyltransferase (Gouaux & Lipscomb, 1990) **(g)** Literature entry 1b8f, *Pseudomonas putida* histidine ammonia-lyase (Schwede *et al.*, 1999) **(h)** Literature entry 1bzy, *Homo sapiens* hypoxanthine-guanine phosphoribosyltransferase (Shi *et al.*, 1999) **(i)** Literature entry 1ses, *Thermus thermophilus* seryl-tRNA synthetase (Belrhali *et al.*, 1994)

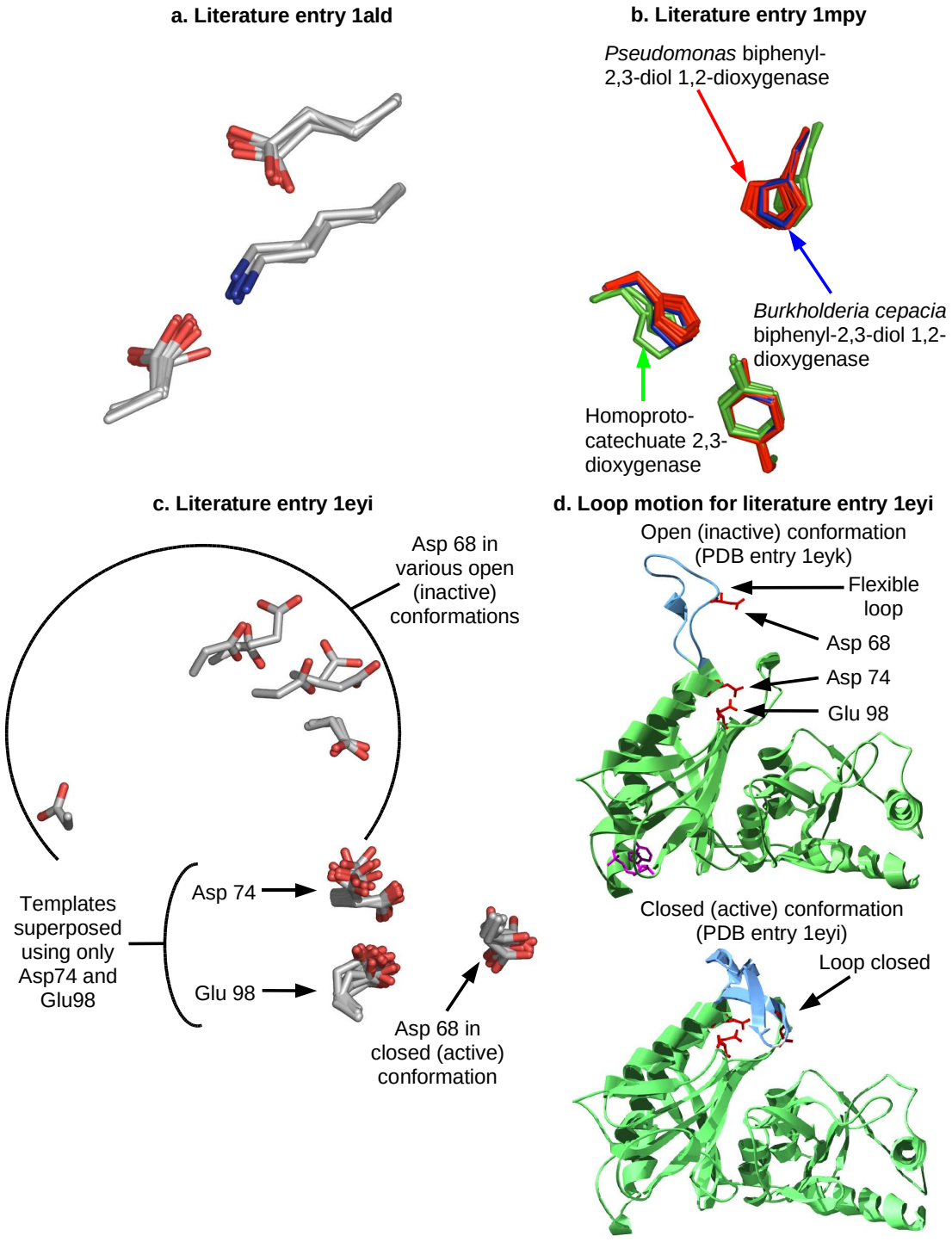


Figure 3.10: Catalytic site structures for example families.
All C_α/C_β templates corresponding to a single template family were superposed. The whole of the sidechain is shown for clarity, although only the C_α and C_β atoms were used in the template. Sections a, b and c were created using Pymol (www.pymol.org). Section d was created using the Swiss-Pdb viewer (Guex & Peitsch, 1997). **(a)** Literature entry 1ald, human aldolase A (Gamblin *et al.*, 1991) **(b)** Literature entry 1mpy, *Pseudomonas putida* catechol 2,3-dioxygenase (Kita *et al.*, 1999) **(c)** Literature entry 1eyi, *Sus scrofa* fructose 1,6-bisphosphatase (Choe *et al.*, 2000). Templates for 1eyi were superposed for this diagram using only the two residues Glu98 and Asp74. The third residue, Asp68, is on a loop which can be in a single closed (active) conformation, or in a variety of open (inactive) conformations. **(d)** Open and closed forms of *Sus scrofa* fructose 1,6-bisphosphatase (Choe *et al.*, 2000), showing the flexible loop.

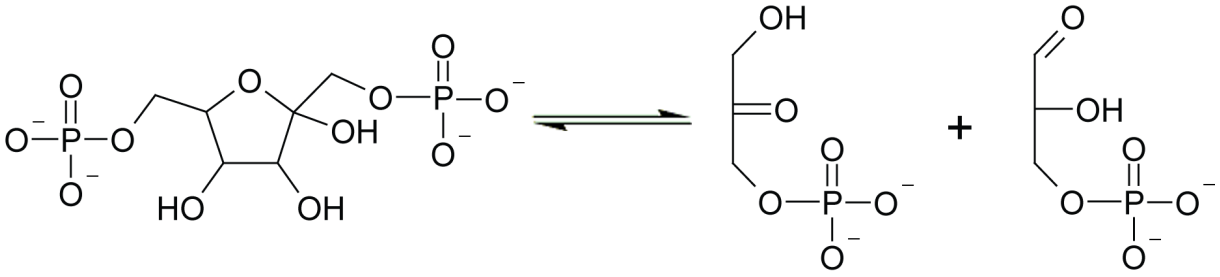


Figure 3.11: Aldolase reaction.

RMSD.

The fructose 1,6-bisphosphatase family (reaction shown in Figure 3.13) is based on the literature entry *Sus scrofa* fructose 1,6-bisphosphatase, with PDB code 1eyi (Choe *et al.*, 2000). This family is an instance of considerable variation in catalytic site structure in the absence of variation in sequence (Figure 3.9c). There are three catalytic residues: Asp 78 and Glu 98 catalyse the generation of a nucleophilic OH^- ion, whereas Asp 68 stabilises the transition state which comes about once the OH^- ion makes a nucleophilic attack on a phosphate group (residue numbering is for 1eyi) (Choe *et al.*, 2000). One of these catalytic residues (Asp 68) is part of a flexible loop. When AMP binds this enzyme at an allosteric site, the enzyme undergoes a conformational transition and this loop becomes disordered. Different crystal structures represent this disordered loop in a variety of conformations, as shown in Figure 3.10c and Figure 3.10d.

3.2.3 Family analysis

The family analysis, outlined in Figure 3.14, aimed to discover how well structural templates could discriminate family matches from random noise. Each CSA family was analysed individually. For each CSA family, a representative template was selected, as mentioned above. This representative template was used to search a non-redundant subset of the PDB, in order to discover the distribution of random matches for that template. These random matches were compared with matches to other members of the CSA family.

The distribution of family and random matches for templates of different sizes is shown in Figure 3.15, which combines the data for all families with a given template size. As noted above, the great majority of family matches have RMSDs below 1 Å. The great

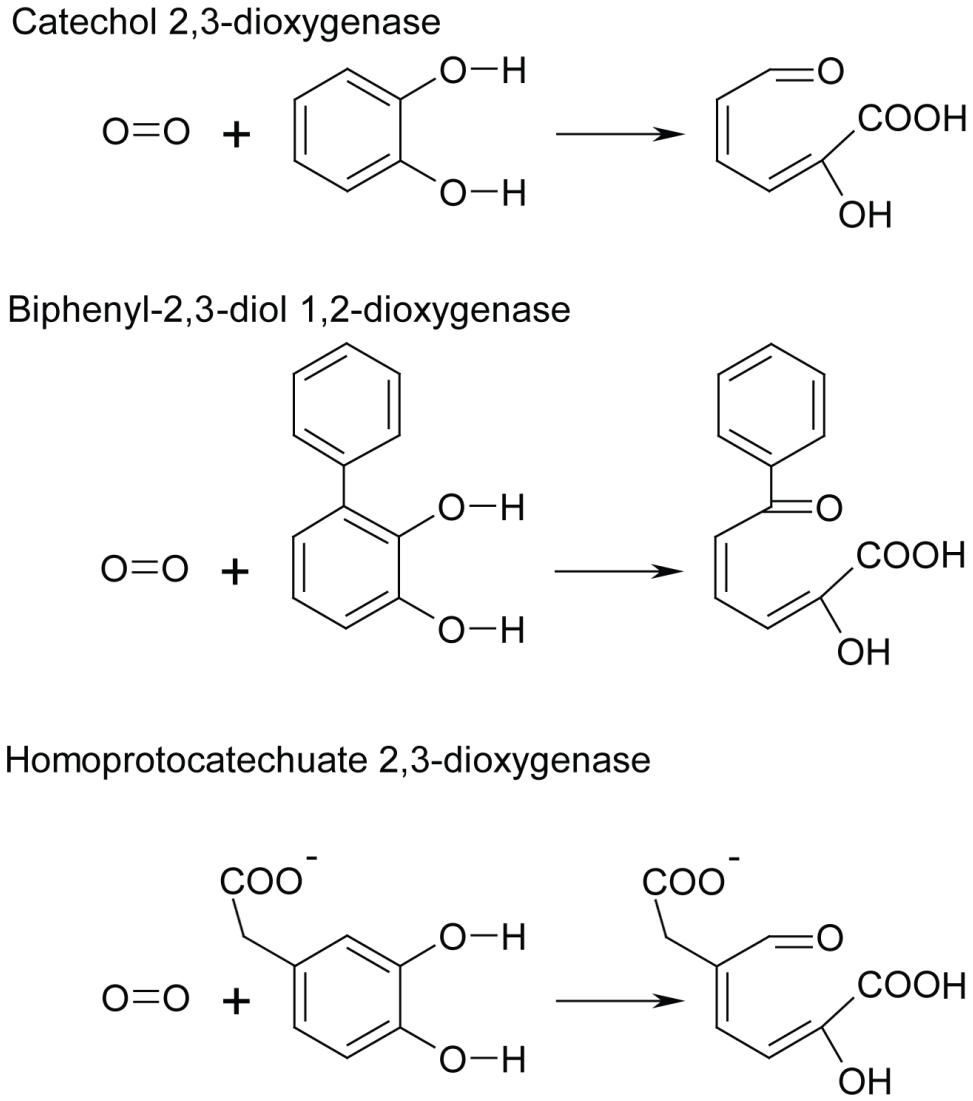


Figure 3.12: Catechol 2,3-dioxygenase family reactions.
Reactions performed by members of the catechol 2,3-dioxygenase family described in the text.

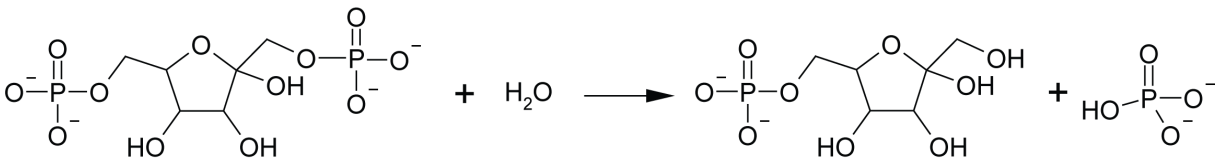


Figure 3.13: Fructose 1,6-bisphosphatase reaction.

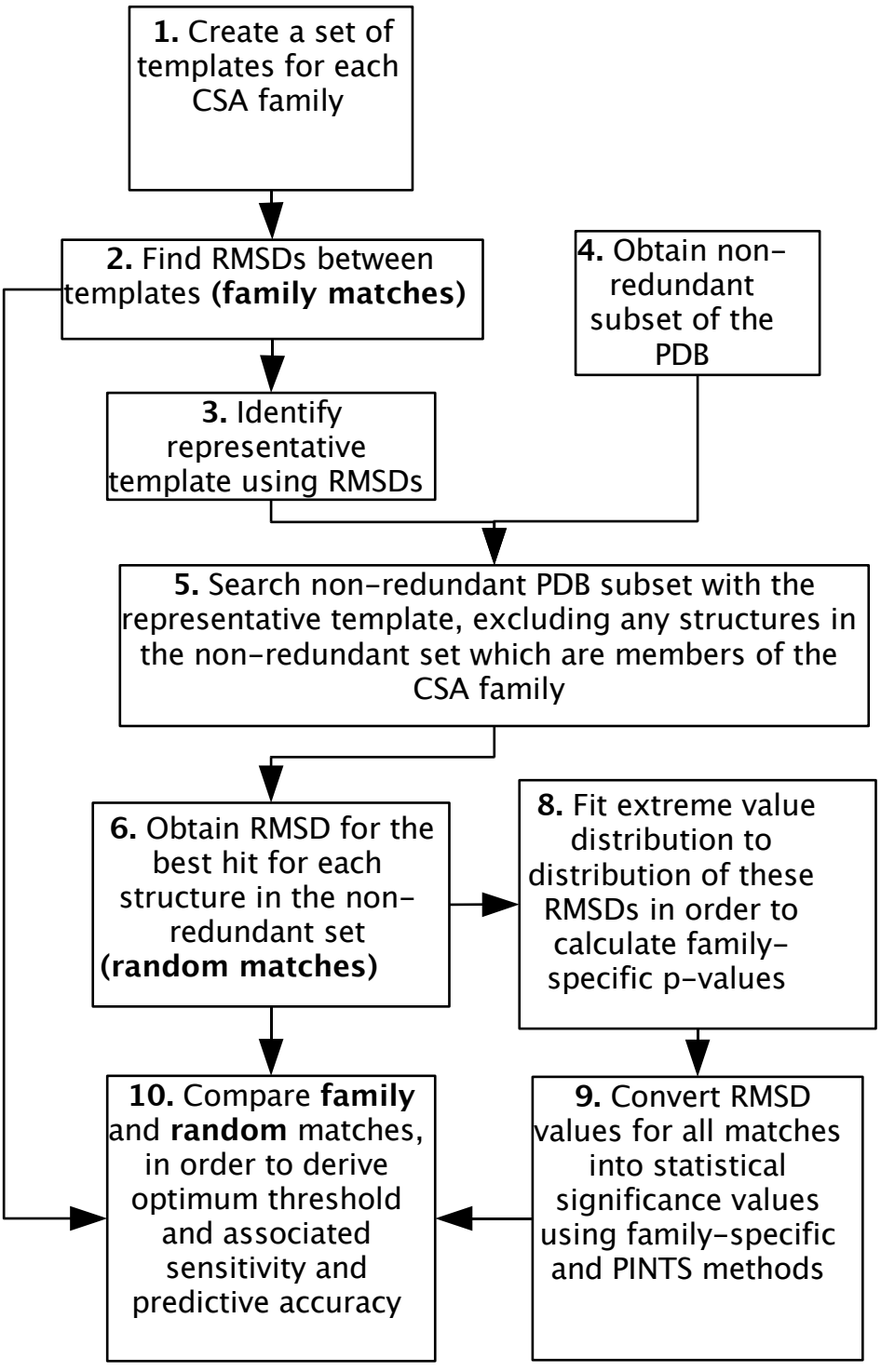


Figure 3.14: Family analysis flowchart.

majority of random matches have RMSDs above 1 Å. C_α/C_β templates tend to have lower RMSD matches than functional atom templates at any given template size. The separation of family matches from random ones is consistently better for C_α/C_β templates than for functional atom templates. The main difference between the different template sizes is that the RMSDs of matches tend to rise as template size increases. Functional atom templates with five residues show markedly worse separation of family from random matches than any other template category. The reason is not clear; in contrast, five-residue C_α/C_β templates have quite distinct separation of family and random matches.

In order to use templates to discriminate between family and random matches, it is necessary to set a threshold level of RMSD: below this threshold, matches are predicted as family members; above it, they are predicted as random. Figure 3.16 shows the effect that altering the threshold has on sensitivity (the proportion of family matches that are below the threshold), predictive accuracy (the proportion of matches below the threshold that are family matches) and Matthews Correlation Coefficient (Matthews, 1975) (MCC, an overall measure of the separation of family matches from random matches). The MCC rises higher, peaks more sharply, and peaks at a lower level of RMSD for C_α/C_β templates than for functional atom templates: C_α/C_β templates are better at distinguishing family members from random matches. Sensitivity rises and predictive accuracy falls more swiftly and at lower levels of RMSD for C_α/C_β templates than for functional atom templates.

The optimal threshold level was defined as the one where MCC was at a maximum. Optimal thresholds can be assigned individually for each CSA family. However, it is more convenient to have a single threshold across all families, since this makes it easier to compare matches from templates based on different CSA families. This overall threshold could be an RMSD, as employed in Figure 3.16. However, expressing the overall threshold in terms of RMSD suffers from the problem that, as discussed above, the significance of a given RMSD (and hence any given threshold) is different for different templates. This problem can be avoided by scoring matches to a given template in terms of their statistical significance, rather than their RMSD. Since a given level of statistical significance has the same meaning for all templates, a single statistical significance threshold might be expected to work better than a single RMSD threshold. Two statistical significance measures were employed here. The first was based on fitting an extreme value distribution

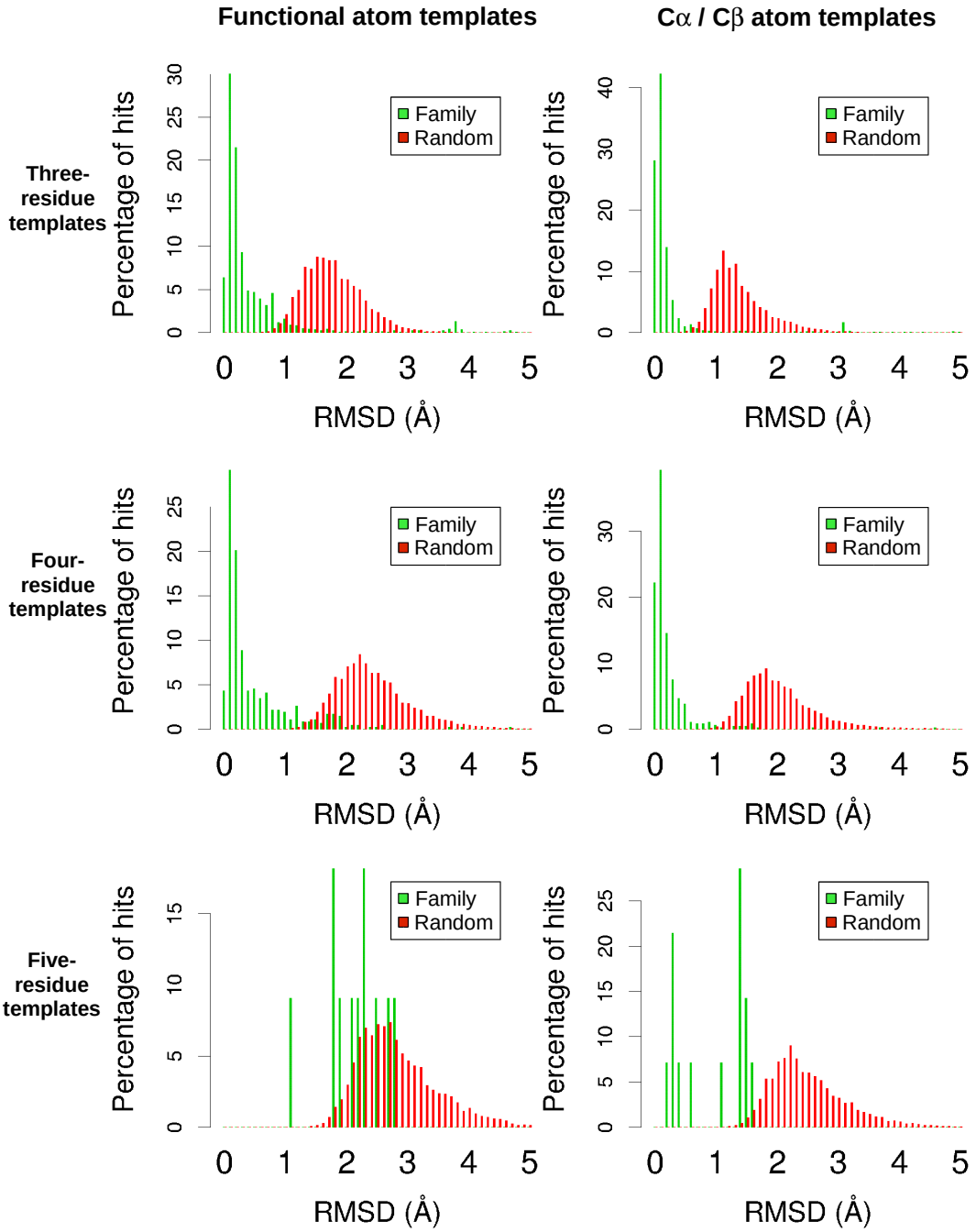


Figure 3.15: RMSD distribution of family and random matches for all CSA families. Each plot shows a histogram of the percentages of family matches that fall in various RMSD bins, overlaid on a histogram of the percentages of random matches falling in those RMSD bins. Each histogram bin covers 0.1 Å. Results are shown separately for families with three-residue templates. families with four-residue templates, and families with five-residue templates. There are many more random matches than family matches, but this is masked by showing the *percentage* of each type of match on the vertical scale.

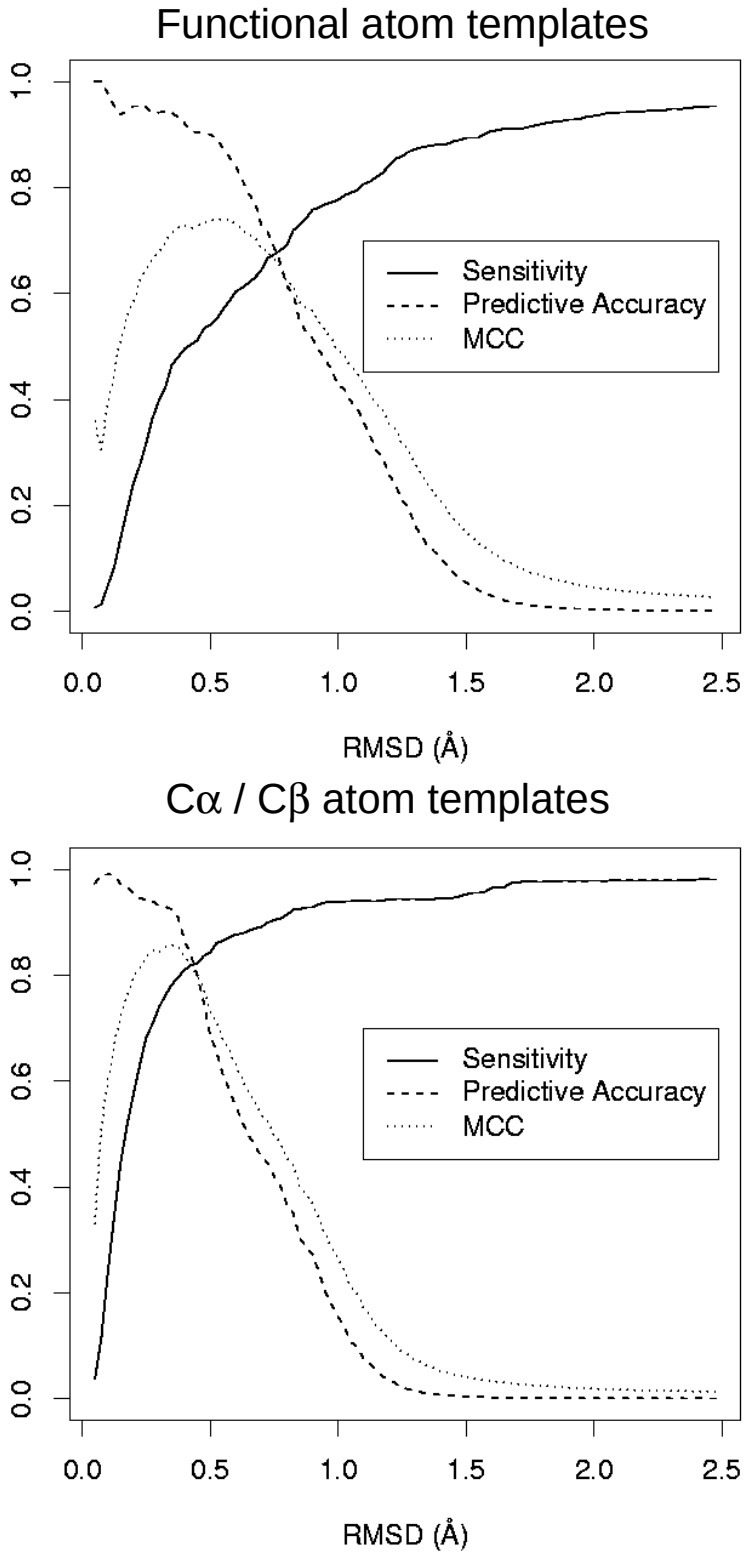


Figure 3.16: Ability of templates to discriminate family matches from random matches at different RMSD levels.

Results from all 147 CSA families are combined to analyse how overall performance varies with the threshold RMSD selected. Y-axis values are averages of all values for individual families at a given RMSD.

to the observed pattern of random matches for the representative template of each family (**family-specific p-value**). The distribution of matches was fitted independently for each representative template. The second statistical measure was the one adopted by the PINTS web server (Stark *et al.*, 2003) (**PINTS p-value**). This uses a formula based on the geometry of template matching. This method is not calibrated separately for templates representing different CSA families.

Table 3.3 shows performance levels for different types of template and different ways of setting a threshold. The C_α/C_β templates outperformed the functional atom templates on every measure. Using individual thresholds, the templates separated family matches from random ones very well: C_α/C_β templates had a MCC of 0.91, and similar scores for sensitivity and predictive accuracy. Setting a single threshold level for all templates led inevitably to a decrease in performance, but performance remained high, with sensitivity, predictive accuracy and MCC all at 0.80 or above for C_α/C_β templates. A single threshold defined in terms of statistical significance performed better than a single RMSD threshold, as expected. However, the improvement in performance was small. There was little difference between the effectiveness of a threshold based on family-specific p-values and the effectiveness of one based on PINTS p-values.

These statistics can be set in context by considering a few individual template families. The examples used here are the same as those used to illustrate the analysis of variation within families. As before, these examples focus on the better-performing C_α/C_β templates.

The aldolase template family (literature entry PDB code 1ald) showed perfect separation of family and random matches (Figure 3.17a). This was simply because catalytic sites in this family vary little in structure; the greatest RMSD was 0.45 Å. Over 60% of all template families similarly showed perfect separation of family and random matches when C_α/C_β atoms were used.

The biphenyl-2,3-diol 1,2-dioxygenase family (literature entry PDB code 1mpy) is an example of a family with a sub-population that was not detected by the representative template (Figure 3.17b). Most of the family are biphenyl-2,3-diol 1,2-dioxygenases (EC 1.13.11.39). However, the subpopulation that was not detected are homoprotocatechuate 2,3-dioxygenases (EC 1.13.11.15), and they have low sequence relationship (20% sequence

	Template type	Threshold level	Mean Matthews Correlation Coefficient	Mean sensitivity	Mean predictive accuracy
Individual thresholds	Functional atoms	-	0.79 (0.26)	0.78 (0.25)	0.87 (0.30)
	C_α/C_β	-	0.91 (0.18)	0.90 (0.18)	0.95 (0.19)
Overall RMSD threshold	Functional atoms	0.60 Å	0.63 (0.34)	0.60 (0.35)	0.72 (0.40)
	C_α/C_β	0.38 Å	0.80 (0.28)	0.80 (0.29)	0.85 (0.31)
Overall family-specific p-value threshold	Functional atoms	2.5×10^{-8}	0.69 (0.33)	0.68 (0.35)	0.75 (0.38)
	C_α/C_β	2.5×10^{-11}	0.83 (0.25)	0.85 (0.25)	0.86 (0.29)
Overall PINTS p-value threshold	Functional atoms	3.2×10^{-6}	0.68 (0.33)	0.67 (0.35)	0.76 (0.38)
	C_α/C_β	3.2×10^{-6}	0.83 (0.26)	0.86 (0.25)	0.84 (0.30)

Table 3.3: Ability of templates to discriminate family matches from random matches. All "mean" values are means of all values for the 147 individual template families. Values in brackets are standard deviations. Consult Methods section for a definition of terms.

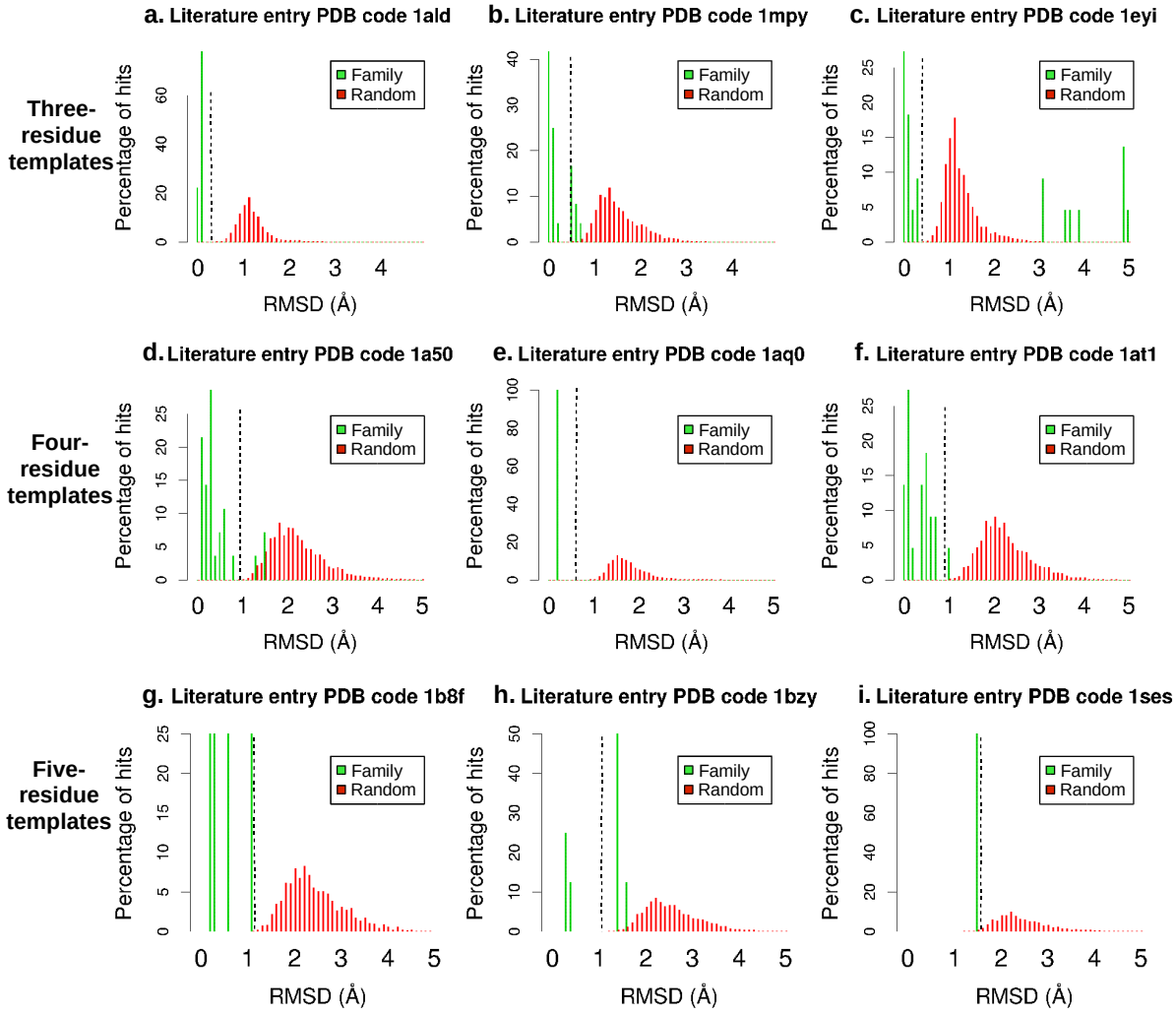


Figure 3.17: Distribution of family and random matches for example families. Each plot shows a histogram of the percentages of family matches that fall in various RMSD bins, overlaid on a histogram of the percentages of random matches falling in those RMSD bins. Each histogram bin covers 0.1 Å. All RMSDs are derived from C_α/C_β templates. The threshold RMSD level for each family is shown as a vertical dashed line. There are many more random matches (> 5000 for most families) than family matches (< 50 for most families), but this is masked by showing the *percentage* of each type of match on the vertical scale. **(a)** Literature entry 1ald, human aldolase A (Gamblin *et al.*, 1991) **(b)** Literature entry 1mpy, *Pseudomonas putida* catechol 2,3-dioxygenase (Kita *et al.*, 1999) **(c)** Literature entry 1eyi, *Sus scrofa* fructose 1,6-bisphosphatase (Choe *et al.*, 2000) **(d)** Literature entry 1a50, *Salmonella typhimurium* tryptophan synthase (Schneider *et al.*, 1998) **(e)** Literature entry 1aq0, *Hordeum vulgare* 1,3-1,4-β-glucanase (Keitel *et al.*, 1993) **(f)** Literature entry 1at1, *Escherichia coli* aspartate carbamoyltransferase (Gouaux & Lipscomb, 1990) **(g)** Literature entry 1b8f, *Pseudomonas putida* histidine ammonia-lyase (Schwede *et al.*, 1999) **(h)** Literature entry 1bzy, *Homo sapiens* hypoxanthine-guanine phosphoribosyltransferase (Shi *et al.*, 1999) **(i)** Literature entry 1ses, *Thermus thermophilus* seryl-tRNA synthetase (Belrhali *et al.*, 1994)

identity or lower) to the other members of the family. As described above, these homoprotocatechuate 2,3-dioxygenases have a slightly larger separation between the two catalytic histidines than other members of this CSA family: this was the structural difference that prevented them being detected. However, they are very close to the RMSD threshold for the family; this highlights the limitations of using a strict cut-off. For this reason, when manually inspecting template matches, it may sometimes be appropriate to relax thresholds, especially if no matches better than threshold are observed.

The fructose 1,6-bisphosphatase family (literature entry PDB code 1eyi) is an example of a family that is difficult to represent using a template (Figure 3.17c). As described above, when AMP binds an allosteric site, a loop containing one of the catalytic residues becomes disordered, inactivating the enzyme. Different crystal structures represent the disordered loop in a variety of conformations (Figure 3.10c). As a result, only the active state of the enzyme (without AMP bound) can be represented by a template. In cases like this, it may prove useful to remove conformationally flexible catalytic residues from the template. If this makes the template too small, then rigid non-catalytic ligand-binding residues could be added to the template.

3.2.4 Library analysis

The family analysis described in the previous section provided an assessment of how well individual template families perform. However, the user of the template library will rarely be asking "How well do individual templates perform?". Instead, they will usually be asking "When I run a single structure against the whole library, is the top match meaningful?"

Here the problem of comparing results between different templates is critical. As described above, the meaning of RMSD is not comparable between templates. This creates potential problems. The most obvious is that in a template library containing small and large templates, the small templates might frequently get lower RMSD matches than those obtained by large templates– even when the matches from the small templates are not meaningful and the matches from the large templates are meaningful. The following library analysis looks at this problem of comparing matches from different templates, and

the extent to which the results can be improved by scoring template matches using a statistical significance measure, rather than RMSD.

The library analysis resembles the anticipated usage of the CSA template library, in that one structure is run against the whole library, as described in Figure 3.18. The template library used here is the set of all the representative templates selected in the family analysis. The set of structures that was searched using this library consisted of all structures included in the family analysis– except the structures from which the representative templates were derived. For each one of these structures, the analysis asked whether the top match from the whole template library corresponded to the representative template for that structure's CSA family. In other words, whether the top match was the right match. As with the family analysis, three different ways of scoring the template matches were used: RMSD, family-specific p-value, and PINTS p-value. As part of the family analysis described above, a threshold level of RMSD was set individually for each template. The library analysis looked at how useful this threshold was in practice– does discarding all matches that are worse than the threshold improve the library analysis results?

One category of template families was excluded from the library analysis: those which involved a Ser-His-Asp catalytic triad as found in trypsin or subtilisin. Because this triad has evolved independently several times in different structural frameworks (often described as "convergent evolution"), there were frequent cross-hits between these families. This resulted in large numbers of matches which the analysis regarded as "random" but which were in fact meaningful. Not only did this produce spuriously low performance levels for these families, it also led to the extreme value distribution tending to fit poorly to the "random" matches. This in turn meant that their family-specific p-value results did not correspond to the real probabilities of random matches. This made it difficult to investigate the effectiveness of p-values in the library analysis unless these families were excluded. Since there were no other known catalytic motifs in the dataset that had evolved independently with a similar frequency, it seems likely that within the present dataset this type of problem is limited to Ser-His-Asp triads. Note that the active sites in this dataset are groups of at least three residues; convergent evolution of active sites does occur on groups of fewer than three residues, such as the pairs of acidic residues which

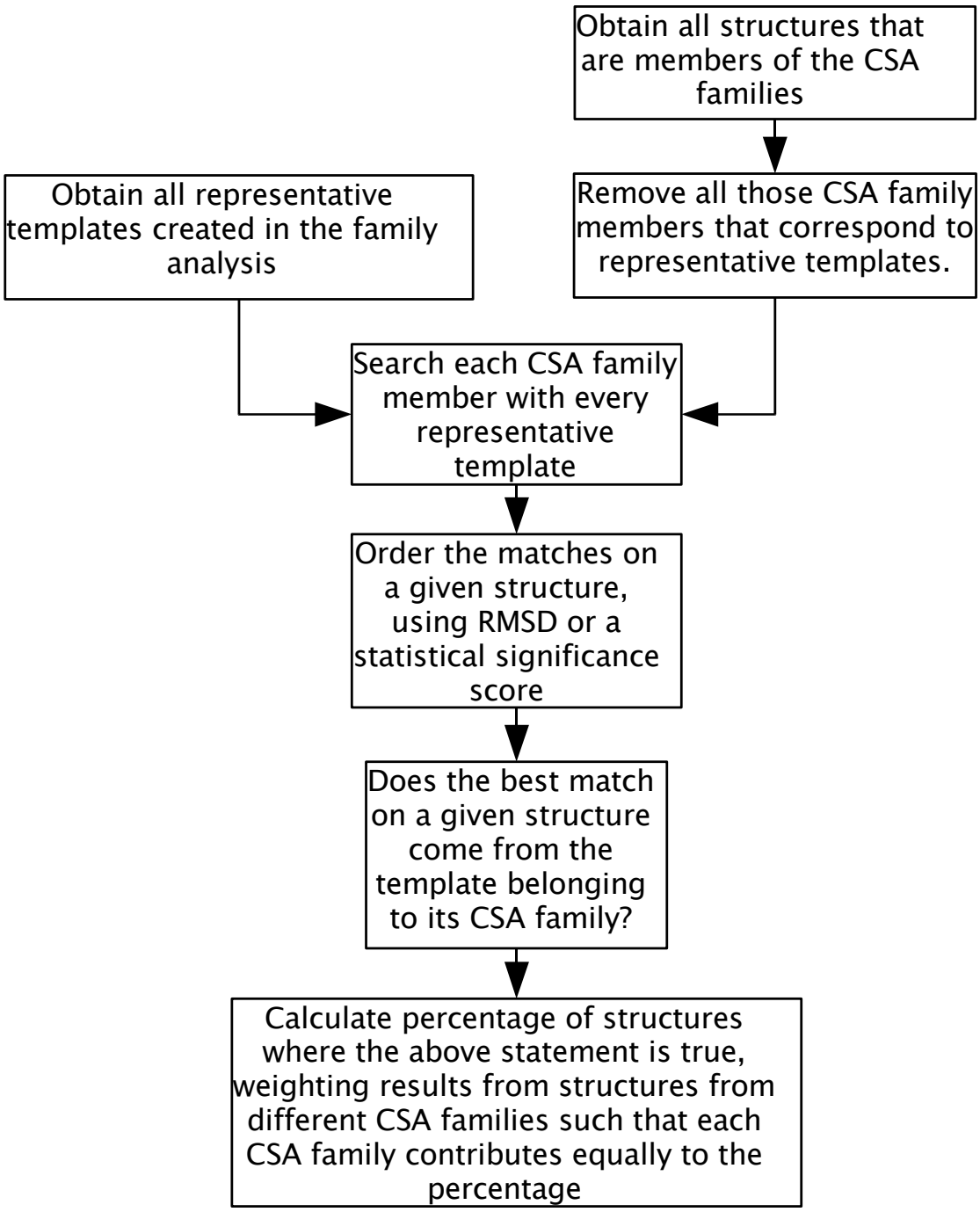


Figure 3.18: Library analysis flowchart.

recur in unrelated glycosyl hydrolases (Davies & Henrissat, 1995). Furthermore, cases of convergent evolution of groups of three or more residues may exist in enzymes outside the present dataset.

The results (Table 3.4) show that the right match is the best match on the great majority of occasions: in over 85% of cases when C_α/C_β templates are used. As with the family results, C_α/C_β templates consistently perform better than functional atom templates. When matches are required to be below the individual template thresholds, this generally decreases performance, but not markedly. Statistical significance scores perform better than RMSD, as expected. However, the improvement is very slight, and only occurs when the template thresholds are ignored.

RMSDs tend to be higher for comparisons that involve more atoms. Therefore one would expect the use of statistical significance to be most helpful for large templates, because this turns the high RMSD values for these large templates into scores that are comparable with those of smaller templates. Separate analyses were carried out for the different template sizes. These analyses ignored thresholds and were only carried out

	Template type	Ignoring thresholds	Discarding matches worse than threshold
RMSDs	Functional atoms	75.6	77.0
	C_α/C_β	87.9	87.5
Family-specific p-values	Functional atoms	83.6	75.2
	C_α/C_β	91.4	87.5
PINTS p-values	Functional atoms	82.3	76.9
	C_α/C_β	91.2	88.4

Table 3.4: Library analysis results.
All values are the percentages of all structures tested where the best match against the library is the correct match.

for C_α/C_β data, using family-specific p-values. For three-residue templates, there was virtually no difference between the performance for RMSD and that for family-specific p-values (89.8% versus 90.3%). For four-residue templates, there was a small difference (86.3% versus 92.7%). It was only for the five-residue templates that there was any considerable benefit from using p-values (25.0% versus 66.7%). There was only a small number of five-residue templates in the whole dataset; this is why there was little difference between RMSD performance and family-specific p-value performance in the overall values given in Table 3.4.

3.3 Discussion

3.3.1 Structural conservation of active sites and the performance of structural templates

The purpose of this chapter was to look at the *scientific* question of the degree of structural conservation of enzyme catalytic sites, and the *technical* question of how well structural templates can recognise these sites. There has been no previous general study of the structural conservation of enzyme active sites. The results show that most catalytic sites are strongly structurally conserved, even between distant relatives. Previous assessments of the effectiveness of templates in recognising functional sites have only looked at their performance for one or a few types of functional site. Most often, this has been the Ser-His-Asp catalytic triad, which is a somewhat unusual case in that it has independently evolved multiple times. Arakaki *et al.* evaluated the performance of a library of structural templates; however, their focus was on the applicability of templates to low-resolution or predicted structures, rather than testing how well templates can recognise known relatives (Arakaki *et al.*, 2004). Thus, the present study is the first to examine a library of structural templates to see how well they can discriminate relatives from random matches. The results show that the discriminative ability of structural templates was high, especially when C_α and C_β atoms were used to represent residues.

Catalytic site structure was highly conserved in most CSA families– even at low levels of sequence identity. Previous studies have noted that catalytic site structure is strongly

conserved between distant relatives for proteins with Ser-His-Asp triads (Wallace *et al.*, 1996) and within the enolase superfamily (Meng *et al.*, 2004). The results presented in this chapter demonstrate that this structural conservation is not merely a feature of these two specific enzyme groups, but is a general feature of most enzymes. The most obvious explanation is that the precise positioning of catalytic residues is frequently critical to their function, and therefore the positions of the catalytic residues are strongly conserved, even when much of the protein sequence has diverged. For example, it has been shown for isocitrate dehydrogenase that a 0.56 Å change in the distance between the hydride acceptor-donor pair of the cofactor and substrate can reduce the reaction rate by almost three orders of magnitude (Mesecar *et al.*, 1997; Koshland, 1998).

Protein structure generally diverges as sequence identity falls (Wilson *et al.*, 2000). However, within most CSA families there was not a significant relationship between catalytic site RMSD and sequence identity. This lack of relationship may be due to strong functional constraints on catalytic site conformation, as discussed in the previous paragraph. A second possible explanation for the lack of relationship is that other factors may be more important than evolutionary divergence in determining catalytic residue position. These include whether the structure has substrate bound or not; whether it is bound to cofactor; and whether there are artificially introduced mutations at other residues affecting catalytic residue conformation. The present study dealt with a large number of template families, and many individual template families had only a small number of members. For these reasons it was impractical to control for the factors described above by, for example, only considering structures that were not bound to substrate. A study that used a structure dataset that was manually curated in this manner might be able to shed more light on how catalytic sites change as proteins evolve.

Both the family and library analyses showed that structural templates were generally able to discriminate well between relatives and random matches. This was possible because of the strong structural conservation of catalytic sites.

As noted above, the catalytic sites of some enzymes can undergo substantial conformational changes, often due to binding of a substrate, inhibitor, cofactor or allosteric effector. The good performance of the template library in the family and library analyses suggests that such conformational changes do not present a major obstacle to the use of

templates in general. However, there are a minority of enzymes where this is a problem, such as *Sus scrofa* fructose 1,6-bisphosphatase. In these cases, it may prove useful to replace the most conformationally variable residues in the template with other conserved residues from the active site.

The chemical functions of catalytic residues are typically carried out by a few atoms near the end of the sidechain. One might expect that the positions of these atoms would be the aspect of the enzyme most conserved through evolution. Yet the level of structural variation of C_α and C_β atoms within families was lower than the variation of functional atoms, and C_α/C_β templates were consistently better at discriminating family matches from random ones. Residue sidechains are more flexible than the main chain of a protein, and they are more liable to alter their conformation in the presence of ligand (Gutteridge & Thornton, 2004). Although these factors mean that C_α/C_β based structural templates work best when detecting close homologues, it is possible that functional atom templates might work better for distant homologues or instances of convergent evolution. For instances of convergent evolution in particular, one would expect functional constraints to lead to similar sidechain positions, but backbone positions would be under no such constraint.

3.3.2 Statistical significance measures

The significance of a given level of RMSD is not the same for different templates. Therefore one would expect statistical significance scores to perform better than RMSD in the library analysis. The improvement was considerable for five-residue templates, but for smaller templates there was little benefit to using statistical significance scores. The great majority of templates in the library analysed in this chapter consisted of three or four residues. In a template library that was more diverse in template size, it would probably be more important to either make use of statistical significance scores, or use different RMSD thresholds for different template sizes.

There was little difference between the performance of family-specific p-values and PINTS p-values. PINTS scoring has the advantage over family-specific p-values that it does not require a set of random matches to be generated for each template. PINTS

scoring is therefore more convenient than family-specific p-values, and equally effective. However, for the reasons described in the previous paragraph, the dataset used in this work is not a good one for understanding the performance of statistical significance measures. These conclusions about the relative performance of different statistical significance measures should therefore be viewed with caution.

The overall thresholds for PINTS p-values and family-specific p-values differed considerably. This was due to the differing shapes of the distributions involved at low values of RMSD. Since the data used for fitting the PINTS and family-specific p-value probability distributions contained few (if any) of the low RMSD values occurring around these threshold levels, the shape of the distributions at these values has no relation to actual data. The probability distribution used for calculating the family-specific p-values was selected empirically. The probability distribution used for the PINTS p-values has a more sound theoretical basis, so the PINTS threshold p-values are more likely to be meaningful.

3.4 Methods

3.4.1 Non-redundant set of CSA families

The CSA is based around a set of structures whose catalytic residues are recorded based on information from the scientific literature. These structures were chosen in order to maximise coverage of different catalytic functions. As described in the previous chapter, proteins of known structure that are related to these literature-based entries have been identified using PSI-BLAST, and the resulting sequence alignment has been used to identify the catalytic residues of these relatives. In this chapter, a literature-based CSA entry plus its relatives identified using PSI-BLAST are referred to as a **CSA family**. There is some overlap between these CSA families.

All analyses in this chapter used a non-redundant set of CSA families, selected from version 2.0 of the CSA. Families were regarded as redundant with one another if they met either of the following criteria:

- Common family members.
- The literature-based entries for the families shared both the same structural homol-

ogous superfamily (as defined by either CATH or SCOP) and the same EC number to four digits.

Where a pair of families were redundant with one another, the family containing the larger number of structures was used. CSA families were only used if at least three sidechain-acting residues were annotated as catalytic, because templates consisting of only two residues return very large numbers of matches and can seldom discriminate family matches from random ones. CSA families which had no members other than the literature entry were not used. One CSA family was excluded because the sequence alignment between the literature entry and other family members was found to be faulty.

Family members were only included in this study if:

- They had the same residue type as the parent literature entry for all catalytic residues.
- They were based on X-ray crystallography, not NMR.
- Their resolution was better than 2.5 Å.

It was not possible to fit an extreme value distribution (see below) to some of the larger families, because they returned an insufficient number of random matches for fitting. All analyses only used those families where it was possible to fit an extreme value distribution to the random results for both C_α/C_β and functional atom templates.

The EC number coverage of the family members was relative to the total numbers of codes in IntEnz release 2 (www.ebi.ac.uk/intenz) and numbers of codes covered by the PDB version released on the 5th October 2004.

3.4.2 Template generation (Figure 3.14 box 1)

Two banks of templates were constructed from the CSA: one bank of templates that represented residues in terms of the positions of their C_α and C_β atoms, and one bank of templates that represented each residue using three functional atoms (Table 3.5). These were the atoms most likely to be involved in catalytic activity. For residues with fewer than three atoms likely to be involved in catalytic activity, atoms were chosen that were at

the distal end of the sidechain from the protein backbone. The templates permit matching of chemically similar residues (Ser matches Thr, Asp matches Glu and Asn matches Gln). They also allow for matching between chemically equivalent atoms in a residue, such as the two oxygens in the carboxyl group of Asp. Those template residues that are annotated in the CSA as acting via their backbone (such as the residues forming the oxyanion hole in the serine proteases) were permitted to match any residue type. Residues that are annotated in the CSA as acting both via their sidechain and their backbone were treated as acting via their sidechain alone.

Within each family, one functional atom template and one C_α/C_β template were created from each family member. RMSDs between all templates in the family were calculated using the McLachlan algorithm (McLachlan, 1982) as implemented in the program ProFit (Martin, A.C.R., www.bioinf.org.uk/software/profit/), allowing for ambiguous atom identifiers as described above (Figure 3.14 box 2). The template with the lowest mean RMSD from all other templates was taken as the representative template for the family (Figure 3.14 box 3).

3.4.3 Similarity within template families

The percentage sequence identity between proteins was determined using ClustalW (Thompson *et al.*, 1994) to align the sequence families, using its default settings, including employing the BLOSUM series of substitution matrices (Henikoff & Henikoff, 1992) and using the slow, full dynamic programming algorithm for initial pairwise alignment. Correlation between sequence identity and RMSD was quantified using Spearman's rank correlation coefficient, a non-parametric method. Statistical significance could only be calculated where there were at least four structures in the family. The analyses that separated templates according to the number of residues they contained only used those templates where all residues were annotated in the CSA as acting catalytically via their sidechains.

3.4.4 Non-redundant PDB subset (Figure 3.14 box 4)

The non-redundant subset of the PDB was based on the non-redundant chain set provided by the NCBI

Residue	Atoms		
ALA	CA	CB	C
ARG	NE	NH1	NH2
ASN	CG	OD1	ND2
ASP	CG	OD1	OD2
CYS	CA	CB	SG
GLU	CD	OE1	OE2
GLY	CA	N	C
GLN	CD	OE1	NE2
HIS	CG	ND1	NE2
ILE	CD1	CG1	CG2
LEU	CG	CD1	CD2
LYS	CD	CE	NZ
MET	CG	SD	CE
PRO	CB	CD	CG
PHE	CZ	CE1	CE2
SER	CA	CB	OG
THR	CA	CB	OG1
TYR	CE1	CZ	OH
TRP	NE1	CZ2	CH2
VAL	CB	CG1	CG2
Main chain carbonyl	C	CA	O
Main chain amide	N	CA	C

Table 3.5: Atoms used in functional atom templates.
Atom types shown in grey for a particular residue are allowed to match one another because they are chemically equivalent.

(www.ncbi.nlm.nih.gov/Structure/VAST/nrpdb.html), using the version released on the 3rd of February 2004. Structures were regarded as redundant with one another if there was a BLAST hit between them with an E-value of 10^{-80} or less (i.e. essentially identical sequences were removed). If any chain occurred in this non-redundant set, the entire parent structure was used. The quaternary structural state of the protein that was used was the one judged by PQS to be most likely to correspond to the biological form of the protein (Henrick & Thornton, 1998). Where a structure was not available from the PQS set, the structure was used in the form stored in the PDB. When the family analysis was

carried out, a match between a template t and a member of the non-redundant PDB subset m was discarded if m was a member of t's CSA family.

3.4.5 Template matching

The program Jess (Barker & Thornton, 2003) was used for template matching (Figure 3.14 box 5). The templates were configured such that Jess only returned matches if all inter-atom distances in the match differed from the equivalent distances in the template by no more than 6 Å.

Users of this template library are likely to follow up only the best match of a given template on a given structure. For this reason, this analysis did not use all matches obtained by a representative template searching against the a particular structure. This analysis only used the best (lowest RMSD) match obtained by a given template on a given structure.

3.4.6 Statistical significance of template matches

The statistical significance of template matches to structures was estimated in two ways (Figure 3.14 boxes 8 and 9). In both cases, the value estimated was the probability of a specific match between one template and one protein structure occurring at random.

The first statistical significance score (PINTS p-value) was calculated using the method of Stark *et al.* (Stark *et al.*, 2003) as used by the PINTS web server. The formula used is based on the geometry of a match with a given RMSD level, taking into account the abundance of different residue types and the different geometry that applies when different numbers of atoms are used to represent each residue. Several parameters in the formula are derived empirically.

The formula is:

$$E = Pa\Phi bR_M^{0.97}[yR_M^2]^{S-1}[zR_M^3]^{T-1} \text{ where } N = 2$$

$$E = Pa\Phi cR_M^{2.93N-5.88}[yR_M^2]^{S}[zR_M^3]^{T} \text{ where } N \geq 3$$

Where E is expected number of matches with this RMSD or better, P is the number

of proteins that were searched with the template, Φ is the product of the percentage abundances of all residues, R_M is RMSD, N is the total number of residues, S is the number of residues of two atoms, and T is the number of residues of three atoms or more. All other variables are empirically derived constants: $a = 3.704\times10^{6}$, $b = 1.277\times10^{-7}$, $c = 1.790\times10^{-3}$, $y = 0.196$, $z = 0.094$.

The expected number of matches, E, was converted into the probability of any matches occurring, P, using the formula:

$$P = 1 - e^{-E}$$

The second statistical significance score (family-specific p-value) was based on fitting a statistical distribution to the distribution of matches observed at random for each template. The statistical distribution was the extreme value distribution in its double exponential form. Before fitting to the extreme value distribution, the RMSD values were subjected to a $\frac{1}{x^2}$ transformation. This transformation was chosen purely empirically, to obtain a better fit between distribution and data. The set of matches observed at random was obtained by searching the non-redundant PDB subset with the template using Jess. The extreme value distribution was only fitted to those results below 2 Å RMSD, since above this level Jess does not return all results, as a result of the constraint on inter-atom distances described above. Where there were fewer than 50 random results with RMSDs below 2 Å, it was not practical to fit an extreme value distribution. When this was the case, the family was excluded from all analyses. The fitting of the extreme value distribution to the transformed RMSD data was carried out by a program written by Jonathan Barker.

3.4.7 Setting a threshold (Figure 3.14 box 10)

Threshold levels for RMSD, family-specific statistical significance and PINTS statistical significance were set such that Matthews Correlation Coefficient was maximised.

3.4.8 Definition of statistical terms

$$Sensitivity = \frac{Tp}{Tp + Fn}$$

$$Predictive\ accuracy = \frac{Tp}{Tp + Fp}$$

$$Matthews\ correlation\ coefficient = \frac{TpTn - FpFn}{\sqrt{(Tp + Fp)(Tp + Fn)(Tn + Fp)(Tn + Fn)}}$$

Where Tp, Tn, Fp and Fn are the true positives, true negatives, false positives and false negatives respectively. Matthews Correlation Coefficient (Matthews, 1975) is an overall measure of how well a binary classification works. It is also known as Φ value (Cramér, 1946).

These calculations posed a potential problem in that the set of family members (true positives and false negatives) included all available structures (within the constraints described above) and was therefore redundant, whereas the set of proteins used to find random matches (true negatives and false positives) was non-redundant. In order to cope with this, the number of family members was scaled to be commensurate with the non-redundant data set in all calculations of predictive accuracy and Matthews Correlation Coefficient. The numbers of true positives and false negatives were scaled such that the total of all family members would be the same as the number of family members that occurred in the complete non-redundant data set. Note that, as described above, these family members occurring in the non-redundant dataset were removed from it before searching for random matches.

3.4.9 Analysing the results of the family and library analyses

The analyses that separated templates according to the number of residues they contained only used those templates where all residues were annotated in the CSA as acting catalytically via their sidechains. This was because those residues which acted via their mainchains were permitted to match to any residue, and thus might be expected to show a different behaviour in their pattern of random matches than residues which were constrained to match a single residue type.

Chapter 4

Using structural templates to analyse zinc and calcium binding sites

4.1 Introduction

The work described in the previous chapter examined the structural variability of catalytic sites, and the usefulness of structural templates for recognising these sites. This type of analysis has applications beyond catalytic sites; there are a range of functional sites in proteins which consist of small groups of residues which lie close together in the protein structure but which may be widely separated in the protein sequence. Metal binding sites are a prominent example. This chapter digresses from the theme of enzyme active sites to examine how the type of analysis described in the previous chapter can be applied to metal binding sites.

Metal ion binding can stabilise a protein by cross-linking points in the protein which may lie distant from one another in the protein sequence. This is particularly useful for small domains which would not otherwise be stable. Calcium and zinc are the ions that are most commonly employed by proteins for this purpose (Petsko & Ringe, 2004).

For the purposes of this chapter, a metal binding site in a protein consists of one metal ion and all protein side chains and water molecules that participate in its first coordination

sphere. The characteristic residue preferences and binding geometries (in terms of angles between residues) of different metals are well known (Glusker, 1999; Harding, 2001, 2004). Zinc ions playing a structural role are bound by sets of four cysteine and/or histidine residues with a tetrahedral geometry (example shown in Figure 4.1a) (Dudev & Lim, 2003); calcium ions are typically bound by seven groups in a pentagonal bipyramidal geometry comprising oxygen atoms from sidechains, backbone carbonyl groups, and water molecules (example shown in Figure 4.1b) (McPhalen *et al.*, 1991).

The residue preferences of metals are largely determined by the polarisability of the metal ion. Less easily polarised ions such as Ca^{2+} prefer to bind less easily polarised residue groups such as the carboxyl group in glutamate and aspartate. More easily polarised ions, such as Zn^{2+}, are able to bind to easily polarised residue groups, such as the sulphydryl group in cysteines (Pearson, 1963; Glusker, 1999).

It is rare for the sequence separation of metal-binding residues to show similar patterns between unrelated proteins (Harding, 2004). A possible exception is the group of proteins that bind calcium using a DxDxDG motif (including EF-hand proteins such as calmodulin) (Lewit-Bentley & Rety, 2000). Recent work suggests that although some of these DxDxDG motifs probably share a common ancestor, some appear to have evolved independently (Rigden & Galperin, 2004).

Proteins with wholly different folds can convergently evolve the same 3D arrangement of residues in a metal binding site. The phrase "convergent evolution" is used here to indicate metal binding sites that do not merely share the same general geometry and residue preferences, but which also have the same (or very similar) residues in the same relative positions. Structural templates have the potential to act as a tool to identify cases of convergent evolution of metal binding sites.

There have previously been investigations into the convergent and divergent evolution of particular metal-binding motifs (such as zinc fingers) and protein families (such as the metallo-β-lactamases) (Wang *et al.*, 1999; Seibert & Raushel, 2005; Krishna *et al.*, 2003; Rigden & Galperin, 2004). There do not appear to have been any attempts to study the evolution of metal binding sites using structural templates, although one study represented metal binding sites with "molegos", which combine sequence and structural information about short motifs in the region of the metal (Schein *et al.*, 2005).

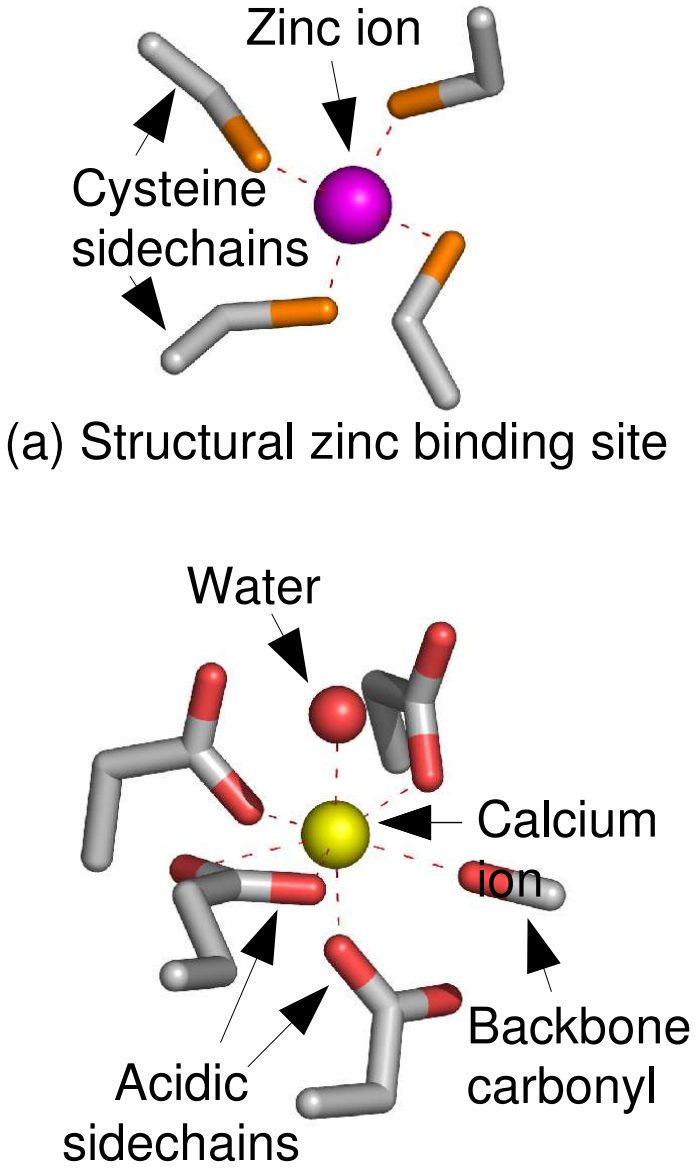


Figure 4.1: Examples of metal binding site structures.
The diagrams were created using Pymol (www.pymol.org). (a) Based on structure 1e7l (Raaijmakers *et al.*, 2001). (b) Based on structure 1g4y (Schumacher *et al.*, 2001).

Metals found in PDB structures may not be bound at biologically meaningful locations; where the site is biologically relevant, the metal which is bound may not be the one which is bound *in vivo* (the "cognate" metal). For these reasons, an analysis of metal binding sites benefits from a manually annotated dataset. Malcolm MacArthur has produced an unpublished manually annotated dataset of biologically relevant metal binding sites. This is based on a set of high-resolution crystal structures, in order to ensure that all the groups observed to be coordinating the metal in the structure genuinely play this role.

This dataset of high-resolution crystal structures was used as a basis for the analysis described in this chapter. The most common types of binding site in this dataset were zinc and calcium binding sites playing structural roles. Consequently, this chapter focuses on those binding site types.

The work described in this chapter looks at the evolution of these structural zinc and calcium binding sites using structural templates, among other techniques. It assesses the extent to which these metal binding sites vary in structure between relatives. It uses structural templates representing metal binding sites to identify groups of unrelated proteins that have convergently evolved common solutions for metal binding. Finally, it examines how frequently relatives of metal binding proteins lack the capacity to bind metal, and identifies the structural basis for this difference between relatives.

4.2 Results

4.2.1 Dataset

The analysis was based on a limited set of 11 calcium binding sites and 7 zinc binding sites that were known to be biologically relevant, taken from the scientific literature. Each of these 18 "literature entries" is described in Table 4.1. These literature entries were derived from a non-redundant set of proteins. In two cases, two separate and different metal binding sites from the same protein were used; these are called "sister" sites. These sister sites are distinguished from one another in this chapter using subscript letters: for example, $1ji1_a$ and $1ji1_b$. For each of the 18 metal binding sites, a set of family members with known structures was identified using PSI-BLAST. Each literature entry plus its

family members is known as a **metal site family**; the 11 calcium binding metal site families had 198 members, whilst the 7 zinc binding metal site families had 152 members. Table 4.1 describes these metal site families.

The number of families was relatively small because they had to meet several requirements, including checking against the scientific literature; see the methods section for full details. In brief, each literature entry had to be non-redundant with other literature entries in terms of their SCOP and CATH protein domain structural classifications, except sister sites; the metal binding sites had to be ones thought to occur *in vivo* (not artefacts of crystallisation); the metal present had to be the cognate metal; the sites had to include at least three non-sidechain residue contacts, in order to permit the creation of specific structural templates; sites could not contain ligands other than protein residues and water; and lastly, the metal site families had to have at least two member structures featuring the bound metal that were based on X-ray crystallography with a resolution better than 2.5 Å.

(Note that the term "ligand" is used in chemistry to indicate a group which coordinates a metal, and is used in biochemistry to indicate one molecule which binds another, particularly a small molecule binding a protein. In this chapter, the term "ligand" is used exclusively in the former sense.)

The calcium binding sites employed a variety of motifs to bind the metal; only one of them used a standard EF-hand motif (calmodulin, PDB entry 1g4y). Most of these calcium binding sites served structural roles. The only exception was calmodulin, where the calcium is regulatory; calcium binding induces a structural change in calmodulin which enables it to bind to a variety of target proteins, altering their activity.

There were four zinc binding site templates in the dataset that each consisted of four cysteines (PDB entries 1e7l, 1k3x, 1m2k and 1n8k). There were two zinc-binding templates that each consisted of three cysteines and one histidine (PDB entries 1btk and 1r5y). The remaining zinc binding site (PDB entry $1jk3_b$) consisted of three histidines and one glutamate. These zinc binding sites all served structural roles. Another important class of zinc binding sites serving structural roles is those which consist of two cysteines and two histidines (C2H2); these include nucleic-acid-binding proteins such the transcription factors TFIIIA, Zif268, and Sp1 (Krishna *et al.*, 2003; Brown, 2005). Unfortunately, the

PDB entry	Ion	Name	CATH	Family members analysed	Maximum sequence identity	Minimum sequence identity
1od3	Ca	Carbohydrate-binding module CsCBM6-3	2.60.120.260	4	100	100
1gk9	Ca	Penicillin acylase	1.10.439.10	24	100	63.5
1h80	Ca	Iota-carrageenase	2.160.20.10	2	100	100
$1ji1_a$	Ca	Alpha-amylase 1	2.60.40.10	5	99	99
$1ji1_b$	Ca	Alpha-amylase 1	2.60.40.10	6	99	31
$1jk3_a$	Ca	Elastase	3.40.390.10	35	98	56
1lyc	Ca	Peroxidase	1.10.520.10	55	100	42
1mct	Ca	Trypsin	2.40.10.10	35	100	65
1oyg	Ca	Levansucrase	-	2	99	99
1r0r	Ca	Subtilisin	3.40.50.200	60	100	61
1g4y	Ca	Calmodulin	1.10.238.10	22	100	87
1btk	Zn	Bruton's tyrosine kinase	2.30.29.30	3	99	98
1e7l	Zn	Endonuclease VII	1.10.720.10	2	99	99
$1jk3_b$	Zn	Elastase	3.40.390.10	53	98	49
1k3x	Zn	Endonuclease VIII	-	12	100	21
1m2k	Zn	Sir 2	3.40.50.1220	12	99	24
1n8k	Zn	Alcohol dehydrogenase	3.90.180.10	54	99	16
1r5y	Zn	tRNA-guanine transglycosylase	3.20.20.105	23	100	22

Table 4.1: Metal site family summary.
The column "PDB entry" gives the identifier for the literature entry for each family in the PDB. Note that there can be more than one metal site family based on a single PDB entry where that PDB entry has two or more structurally unrelated metal binding sites. When this is the case, the two sites are differentiated by being followed by a letter subscript, e.g. $1ji1_b$. The column "CATH" gives the classification number of the literature member's metal binding domain in the CATH structural classification (Pearl *et al.*, 2005). The number of metal site family members is limited to those that met all the criteria for inclusion described in the Methods section; it includes the original literature entry. The sequence identity figures apply to pairwise comparisons between the literature member for the family and each other family member.

only C2H2 structure that occurred in the manually annotated dataset of high-resolution structures used as the basis for the dataset used in this chapter was the structure with PDB entry 1llm, which was not analysed because it is a designed chimeric protein rather than a naturally occurring one (Wolfe *et al.*, 2003).

Two different types of structural template were created for each metal site family. The first represented residues in terms of the positions of their C_α and C_β atoms, and was therefore a reflection of backbone orientation. The second type represented each residue using only those atoms directly involved in liganding the metal. This allowed an investigation into how these different atom subsets change in structure between relatives. These templates resembled those described in Chapter 3 in all respects other than the atom subsets used.

4.2.2 Structural variation of metal binding sites

To understand the constraints on metal binding site construction, an analysis was carried out of the extent to which the binding sites vary within the metal site families, analogous to the analysis of structural variation in catalytic sites described in the previous chapter. This analysis involved calculating the atom coordinate root mean square deviation (RMSD) between each site and the site for the literature entry for that family. As in the rest of this chapter, only those protein structures where metal was present in the binding site were analysed.

Two potential causes of structural differences between metal binding sites were examined:

- Evolutionary divergence. This was analysed by using the percentage sequence identity between a given metal site family member and the appropriate literature entry as a measure of evolutionary divergence.

- Resolution of the crystal structure (low resolution relatives may differ structurally from the literature entry due to uncertainty in atom location).

The effects of evolutionary divergence and resolution on the RMSD of metal binding sites are summarised in Figures 4.2 and 4.3 respectively. The distribution of RMSDs can

most easily be seen in Figure 4.2, which shows how RMSD varies with sequence identity to the literature entry for all metal site families. There is relatively little variation in structure at any level of sequence identity, with the great majority of structures deviating less than 0.6 Å RMSD from the literature entry for their metal site family.

There are 11 unusual cases where the structural difference between the metal-binding atoms of two relatives is greater than 0.6 Å RMSD. Ten of these are from calcium-binding sites where there is an apparent flip of the amide group of an Asn or Gln such that the oxygen and nitrogen atoms are in the opposite positions. These all come from the metal site families with literature entries 1r0r and 1od3. The remaining case of structural difference greater than 0.6 Å is from a zinc-binding site (the site from PDB entry 1qic differs from literature entry $1jk3_b$), where there is a difference in position of one of the liganding histidines relative to the rest of the binding site, for reasons that are unclear (Pavlovsky *et al.*, 1999). The C_α/C_β atom comparisons include only one case of structural difference greater than 0.6 Å.

Calcium binding sites show more variation in the structure of the metal binding atoms than do zinc binding sites. However, this is largely accounted for by the outlying data points discussed above. For zinc-binding sites, there is a statistically significant correlation suggesting that the metal binding atoms of closer relatives are more structurally similar (Figure 4.2b; statistical significance threshold $\alpha = 0.05$; applying Bonferroni correction (Bonferroni, 1936) for the 8 correlation tests in this figure gives $\alpha = 6.25 \times 10^{-3}$). However, this correlation is extremely weak. There is no such correlation for the metal binding atoms from calcium binding sites (Figure 4.2a).

For the C_α/C_β atoms of zinc binding sites, there is a pronounced (and statistically significant) negative correlation between RMSD and sequence identity (Figure 4.2d). For the calcium binding sites, no such relationship is evident (Figure 4.2c). It is possible that a relationship is not evident because there happen to be few members of the calcium binding families in the current dataset which have levels of sequence identity to the literature entry below 50%. However, a relationship between these two variables is evident for zinc even when one only considers those family members with sequence identities above 50%, so it may be that there is some more fundamental difference between the calcium and zinc datasets in this regard. The lack of distant relatives of the calcium literature entries

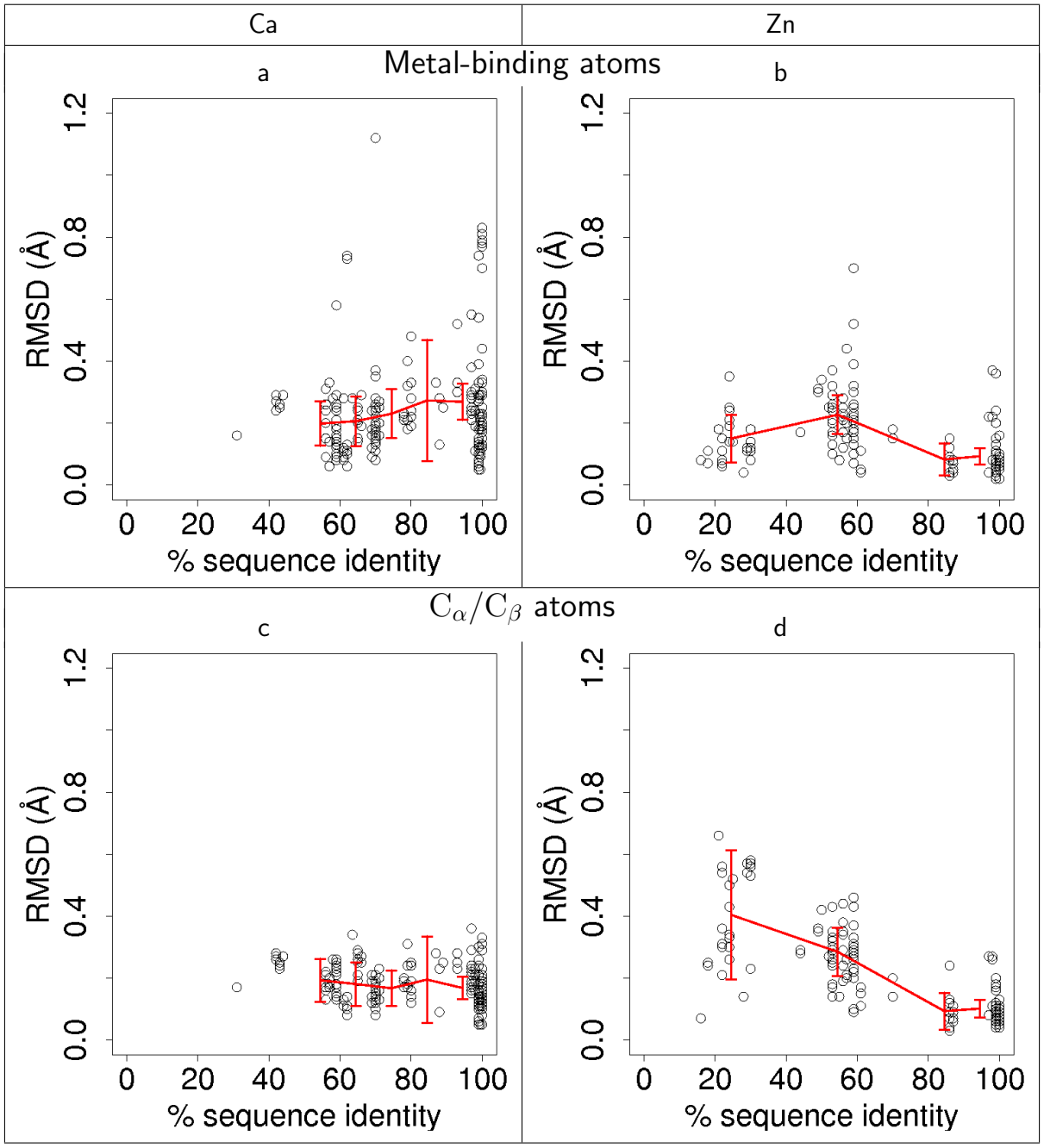


Figure 4.2: Effect of evolutionary divergence on metal binding site structure. Each point on one of the scatter plots corresponds to a single comparison between the literature entry for a given metal site family and one of that metal site family's members. There are 198 such comparisons for all calcium binding families and 152 for zinc. The line graphs show mean values for data points falling into evenly spaced bins. Error bars show 95% confidence interval for the mean. Bins for mean values are at 10% intervals. Bins with fewer than ten data points are not included in the line graphs. (a) Metal-binding atoms from calcium-binding sites. (b) Metal binding atoms from zinc-binding sites. (c) C_α/C_β atoms from calcium-binding sites. (d) C_α/C_β atoms from zinc-binding sites.

appears to be the reason why the RMSD levels for calcium binding sites are, on average, lower than those for zinc binding sites.

The resolution of calcium binding family members has no statistically significant effect on the extent to which they differ structurally from the literature entry; this is true both for metal binding atoms (Figure 4.3a) and C_α/C_β atoms (Figure 4.3c). For zinc binding families, however, there is a statistically significant correlation, albeit a very weak one, for both sets of atoms (Figures 4.3b and 4.3d). Large variations in structure (RMSD $>$ 0.4 Å) are rare when the resolution is better than 1.5 Å.

4.2.3 Water molecule structural variation compared to that of protein sidechains

The templates used in the analysis described above included water molecules where they were bound to the metal. Water molecules were found in eight of the 18 literature entry binding sites in the current dataset; all of these were calcium binding sites. In order to compare the structural variability of water to that of residues, those sites where water was present were represented using templates that lacked the waters, and RMSDs to metal site family members were calculated as described above.

The overall mean RMSD (mean of all means for individual metal site families) can be used to summarise the RMSDs over all metal site families. For metal-binding-atom templates, this overall mean was almost identical for water-containing templates and templates from which water had been removed: 0.32 Å. The presence or absence of water also had little effect on C_α/C_β templates; the RMSD was 0.17 Å with water present, and 0.15 Å without.

Alternatively, one can consider the structural variability of waters in terms of the distance of an individual water atom in a metal site family member from the equivalent water in the literature binding site when the two metal binding sites are optimally superposed. The mean distance for waters in metal-binding-atom templates was 0.28 Å (against 0.24 Å for non-water atoms); figures for C_α/C_β templates show a slightly larger separation (0.32 Å against 0.18 Å). For both template types, a t-test shows there was a statistically significant difference between water and non-water distances at the $\alpha = 0.05$ level. Thus,

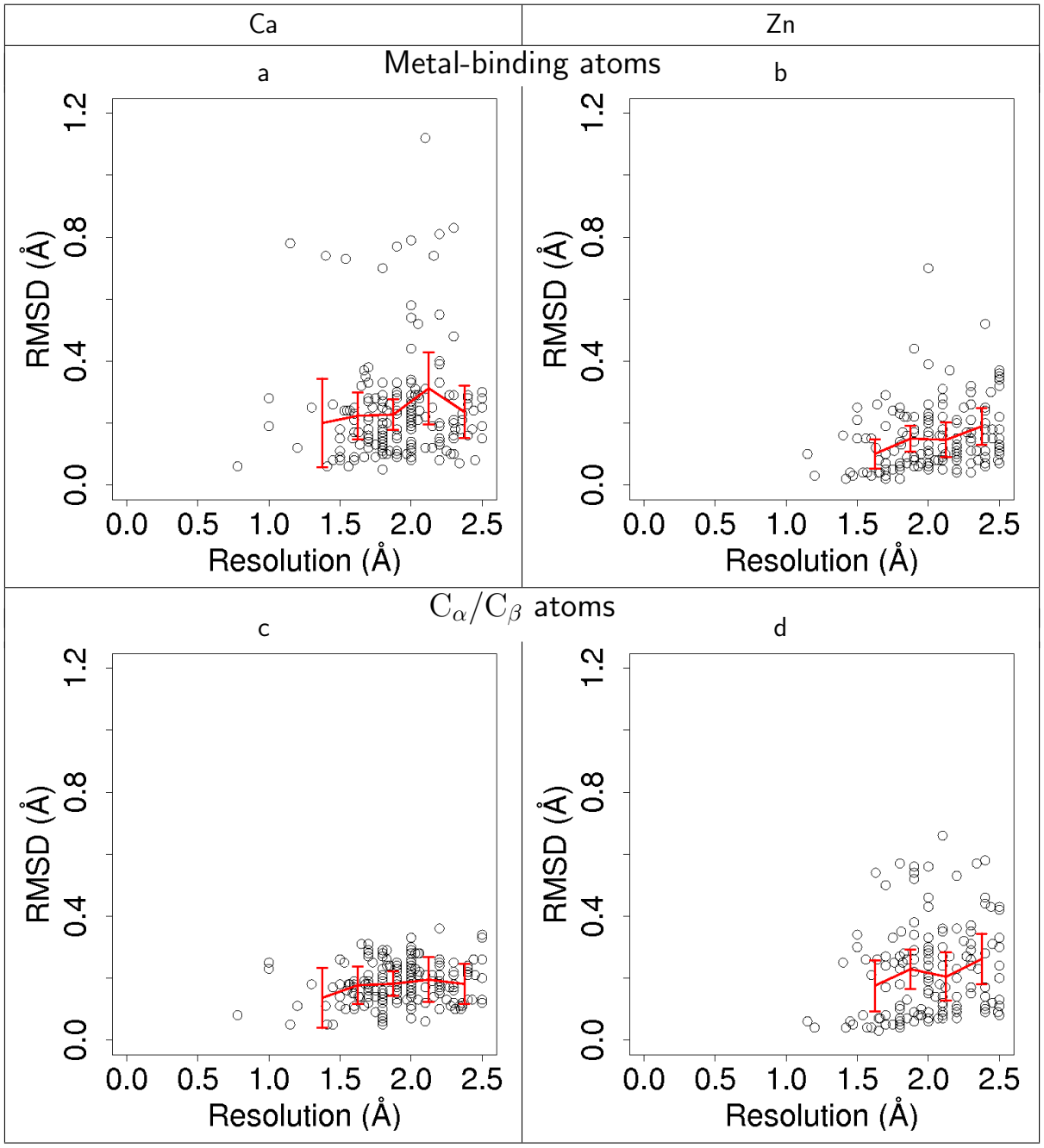


Figure 4.3: Effect of resolution on metal binding site structure.
Each point on one of the scatter plots corresponds to a single comparison between the literature entry for a given metal site family and one of that metal site family's members. There are 198 such comparisons for all calcium binding families and 152 for zinc. The line graphs show mean values for data points falling into evenly spaced bins. Error bars show 95% confidence interval for the mean. Bins for mean values are at 0.25 Å intervals. Bins with fewer than ten data points are not included in the line graphs. (a) Metal-binding atoms from calcium-binding sites. (b) Metal binding atoms from zinc-binding sites. (c) C_α/C_β atoms from calcium-binding sites. (d) C_α/C_β atoms from zinc-binding sites.

as expected, metal-binding waters were slightly more structurally variable than non-water ligands.

4.2.4 Structural template matches

Structural templates were used to detect similar metal binding sites in distant relatives, and in non-relatives. The results provide an insight into the structural conservation of metal binding sites over long evolutionary distances, and how often unrelated proteins develop the same solution for metal binding.

The analysis of structural template results presented here differs from the analysis of structural templates representing catalytic sites described in the previous chapter. The work in the previous chapter examined the ability of structural templates to recognise relatives of catalytic sites. The task of recognition of similar sites is less useful for these structural metal binding sites; a related structure where the metal is in place needs no recognition. Whilst it is possible for structures of metal binding proteins to lack the metal, this was rare for metal site family members in the current dataset, so it was not practical to investigate the ability of structural templates to detect these unoccupied sites. However, it is easier to recognise distant relatives and cases of convergent evolution for metal binding sites than for catalytic sites, since one merely needs to check whether the metal is present. For that reason, the analysis described below is more focused on analysing the different types of match achieved by structural templates, including matches to distant relatives and cases of convergent evolution.

For each metal site family, a template based on the literature entry was used to search a non-redundant subset of the PDB. Only the templates based on metal-binding atoms were used for this analysis, in order to focus on similarities in the first-shell coordination of metals. The templates themselves did not include the metal ions. After the template matches had been obtained, those matches with an RMSD better than 1.5 Å were checked for the presence of metal ions. If all residues were within 3 Å of the same metal ion, then that match was regarded as having a metal ion present.

Template matches to this non-redundant dataset can be divided into the following categories, each of which is assigned a letter, which is used for reference throughout this

section and in figure 4.4.

- **(a)** Members of the literature entry's metal site family: structures that had enough sequence similarity to the literature entry to be identified by a PSI-BLAST search.
- Structural relatives of the literature entry not identified by sequence-based PSI-BLAST searching. These belonged to the same CATH or SCOP homologous superfamily as the literature entry. These can be divided into:
  - **(b)** Cases where a metal was adjacent to all the residues matched.
  - **(c)** Cases where such a metal was not present.
- Non-relatives of the literature entry. These can be divided into:
  - **(d)** Matches without any metal present in the binding site. The great majority of these were probably random matches without biological meaning. These matches provide a "background distribution" showing how frequently the conformation of residues found in the template occurred at random.
  - **(e)** Matches with a metal present in the binding site that was different to the metal found in the literature entry. These matches show how often the same configuration of metal-binding residues was used to bind different metals. In some cases, these matches represented sites that naturally bind the same metal as that occurring in the literature entry, but which bound a different (non-biological) metal in the structure that was matched.
  - **(f)** Matches with the same metal present in the binding site as that which was present in the literature entry. These represented cases of convergent evolution.

Figure 4.4 shows the distribution of RMSDs for each class of match. Figure 4.4a shows the distribution of RMSDs within metal site families. This is the same RMSD information that was shown in Figure 4.2, presented in a different manner. As previously described, metal binding sites exhibited little structural variation between metal site family members. Most family matches had low RMSDs; there were very few family member matches with RMSDs above 0.5 Å. RMSDs for calcium were slightly higher than those for zinc.

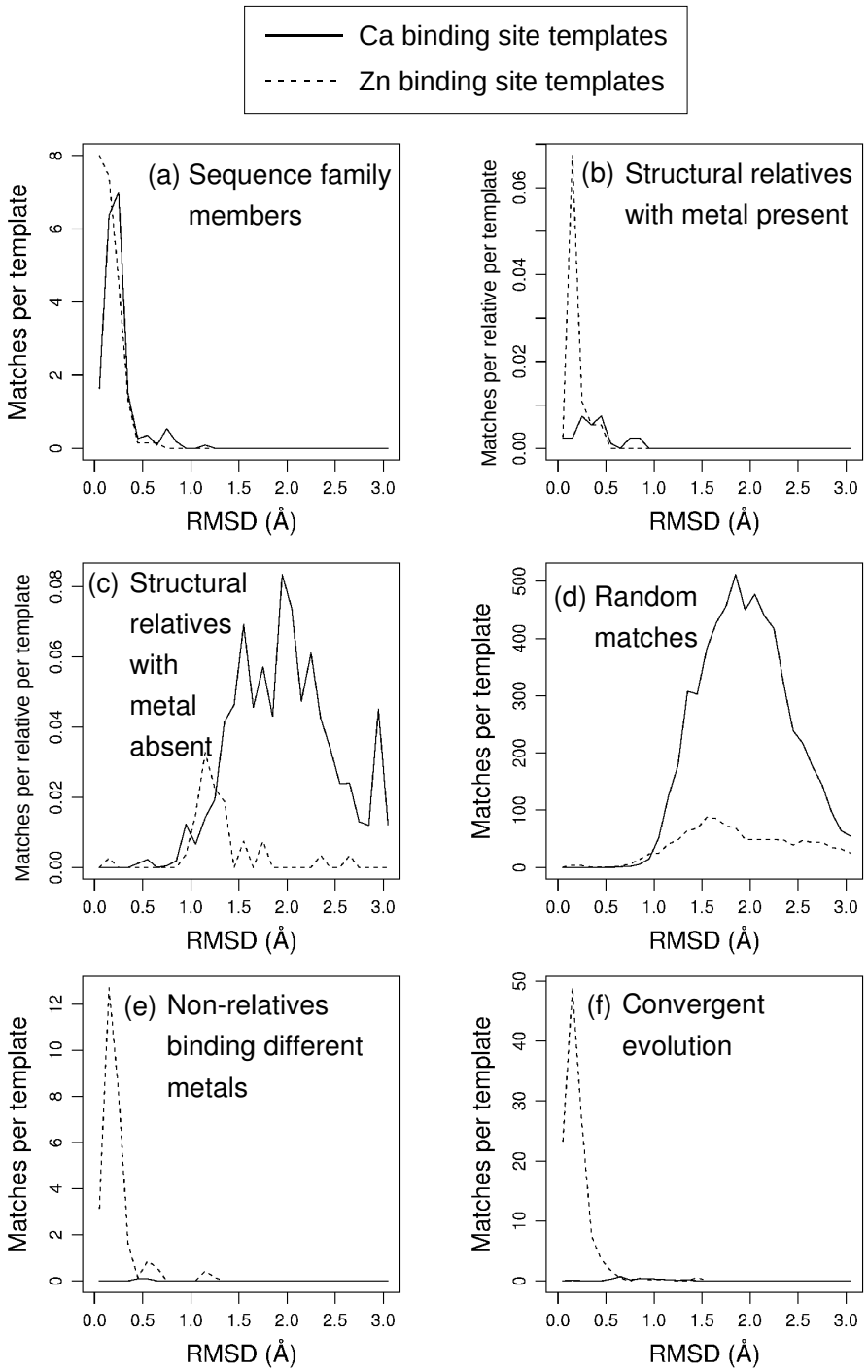


Figure 4.4: RMSD distribution of template matches.

Each histogram shows the distribution of RMSDs of matches to the structural templates. Note that the y-axis scale differs between plots. All frequencies are per template. For the plots relating to structural relatives (plots b and c), the frequencies for each template were normalised by the number of structural relatives of that template which were searched, before being normalised for the number of templates. (a) Matches of templates to members of their own metal site families. (b) Matches to non-relatives without metal present (random matches). (c) Matches to structural relatives with metal present. (d) Matches to structural relatives without metal present. (e) Matches to non-relatives with a different metal. (f) Matches to non-relatives with the same metal (cases of convergent evolution).

Most random matches differed from the template to a much greater extent than did the matches to family members; there was little overlap of the RMSD distributions of the two categories. In other words, the conformations of residues seen in metal binding sites occurred at random relatively rarely. The distribution of random matches peaked at 1.8 Å for zinc and 1.6 Å for calcium, with only a small proportion of random matches below 0.5 Å (Figure 4.4d). The reason that the frequency of random matches did not continue increasing indefinitely with higher RMSD was that the program used for template matching, Jess (Barker & Thornton, 2003), does not return matches where the inter-atom distances in the match differ from those in the template by more than 6 Å. Distributions with peaks at lower RMSD values than 1.8 Å probably represent genuine falling off in numbers of results with higher RMSD values. Although the shape of the random match distribution is similar for both calcium and zinc, there are many more random matches for calcium than for zinc. This is probably because the zinc sites include multiple cysteines; this is a relatively rare residue. By contrast, the calcium sites tend to include the more common aspartates, and glutamates, as well as the ubiquitous backbone carbonyls and water molecules.

There was a small population of "random" matches for calcium with RMSD below 0.5 Å. These may correspond to cases of convergent evolution where the metal happens to be absent in the structure matched, although this is difficult to confirm.

The matches to structural relatives where metals were present (Figure 4.4b) had an RMSD distribution very similar to that of family members. This suggests that metal binding sites in distant relatives remain as structurally similar as those in close relatives. There were 29 such matches for calcium, and 34 for zinc. This total is very small compared to the number of structures searched which were structural relatives of the family: 886 for calcium and 183 for zinc. The generally low RMSD of the matches suggests that the low total number of matches was not due to structural change. The main causes appear to be that:

- Many structures of metal-binding proteins do not have the metal present.
- Many structural relatives of metal-binding proteins do not bind to the same metals themselves. This point is discussed in more detail below.

- Some relatives use different residue types for metal binding, and thus cannot be detected by structural templates. This is particularly an issue for calcium binding sites, where a change from Asp/Glu to Asn/Gln sometimes occurs, but is not detected by the templates. More flexible templates would remedy this problem.

- Some relatives are falsely identified as not having metal present because the best template match is not a match to the metal-binding atoms.

It is difficult to assess the number of structures falling into each of these categories, although all of them occur in this dataset. The proportion of relatives matched was considerably smaller for calcium than for zinc. This is probably because several of the issues mentioned above are more likely to apply to calcium binding proteins than to zinc binding proteins. The zinc binding proteins in the current dataset have zincs that play an important role in protein structure, and it is therefore less likely that these metals will be absent in structures of relatives which bind zinc, and it is also less likely that there will be relatives that do not bind zinc *in vivo* (see the section on loss of metal binding through evolution below).

The matches to structural relatives which lack metals (Figure 4.4c) for calcium-binding templates showed a broadly similar RMSD distribution to the random matches, reaching a plateau above 1.6 Å. The equivalent distribution for zinc-binding templates peaks at the lower value of 1.1 Å, and has fewer matches per relative for each template. There are probably fewer zinc matches for this category for the same reason that there are fewer random zinc matches than random calcium matches. Most of the matches for zinc-binding templates have lower RMSDs than the calcium-binding template matches. These low-RMSD matches to zinc-binding templates almost all derive from a single zinc-binding template, $1jk3_b$. This template includes three histidines (each represented in the template by a single nitrogen atom), and an aspartate which binds metal in a monodentate manner. These non-metal matches correspond to a quirk of the behaviour of this particular template: the geometry of the template permits two nitrogen atoms in the template (derived from separate residues) to match to two separate nitrogen atoms in a single histidine. This results in a set of matches to metal binding sites with a different configuration, which are treated as sites with no metal present because one of these two nitrogen atoms in the

histidine is not directly involved in metal binding. All but three of the zinc matches in Figure 4.4c between 0.9 Å and 1.4 Å are of this type. The lack of a peak at similar levels of RMSD in other distributions suggests that this type of template match is unusual.

The matches made by zinc-binding templates to non-relatives binding different metals (Figure 4.4e) generally showed a similar RMSD distribution to the family members: it is possible for different metals to be bound by similar groups of residues in very similar conformations.

The great majority of the zinc-binding template matches to sites binding other metals are made by the four templates which each consist of four cysteines. These have a total of 190 matches, although in most cases, all four templates make the same match, and as a result there are only 50 unique PDB entries matched. Of these 50 unique PDB entries, there are 23 matches to sites binding iron, 20 for cadmium, four for mercury, and one each for gallium, nickel and silver. These are the metals present in the PDB files, and they may not be biologically relevant. Of the iron matches, 19 are to rubredoxin-like domains, where the iron is involved in electron transfer (Auld, 2004). The cadmium matches include 14 to metallothionein domains. The functions of metallothioneins include sequestering cadmium and other heavy metals in order to protect against their toxicity (Klaassen *et al.*, 1999). Cadmium and zinc have similar ligand preferences, and cadmium toxicity may be due to cadmium displacing zinc from binding sites (Dudev & Lim, 2003).

The zinc-binding template consisting of three histidines and one glutamate (literature PDB entry 1jk3$_b$) had two matches to iron binding sites, one to a cobalt binding site, and one to a nickel binding site. There are no matches made by the zinc templates which consist of three cysteines and one histidine.

There are only two matches between calcium-binding templates and sites binding other metals. The template with literature PDB entry 1mct matches a manganese binding site of uncertain function (Kumar *et al.*, 1996), and the template with literature PDB entry 1g4y matches a magnesium binding site, although it is possible that this latter site binds calcium *in vivo* (Le Du *et al.*, 2001).

The 790 matches to non-relatives binding the same metal (cases of convergent evolution) for zinc-binding templates also showed a similar RMSD distribution to the family members (Figure 4.4f). This suggests that this type of convergent evolution of metal

binding sites happens frequently, and that convergent evolution generally leads to very similar site structures. Convergent matches to calcium-binding templates did occur, but were much rarer (only 30 cases) and tended to have higher RMSDs than those observed for the zinc binding sites.

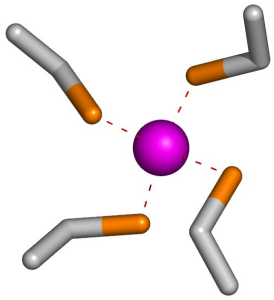
4.2.5 Convergent evolution

The results above show that unrelated proteins can use the same residues in the same geometry for metal binding. Where this convergent evolution occurs, it suggests that the conformation of residues used is particularly favourable for metal binding. Table 4.2 presents details of these archetypal residue conformations and the protein domains that have convergently evolved them, including diagrams of the metal binding sites.

Table 4.2: Convergent evolution.
Each table section represents a group of convergently evolved metal binding sites. Each section opens with a diagram showing the structure of the metal binding site of one of the literature entries. The table below this gives the details of the convergent groups, organised by the CATH code of the metal binding domain. The domains that are among the 18 families analysed in this chapter are at the top of the table, with other domains below, separated by a double horizontal line. Each row gives the CATH code for the domain (where it exists), a description of the domain (taken from CATH where possible; otherwise taken from SCOP), an example PDB code for that domain, and then the residues that form the metal binding site for the example case. Equivalent residues are given in the same column as one another. For each residue, the one-letter amino acid code is given followed by the residue numbering in the PDB file. Where the residue has a chain letter, this is given afterwards in brackets. Water molecules are not listed. Where a residue binds the metal by more than one atom, each atom is given a separate column, because other convergent domains may use two different residues for these contacts. Note that although the same CATH code can occur in different convergent groups, this does not mean that these sections should be assimilated to a single convergent group, because there can be multiple unrelated metal binding sites in a given CATH homologous superfamily.

Table 4.2 continued.

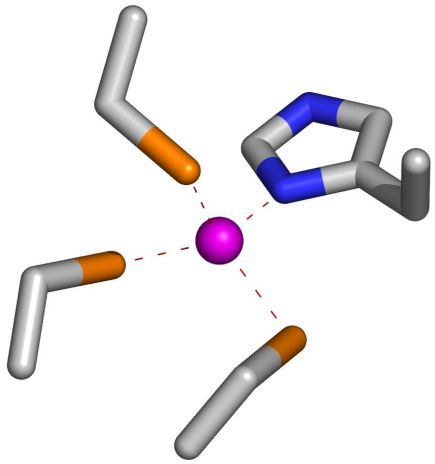
(a) Zinc: Cys-Cys-Cys-Cys sites



CATH code	Domain description	Example PDB	Residues			
3.90.180.10	Medium-chain alcohol dehydrogenases	1n8k	C97(A)	C100(A)	C103(A)	C111(A)
3.40.50.1220	TPP-binding domain	1m2k	C124(A)	C127(A)	C145(A)	C148(A)
-	Glucocorticoid receptor-like (DNA-binding domain)	1k3x	C237(A)	C240(A)	C257(A)	C260(A)
1.10.720.10	His-Me finger endonucleases	1e7l	C23(A)	C26(A)	C58(A)	C61(A)
1.10.10.60	Homeodomain-like	1ef4	C6 (A)	C9 (A)	C44 (A)	C43 (A)
2.10.110.10	Cysteine Rich Protein	1cxx	C147 (A)	C150 (A)	C168 (A)	C171 (A)
2.10.230.10	Chaperone	1exk	C17 (A)	C14 (A)	C67 (A)	C70 (A)
2.170.120.12	RNA Polymerase Alpha Subunit	1i50	C88 (C)	C92 (C)	C95 (C)	C86 (C)
2.170.150.10	Metal Binding Protein	1hxr	C23 (B)	C97 (B)	C26 (B)	C94 (B)
2.20.25.10	Zinc beta-ribbon	1pft	C27	C8	C11	C30
2.20.25.20	Casein kinase II beta subunit	1qf8	C140 (A)	C137 (A)	C114 (A)	C109 (A)
2.20.28.10	Rubredoxin-like	1h7v	C10 (A)	C46 (A)	C13 (A)	C43 (A)
2.20.28.20	Methionyl-tRNA synthetase (MetRS)	1qqt	C158 (A)	C148 (A)	C161 (A)	C145 (A)
2.30.170.10	Metallothionein	1jjd	C14 (A)	C47 (A)	C54 (A)	C52 (A)
2.30.30.20	Aspartate carbamoyltransferase	1d09	C138 (B)	C114 (B)	C141 (B)	C109 (B)
2.40.50.140	Nucleic acid-binding proteins	1ltl	C154 (A)	C132 (A)	C135 (A)	C157 (A)
2.60.11.10	Cytochrome C Oxidase	2occ	C85 (F)	C62 (F)	C60 (F)	C82 (F)
3.10.370.10	Transcription Factor TfIIh p44 Subunit	1e53	C348 (A)	C368 (A)	C345 (A)	C371 (A)
3.20.20.80	Glycosidases	1kwg	C150 (A)	C152 (A)	C155 (A)	C106 (A)
3.20.70.20	PFL-like glycyl radical enzymes	1h7a	C561 (A)	C564 (A)	C543 (A)	C546 (A)
3.30.160.60	Classic Zinc Finger	1rmd	C26	C46	C49	C29
3.30.40.10	Zinc/RING finger domain	1e4u	C53 (A)	C33 (A)	C56 (A)	C31 (A)
3.30.50.10	Erythroid Transcription Factor GATA-1	1by4	C1155 (A)	C1135 (A)	C1152 (A)	C1138 (A)
3.30.60.40	DNA-binding domain of intron-encoded endonucleases	1i3j	C151 (A)	C153 (A)	C164 (A)	C167 (A)
3.40.10.10	DNA Methylphosphotriester Repair Domain	1adn	C42	C72	C69	C38
3.40.50.300	P-loop containing nucleotide triphosphate hydrolases	1jr3	C59 (E)	C50 (E)	C62 (E)	C65 (E)
3.40.50.620	Nucleotidylyl transferase	1gax	C179 (A)	C347 (A)	C344 (A)	C176 (A)
3.40.50.720	NAD(P)-binding Rossmann-like Domain	1e3j	C96 (A)	C110 (A)	C99 (A)	C102 (A)
3.90.148.10	Adenovirus Single-stranded Dna-binding Protein	1anv	C396	C467	C398	C450
3.90.530.10	Nucleotide Excision Repair Protein Xpa (Xpa-mbd)	1d4u	C8 (A)	C29 (A)	C32 (A)	C11 (A)
4.10.10.10	Metallothionein Isoform II	4mt2	C7	C26	C15	C13
4.10.240.10	CD2-Gal4	2alc	C49 (A)	C42 (A)	C12 (A)	C39 (A)
4.10.830.10	30s Ribosomal Protein S14	1hr0	C40 (N)	C43 (N)	C24 (N)	C27 (N)

Table 4.2 continued.

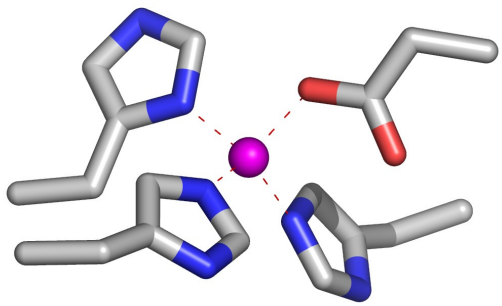
(b) Zinc: His-Cys-Cys-Cys sites



CATH code	Domain description	Example PDB	Residues			
3.20.20.105	tRNA-guanine (tRNA-G) transglycosylase	1r5y	H349(A)	C318(A)	C323(A)	C320(A)
2.30.29.30	Pleckstrin-homology domain (PH domain)/Phosphotyrosine-binding domain (PTB)	1btk	H143(A)	C154(A)	C155(A)	C165(A)
1.10.1170.10	Inhibitor Of Apoptosis Protein (2mihbC-IAP-1)	1g3f	H320 (A)	C327 (A)	C300 (A)	C303 (A)
1.20.1280.50	RING/U-box	1ldk	H1082 (C)	C1053 (C)	C1068 (C)	C1056 (C)
2.10.110.10	Cysteine Rich Protein	1cxx	H141 (A)	C144 (A)	C120 (A)	C123 (A)
2.170.220.10	Nucleotidylyl transferase	1a8h	H147	C130	C127	C144
2.20.25.10	Zinc beta-ribbon	1dl6	H18 (A)	C34 (A)	C15 (A)	C37 (A)
2.40.10.10	Trypsin-like serine proteases	2hrv	H114 (B)	C112 (B)	C54 (B)	C52 (B)
2.60.40.720	p53-like transcription factors	1tsr	H179 (C)	C176 (C)	C242 (C)	C238 (C)
3.30.160.60	Classic Zinc Finger	1rmd	H43	C64	C61	C41
3.30.40.10	Zinc/RING finger domain	1vfy	H203 (A)	C176 (A)	C200 (A)	C179 (A)
3.30.60.20	Cysteine-rich domain	1tbn	H102	C135	C151	C132
3.30.710.10	Potassium Channel Kv1.1	3kvt	H75 (C)	C81 (A)	C102 (C)	C103 (C)
3.40.960.10	Endonuclease	1cw0	H71 (A)	C117 (A)	C73 (A)	C66 (A)
3.90.148.10	Adenovirus Single-stranded DNA-binding Protein	1anv	H286	C339	C284	C355
3.90.430.10	Activator Of Metallothionein 1	1co4	H25 (A)	C23 (A)	C14 (A)	C11 (A)
3.90.580.10	DNA Primase	1d0q	H43 (B)	C40 (B)	C64 (B)	C61 (B)
3.90.75.10	Homing Intron 3 (I-ppo) Encoded Endonuclease	1a73	H134 (A)	C138 (A)	C132 (A)	C125 (A)

Table 4.2 continued.

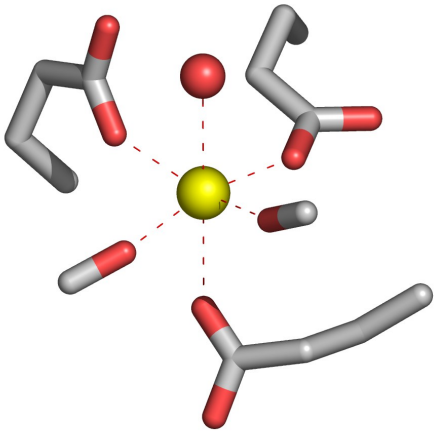
(c) Zinc: Asp-His-His-His sites



CATH code	Domain description	Example PDB	Residues			
3.40.390.10	Collagenase (Catalytic Domain)	1jk3	H168(A)	D170(A)	H183(A)	H196(A)
3.20.20.70	Aldolase class I	1dos	H110 (A)	E174 (A)	H226 (A)	H264 (A)
3.60.15.10	Metallo-beta-lactamase	1qh5	H54 (A)	D134 (A)	H110 (A)	H56 (A)

Table 4.2 continued.

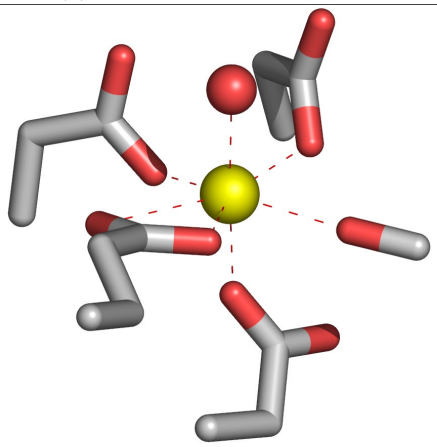
(d) Calcium: trypsin-like sites



CATH code	Domain description	Example PDB	Residues				
2.40.10.10	Trypsin-like serine proteases	1mct	E70(A)	N72(A)	V75(A)	E77(A)	E80(A)
1.10.238.10	EF-hand	1qls	E36 (A)	K31 (A)	N28 (A)	E36 (A)	D26 (A)
1.50.10.10	Six-hairpin glycosidases	1clc	D361	D401	S358	D361	D362
2.10.25.10	Laminin	1nzi	D116 (A)	F135 (A)	G138 (A)	D116 (A)	E119 (A)
2.130.10.10	YVTN repeat-like/Quinoprotein amine dehydrogenase	1qni	D221 (A)	Y204 (A)	T215 (A)	E207 (A)	E207 (A)
2.60.120.200	Concanavalin A-like lectins/glucanases	1h30	D656 (A)	R440 (A)	E331 (A)	D329 (A)	D329 (A)
2.60.40.1190	CBD9-like	1i82	D154 (A)	A155 (A)	V62 (A)	D74 (A)	D60 (A)
3.20.20.80	Glycosidases	1ava	D142 (A)	F143 (A)	A146 (A)	D148 (A)	D127 (A)
3.40.390.10	Collagenase (Catalytic Domain)	1hfs	D107	D182	E184	D182	D107
3.40.50.1820	Alpha/beta-Hydrolases	1rp1	D195	E187	R190	D192	D195

Table 4.2 continued.

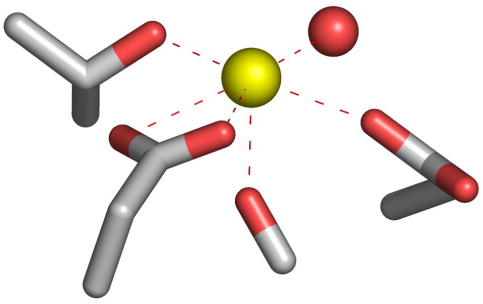
(e) Calcium: EF-hand-like sites



CATH code	Domain description	Example PDB	Residues					
1.10.238.10	EF-hand	1g4y	D20(R)	D22(R)	D24(R)	T26(R)	E31(R)	E31(R)
3.10.100.10	Mannose-Binding Protein A	1b08	D2330 (C)	E2301 (C)	D2324 (C)	E2329 (C)	D2297 (C)	D2297 (C)

Table 4.2 continued.

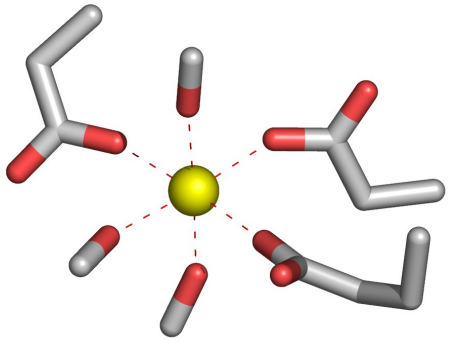
(f) Calcium: iota-Carrageenase-like sites



CATH code	Domain description	Example PDB	Residues				
2.160.20.10	Single-stranded right-handed beta-helix	1h80	T438(A)	D445(A)	D445(A)	Y446(A)	D447(A)
3.10.100.10	Mannose-Binding Protein A	1j34	S241 (B)	E320 (B)	E320 (B)	S241 (B)	E247 (B)

Table 4.2 continued.

(g) Calcium: collagenase-like sites



CATH code	Domain description	Example PDB	Residues					
3.40.390.10	Collagenase (Catalytic Domain)	1jk3	D175(A)	G176(A)	G178(A)	I180(A)	D198(A)	E201(A)
1.10.238.10	EF-hand	1scm	D22 (C)	G23 (C)	D22 (C)	A27 (C)	D19 (C)	D25 (C)
2.60.40.10	Immunoglobulins	1ji1	D42 (A)	N6 (A)	D42 (A)	A2 (A)	D4 (A)	D96 (A)
3.40.50.200	Subtilisin-like	1thm	D5	V82	I89	T87	D47	D47

As mentioned above, all of the zinc binding sites in the dataset played a structural role. Structural zinc binding sites typically consist of cysteine and histidine residues, with a total of four ligand residues in a tetrahedral arrangement; six of the seven single-zinc binding sites in the current dataset were of this type. These Cys/His zinc binding sites have been well studied, and are known to have convergently evolved on a large number of occasions (Krishna *et al.*, 2003). A structural classification of these sites is available, based on the secondary structural elements associated with binding (Krishna *et al.*, 2003). Catalytic zinc binding sites also exist; these differ from structural zinc binding sites in their preferred geometry and binding residues. There were no catalytic zinc binding sites in this dataset.

The four zinc binding templates which each consisted of four cysteines (literature PDB entries 1e7l, 1k3x, 1m2k and 1n8k; Table 4.2a) matched 29 different CATH homologous superfamilies, implying that this type of zinc binding site has independently evolved at least 33 times. All the templates had matches with the same set of CATH families. The two zinc-binding templates that each consisted of three cysteines and one histidine (literature PDB entries 1btk and 1r5y; Table 4.2b) matched 16 CATH homologous superfamilies. Again, both templates converged with the same set of families as one another.

The remaining zinc binding site (literature PDB entry $1jk3_b$) consisted of three histidines and one glutamate (Table 4.2c). The presence of acidic residues is more typical of catalytic zinc binding sites and, indeed, the template search identified convergent evolution with two catalytic sites. It is interesting that despite having similar conformations (RMSD < 0.7 Å), not only do these sites perform different functions, but also the 1jk3 site is a single zinc site, whereas the sites it structurally matches are a part of two-zinc sites.

There were cases of convergent evolution of calcium binding sites (Table 4.2d-g), but they were considerably rarer than those for zinc (Figure 4.4f). Most of the calcium binding sites had one or no cases of convergent evolution as recognised by the templates. The exceptions were the families with literature entries $1jk3_a$ (human elastase; Table 4.2g) and 1mct (porcine trypsin; Table 4.2d). The nine homologous superfamilies that matched 1mct include the family of literature entry $1ji1_a$ (*Thermoactinomyces vulgaris* alpha-amylase 1). The pattern formed by the template for 1mct thus appears to be relatively common;

it consists of three oxygens from acidic sidechains, two backbone oxygens, and one water. These are typically arranged with the acidic sidechains on one side of the ion, and the backbone oxygens and water on the other side. The calciums in these convergent groups generally played a structural role, as do most calcium ions.

One probable case of convergent evolution of calcium binding sites has been discussed by Rigden & Galperin (2004). They examine similarities between instances of the canonical EF-hand calcium binding motif (featuring a DxDxDG sequence) that occur in apparently unrelated domains. Rigden *et al.* identify 13 DxDxDG families. Only one of these coincides with one of the 11 calcium binding families analysed in this chapter: the metal site family whose literature entry PDB code is 1g4y (rat calmodulin). This metal site family does have a single convergent evolution match based on the template search (Table 4.2e), but this match does not correspond to any of the domains in Rigden's dataset. Of Rigden's remaining twelve families, nine have mutations relative to 1g4y, mostly Asp to Asn. These mutations make it impossible for the structural template to detect the relatives. In two more of Rigden's families, the template almost matches the site, but picks up the wrong oxygen atoms from the carboxyl groups of acidic residues that bind the substrate in a monodentate manner. The final one of Rigden's families has no structures available with metal present, so the template search naturally fails to match it. This suggests that the templates could benefit from refinements such as permitting matching between acidic residues and amide residues, and measures to deal with multiple possible matches to carboxyl oxygens.

4.2.6 Metal loss over evolution

In the course of evolution, a metal binding site can be lost entirely from a protein, with consequent changes to the protein's structure and function. This section of the analysis examines how frequently such losses occurred, what the structural changes underlying the loss of metal binding were, and how the protein function changed when a metal was lost.

Note that where the term "loss of metal binding" is used, it denotes loss relative to the original literature entry; it does not imply that the common ancestor of the two proteins is known to bind metal. This investigation did not look into the ancestral state; this would

probably not be practical at the superfamily level, because relationships are so distant that it is difficult to establish a robust phylogeny.

Loss of metal binding occurs frequently in distant relatives. Loss was observed in CATH homologous superfamily relatives in 9 of the 11 calcium-binding metal site families, and in 4 of the 7 zinc-binding metal site families. These CATH homologous superfamilies can be subdivided into families which have at least 35% sequence identity with one another (CATH S35 families). Of the S35 families relating to calcium-binding sites, 207 consisted entirely of metal binding members, 85 consisted entirely of non-binding members, and 7 included some members that did bind metal and some that did not. The S35 families relating to zinc-binding sites either consisted entirely of metal binding members (40) or entirely of non-binding members (22). It can be seen from this that S35 families tend to consist either of all metal-binding members or of all non-binding members.

4.2.7 Structural basis and functional consequences of metal loss

Loss of metal binding results from structural changes in the protein: either point substitutions of the metal-binding residues, or loss of that part of the protein which binds metal. The structural cause of loss of metal binding was investigated for each of the 13 metal site families where loss occurred among homologous superfamily relatives. It is possible for further structural changes to accumulate over the course of divergent evolution; for instance, the ability to bind metal might be lost by substitution of individual residues, and then subsequently the secondary structural elements that had been responsible for metal binding might be lost. For this reason, the immediate structural reason for loss of metal binding was examined by considering the closest relative to the literature entry that did not bind metal (Figure 4.5).

For calcium binding sites, the majority of cases of metal loss were accompanied by residue substitution. An example of calcium loss by substitution is shown in Figure 4.6a. The structure with calcium is rat calmodulin (PDB entry 1g4y (Schumacher *et al.*, 2001)); the structure without calcium is human p11 (PDB entry 1a4p (Rety *et al.*, 1999)). The calcium in calmodulin plays a regulatory role; p11, which lacks the calcium, resembles a "permanently active" form of a calcium binding regulatory protein (Rety *et al.*, 1999).

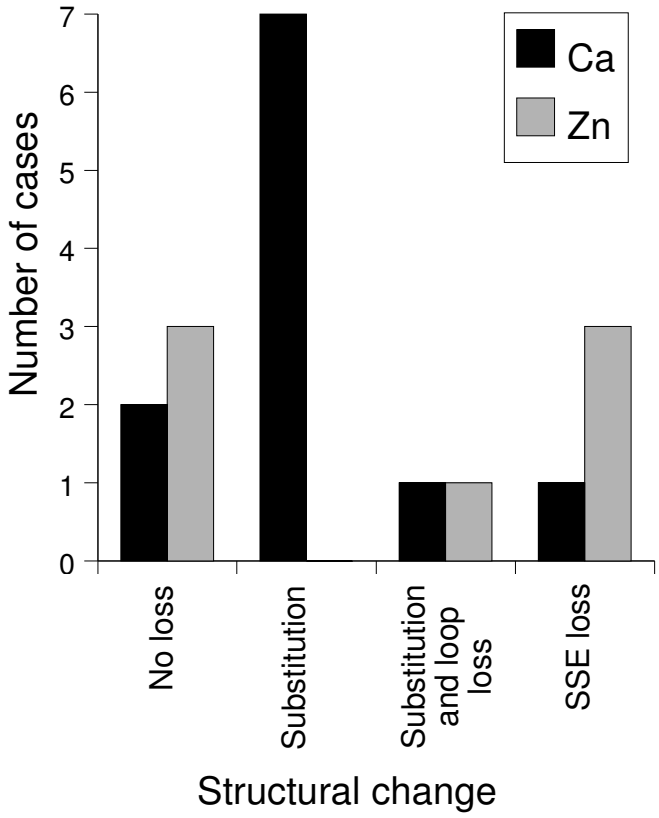


Figure 4.5: Structural changes accompanying metal loss.
The graph shows the type of structural change observed at the metal binding site in the closest relative that does not bind metal. Structural changes were divided into residue substitution, substitution combined with loop loss, and loss of one or more secondary structural elements associated with metal binding. Where all members of the CATH family have metal present at the binding site, this is recorded as "no loss". The results are broken down by the type of metal bound.

For zinc binding sites, a greater proportion of families had no members that did not bind metal. Where metal was lost, this was always accompanied by the loss of secondary structural elements or loops. An example of zinc loss accompanied by secondary structural element loss is shown in Figure 4.6b. The structure with zinc is horse alcohol dehydrogenase (PDB entry 1n8k (Rubach & Plapp, 2003)); the structure without zinc is *Escherichia coli* quinone oxidoreductase (PDB entry 1qor). The functional significance of the lack of the zinc binding loop in quinone oxidoreductase is not clear (Thorn *et al.*, 1995).

Most of the metals in this dataset serve a structural role. In some cases, the loss of the

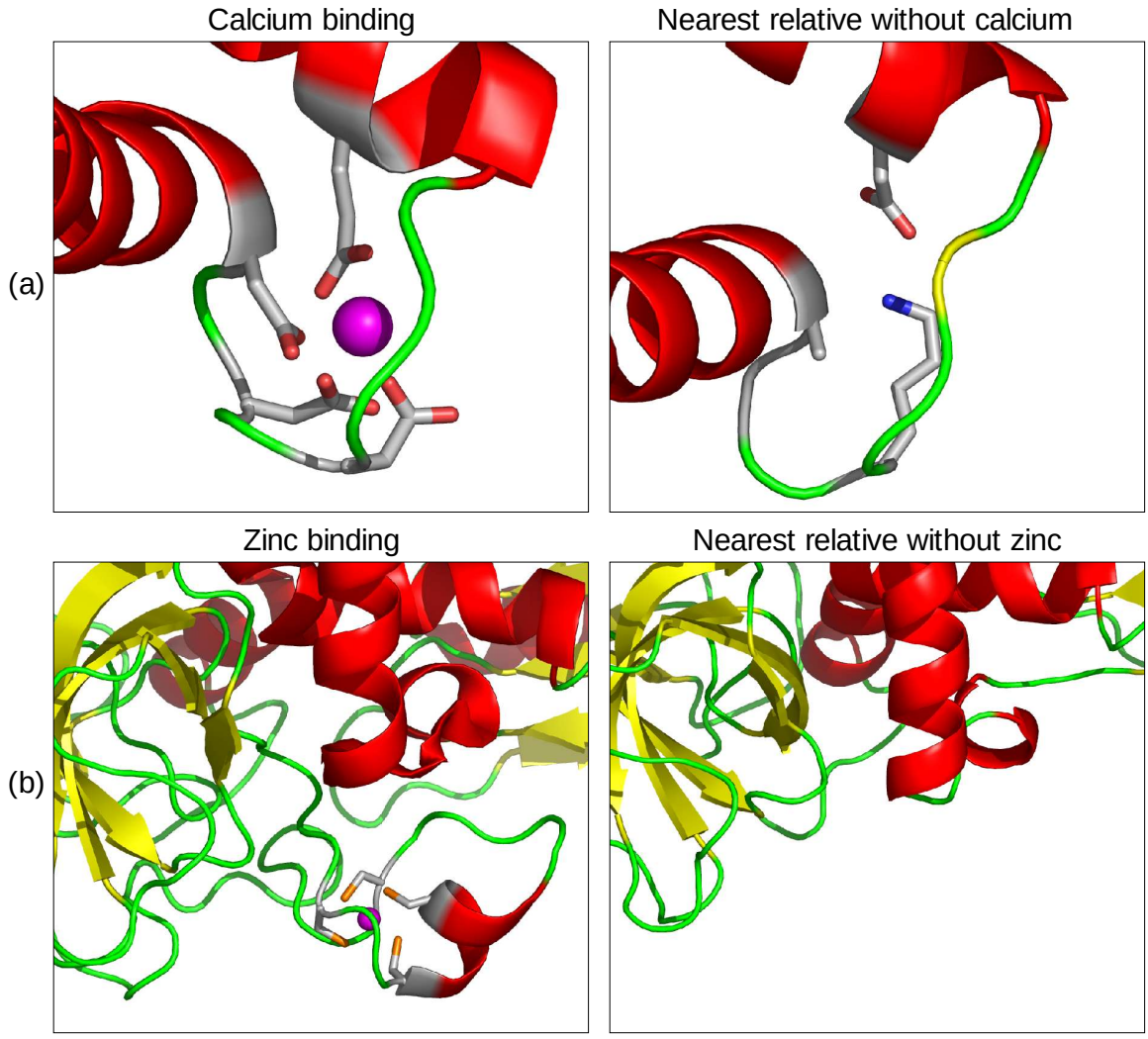


Figure 4.6: Examples of structural changes accompanying metal loss.
Residue sidechains binding metal (or their equivalents in relatives not binding metal) are shown as sticks. The diagram was created using Pymol (www.pymol.org). (a) Calcium loss due to residue substitution. The structure with calcium is PDB entry 1g4y (Schumacher *et al.*, 2001) (rat calmodulin); the structure without calcium is PDB entry 1a4p (Rety *et al.*, 1999) (human p11). (b) Zinc loss due to secondary structural element loss. The structure with zinc is PDB entry 1n8k (Rubach & Plapp, 2003) (horse alcohol dehydrogenase); the structure without zinc is PDB entry 1qor (Thorn *et al.*, 1995) (*Escherichia coli* quinone oxidoreductase).

metal was associated with loss of the surrounding secondary structural elements and/or loops whose structure the metal had stabilised. However, in some cases, a structural metal was lost, yet the protein region which it stabilised was retained. How is this possible? In several cases, although the only change at the points where the metal was bound was a residue substitution, there was considerable loss of secondary structural elements lying immediately adjacent (families 1gk9, $1ji1_b$, $1jk3_a$). In other cases, the structural change appeared very small (families $1ji1_a$, 1lyc, 1mct, 1od3).

There was one case in this dataset where the metal had a regulatory function (calmodulin, family 1g4y). Here the closest relative without metal, 1a4p, appeared to serve as a permanently activated form of the regulatory protein, as mentioned above (Rety *et al.*, 1999).

4.3 Discussion

There was very little variation in calcium or zinc binding site structure between sequence relatives, either for metal-binding atoms or C_α/C_β atoms. Evolutionary divergence and resolution mostly had little effect on binding site structure. This suggests that the chemistry of metal binding strongly constrains the positions of the ligand residues, as implied by studies of the geometry of metal binding in small-molecule structures (Harding, 1999, 2000). The exception was that the C_α/C_β atoms of zinc binding sites differed in structure more between relatives that were more distantly related. Evidently it is possible for the protein backbone to alter conformation over the course of evolution (shifting the C_α/C_β atoms) whilst leaving the metal-binding atoms in the same locations due to a reorientation of the sidechains. It is not clear why this same relationship was not evident for C_α/C_β atoms in calcium binding sites; it may simply be that the pattern was not evident because the calcium-binding family members had only a limited range of sequence identities to their literature members.

Because metal binding sites are structurally conserved, structural templates are able to recognise equivalent metal binding sites in distant relatives, provided that these relatives use the same set of residues for metal binding. One could also approach this problem using protein structure alignment programs such as SSM (Krissinel & Henrick, 2004) or

CE (Shindyalov & Bourne, 1998); it would require further analysis to determine which approach is most useful. The two may be complementary.

The templates recognised large numbers of instances of convergent evolution in both zinc and calcium binding sites. The zinc sites had more cases of convergent evolution; they were especially common among the four-cysteine sites. Cysteine is preferred by zinc at sites serving a structural role, accounting for the greater frequency of four-cysteine sites compared to Cys-Cys-Cys-His sites (Auld, 2004). The ability of templates to consistently recognise these sites is clearly a reflection of the sites' simplicity. A tetrahedral binding geometry is used for most sites, and this binding varies little from an ideal tetrahedral geometry. Though the convergent evolution of these structural zinc sites is well known (Krishna *et al.*, 2003), the use of structural templates is a convenient means of collecting together instances and quantifying their structural differences.

Cases of convergent evolution of calcium binding sites were less frequent. A number of calcium binding sites were matched by a template based on trypsin (literature PDB entry 1mct). This may reflect a preference for the calcium binding conformation that occurs in this site. However, this is a relatively small template (six ligand atoms), which is capable of partial matches to larger sites, and which can match to acidic sidechains binding in either a bidentate or monodentate manner. This flexibility in the template may contribute to the large number of matches.

The small total number of cases of convergent evolution of calcium binding sites is partly a reflection of the fact that calcium binding sites can be constructed from oxygen atoms from a number of sources: acidic sidechains, amide sidechains, hydroxyl groups from serine and threonine sidechains, backbone carbonyl groups, and water molecules. The lack of template matches implies a lack of patterns in the way that these building blocks can be assembled, except perhaps a tendency to use trypsin-like binding sites. This is probably also the reason that there are few matches between calcium-binding templates and binding sites for metals other than calcium.

The small number of calcium-binding template matches representing convergent evolution may also be a reflection of the nature of the structural templates employed. More flexible structural templates would permit similarities to be recognised between more diverse binding sites. The analysis of template matches suggested that a larger number of

distant relatives of calcium-binding sites might be matched if the templates permitted matching between oxygen atoms in Asp/Glu and Asn/Gln; this is also suggested by the comparison with the analysis of DxDxDG motifs by Rigden & Galperin (2004). More generally, templates that looked for a given configuration of oxygen atoms around calcium ions (without requiring specific residue types) would almost certainly obtain more matches. There is a danger, however, that such templates might prove so generic that they would match all calcium binding sites with a given coordination number. The current templates permit matching of atoms from different residues (in the template) to different atoms within a single residue (in the target); this behaviour led to a small number of misleading matches for the template with literature PDB entry $1jk3_b$. It might be useful to prevent this type of template match, although it appears to only affect a small number of cases. The template matching was carried out by Jess, which imposes a limit (6 Å in this study) on the maximum difference in inter-atom distances between the template and the target. Whilst this limit could be increased, it seems unlikely that any essentially similar pair of metal binding geometries would have this large a difference in inter-atom distances.

Metal-binding proteins often have relatives that lack the equivalent metal binding site. These are usually distant relatives; there were very few cases where proteins with more than 35% sequence identity to one another differed in whether they bound metal. Similarly, the comparison of metal binding in different proteomes by Dupont *et al.* (2006) found that in the SCOP classification, around half of those superfamilies (roughly equivalent to CATH homologous superfamilies) with any metal-binding members also contained members which did not bind metal, yet only 10% of SCOP families (which tend to have sequence similarity levels above 30% (Murzin *et al.*, 1995)) with metal-binding members also had non-metal-binding members.

The structural basis of loss of metal binding differs between calcium and zinc ions. Calcium loss is generally a result of substitution of the metal-binding residues. Zinc loss occurs more rarely, and when it occurs it is generally a result of deletion of the part of the protein responsible for metal binding. The explanation for these differences may be that structural zinc ions are frequently necessary for protein folding, so that the ability to bind metal cannot be lost without significant change to the protein structure.

A more comprehensive assessment of patterns of metal loss and gain could be made by establishing a phylogeny of protein structures within superfamilies. However, relationships between superfamily members are frequently only discernible at the structural level. This makes it difficult to establish a phylogeny by conventional sequence-based means, although some structure-based phylogenies are available in the PALI (Balaji *et al.*, 2001) and PASS2 (Bhaduri *et al.*, 2004) databases.

The work described in this chapter identified archetypal metal binding sites that have been evolved independently in unrelated proteins. Zinc binding sites have convergently evolved on a large number of occasions. The more complex binding sites for calcium have also convergently evolved, but more rarely. It seems that, in the course of evolution, proteins have explored many different solutions for calcium binding; although chemistry sets the rules for residue and geometry preferences, biology samples relatively freely within those parameters. Once a metal binding site has evolved in a given protein fold, its residue conformation remains constant even when much of the protein sequence has changed. Furthermore, metal binding sites are not easily lost in the course of protein evolution. However, such loss is seen in a variety of metal binding families. Where the metal is critical to the protein structure, its loss entails major structural changes.

4.4 Methods

4.4.1 Non-redundant set of metal site families

The dataset was based around a set of calcium or zinc binding structures whose metal-binding residues were recorded by Malcolm MacArthur based on information from the scientific literature. These "literature entries" all had a resolution of 1.6 Å or better; these high-resolution structures were selected using the PISCES server (Wang & Dunbrack, 2003). Only residues or waters within 3 Å of the metal qualified as metal-binding.

If there were two or more separate and unrelated metal binding sites within a given literature Protein Data Bank (PDB (Berman *et al.*, 2007)) entry, these were treated as separate literature entries.

Proteins of known structure that were related to these literature entries were identified

by using PSI-BLAST (Altschul *et al.*, 1997) to search the sequences of all proteins in the PDB, using a cut-off E-value of 5×10^{-4} for inclusion of a sequence in the profile and the final results, and 20 iterations. The resulting sequence alignment was used to identify the metal-binding residues of these relatives. This assignment of metal-binding residues was confirmed by checking that the specified residues were within 3 Å of the metal, and that the metal bound was the same as that in the literature entry. In this chapter, a literature-based entry plus its sequence relatives identified using PSI-BLAST are referred to as a *metal site family*.

The execution of PSI-BLAST and identification of these relatives was carried out by Craig Porter using scripts derived from those he created for identifying homologous entries in the CSA; the remaining aspects of assembling the dataset described below were carried out by the present author.

Metal site families were only used in this analysis if they met the following criteria:

- They had members other than the literature entry
- They had at least three residues that ligate metal via their sidechains (because templates consisting of only two residues return very large numbers of matches and can seldom discriminate metal site family matches from random ones)
- The bound metal in the literature entry was thought to be present *in vivo*, and was known to serve a structural role
- The metal was only bound by residues and water; no non-protein ligands other than water were involved. This excludes groups such as haem or molybdopterin

Individual family members were only included in this study if:

- They had the same residue type as the parent literature entry for all metal-binding residues, with the exception of residues that liganded metal via backbone atoms. Additionally, Ser was allowed to match Thr; Glu to match Asp; Gln to match Asn.
- They had the same number of liganding residues as the literature entry (though individual residues were permitted to change from monodentate to bidentate binding)

- They bound the same metal as the literature entry, and this metal was present in the structure
- They were based on X-ray crystallography, not NMR. This was because the resolution of NMR-based structures is usually lower than that of structures based on X-ray crystallography, and this can distort the binding site geometry.
- Their resolution was better than 2.5 Å.

All analyses in this chapter used a non-redundant set of metal site families. Because the original dataset obtained from the PISCES server had some redundancy at the homologous superfamily level, it was necessary to eliminate some families in order to achieve the non-redundant set of families. Metal site families were regarded as redundant with one another if their literature entries shared a CATH (Pearl *et al.*, 2005) or SCOP (Murzin *et al.*, 1995) code, or if the two metal site families had members in common. The exception to this rule was where two metal site families were based on different metal binding sites in a single PDB entry. Both sites were included unless they appeared to be related to one another through an ancestral domain duplication. Where a pair of metal site families was redundant with one another, priority was given to metal site families based on PDB entries that had larger numbers of structurally unrelated metal binding sites. After that, priority was given to the metal site family containing the larger number of structures.

4.4.2 Structural templates

Two banks of structural templates were constructed from the literature entries: one bank of templates that represented residues in terms of the positions of their C_α and C_β atoms, and one bank of templates that represented each residue using the atoms directly involved in liganding metal (these were required to lie within 3 Å of the metal). Neither type of template represented the metal ion itself. The templates permitted matching of chemically similar residues (Ser matched Thr, Asp matched Glu and Asn matched Gln). They also allowed the two oxygens in the carboxyl group of an Asp or Glu residue to match either way around, because they are chemically equivalent. Those template residues that bound metal via the protein backbone (rather than the residue sidechain) were permitted to

match any residue type.

Templates including water were capable of matching to any water molecule, including waters described in PDB files as oxygens in a pseudo-residue that included the metal ion. In both the C_α/C_β templates and the metal-binding-atom templates, the water molecule was represented by its oxygen atom coordinates.

4.4.3 Using structural templates to look at divergent evolution

Within each metal site family, one template of each type was created from the literature entry. RMSDs between the literature metal binding site and the metal binding sites of all family members were calculated by running this template against all family members using the program Jess to check for a structural match. It was then checked that this match was with those residues annotated as being metal binding by the PSIBLAST runs.

4.4.4 Similarity within template families

The percentage sequence identity between proteins, and the correlation between sequence identity and RMSD, were calculated in the same manner as described in the previous chapter.

4.4.5 Non-redundant PDB subset

The non-redundant subset of the PDB used for template searching was assembled in the same manner as the non-redundant PDB subset described in the previous chapter. However, the non-redundant subset used in this chapter used a different version of the non-redundant chain set provided by the NCBI– the version released on the 4th of January 2005.

4.4.6 Template matching

The program Jess (Barker & Thornton, 2003) was used for template matching. The templates were configured such that Jess only returned matches if all inter-atom distances in the match differed from the equivalent distances in the template by no more than 6 Å.

The exception was for matching metal site family members, where a distance threshold of 8 Å was used. (This higher threshold would not have been feasible if used over the non-redundant subset of the PDB.)

Only the best (lowest RMSD) match obtained by a given template on a given structure was retained.

Since the templates did not represent the metal ion itself, a template match did not necessarily imply the presence of a metal ion. Matches with an RMSD better than 1.5 Å were checked for the presence of metal ions. If all residues were within 3 Å of the same metal ion, then that match was regarded as having a metal ion present. The distribution of RMSDs (shown in the Results section) suggested that there were likely to be few matches featuring metal ions with RMSDs above 1.5 Å. Note that it was possible for a matched metal site to feature *more* ligands than the template that matched it.

4.4.7 Loss and gain of metal binding

For each of the metal site families, all homologous structures from the same CATH protein domain superfamily were checked for the presence of metal. This checking process consisted of checking the PDB structure for the occurrence of the metal, and also checking the UniProt sequence database (UniProt Consortium, 2007) entry for the protein for mention of the metal in its "feature table" or "cofactor" fields. Note that this system will occasionally falsely identify proteins as lacking or having metal in a given binding site. It will falsely identify a protein as lacking the metal if its structure lacks bound metal (because it is the apo form of the protein, or because a non-cognate metal is present) and if the UniProt file is also misannotated. It will falsely identify a protein as having metal in a given binding site if the protein binds that metal, but at a different location to the one being analysed.

The resulting cases of metal loss/retention were grouped by their CATH S35 families (groups of proteins with at least 35% sequence identity to one another). Any CATH S35 family that apparently included some structures with metal and others without was manually checked, in order to discover whether this really was a case of close relatives varying in whether they bound metal.

Furthermore, all the CATH S35 families that had been identified at the previous stage of analysis as having at least one member lacking metal were examined, and each family was checked using one arbitrarily selected representative. This representative was aligned to the literature entry using the 3D structural alignment server SSM (Krissinel & Henrick, 2004). The region on the relative that was equivalent to the metal binding region on the protein was manually examined to ensure that the metal-binding residues had been deleted or substituted, which confirmed that loss of metal binding had occurred.

In order to investigate the structural causes of metal loss, within each of the metal site families, the representative (described in the previous paragraph) from each CATH S35 family that did not bind metal was compared with the template structure using SSM in order to find the closest relative that did not bind metal. The closest relative was taken as being the one with the highest SSM Q-value (a measure of the quality of the C_α alignment between the two structures).

The region on this relative that was equivalent to the metal binding region on the protein was manually examined, and assigned one of the following reasons for metal loss:

- Substitution of residues responsible for metal binding
- Loss of a loop responsible for metal binding
- Some residues responsible for metal binding are substituted, while others are deleted due to being part of a loop that has been lost
- Loss of one or more secondary structural elements responsible for metal binding.

Other structural reasons for loss of metal binding are conceivable, but these are the only reasons which occurred in the structures examined.

Chapter 5

Geometry of interactions between catalytic residues and substrates

5.1 Introduction

This chapter returns to the analysis of catalytic residue geometry. Whereas Chapter three was concerned with analysing changes in the overall conformation of groups of catalytic residues between homologous proteins, this chapter examines the geometry of individual catalytic residues relative to their substrates. The motivation for this analysis can best be explained by setting it in the context of the role that protein structures typically play in inferring catalytic mechanism.

Elucidating the catalytic mechanism of an enzyme is a complex process of deduction from various types of evidence, including site-directed mutagenesis, kinetic studies, chemical modification of amino acid residues, and a range of other techniques. The three-dimensional structure of the enzyme contributes key information, as well as providing a framework for the interpretation of other forms of evidence. Enzyme structures provide more mechanistic information when they include one or more bound substrates, products, or analogues, because such structures show where amino acid residues lie in relation to the substrate. However, there are many enzymes for which structures of this type are available, but whose mechanisms remain uncertain. It is therefore useful to be able to make as accurate an assessment of enzyme mechanism as possible, based on a structure

of enzyme with bound substrate.

As described in Chapter one, it is difficult to obtain structures of enzymes complexed with their substrates, because the enzyme will convert substrate to product. As detailed in that chapter, it is generally necessary to use a complex with product, or to sabotage catalysis in some manner: for example, by omitting a substrate (where there are several) or a cofactor, by employing a competitive inhibitor instead of substrate, by using an inactive mutant form of the enzyme, or by providing a cofactor in the wrong oxidation state for a reaction to proceed (Fersht, 1999; Price & Stevens, 1999).

The geometry of non-enzymatic chemical reactions has been explored using crystallography to assess the likely trajectories of the atoms involved (Bürgi & Dunitz, 1983). This work has occasionally been invoked to explain the geometry of catalytic residues observed in enzymes (Huber & Bode, 1978). However, there have been no general empirical studies of the detailed geometry of catalytic residues relative to the groups with which they chemically interact. Theoretical treatments of catalysis have necessarily tended to focus on energetics and generalised stabilisation of the transition state, rather than the geometry of individual residues (Bugg, 2001; Kraut *et al.*, 2003; Garcia-Viloca *et al.*, 2004).

Some catalytic residues function by accepting or donating protons. This raises the question of whether these residues have a geometry with regard to their substrate that resembles the geometry of a hydrogen bond, or whether they are geometrically distinctive. Furthermore, residues which stabilise charge on an intermediate or transition state, or which modify the pK_a of a second catalytic residue, often do so by accepting or donating a hydrogen bond. Whilst these residues are quite distinct in function from those that transfer protons, the same question arises: to what extent does their geometry resemble that of non-catalytic hydrogen bonds?

The geometry of hydrogen bonding has been extensively studied, both in high-resolution structures of small molecules (Taylor & Kennard, 1984) and in proteins (Momany *et al.*, 1975; Baker & Hubbard, 1984; Gorbitz, 1989; McDonald & Thornton, 1994; Kortemme *et al.*, 2003). Both types of data source provide a similar picture of hydrogen bonding geometry. Figure 5.1 illustrates this geometry, and how it relates to hydrogen bonds between enzyme and substrate. Hydrogen-bonding geometry can be described in terms of

three parameters. The first parameter is the H···A distance, which tends to around 2 Å. The second parameter is the D-$\hat{\text{H}}$···A angle, which tends towards 180° (hydrogen bonds tend to have the donor, hydrogen, and acceptor in a straight line). The number of possible hydrogen bond conformations becomes smaller as the D-$\hat{\text{H}}$···A angle approaches 180°, so in practice the average value for this angle is around 160° despite the preference for linearity (Taylor & Kennard, 1984). The third parameter is the H···$\hat{\text{A}}$-AA angle, which varies considerably but averages around 120°. Hydrogen bonds have a weak tendency to be oriented so that the hydrogen is aligned with the lone pair on the acceptor, and this angle distribution reflects that.

Is the same hydrogen bonding geometry found in catalytic residues whose functions involve hydrogens? It has been observed that there is non-optimal hydrogen bonding geometry in serine proteases between the catalytic serine and the histidine which abstracts a proton from it. Dodson & Wlodawer (1998) propose that this poor geometry favours proton transfer by reducing the energy of the hydrogen bond.

It has been proposed that low barrier hydrogen bonds (LBHBs) play an important role in stabilising intermediates and transition states in enzymes (Gerlt & Gassman, 1993; Cleland *et al.*, 1998). LBHBs are a type of short, strong hydrogen bond where the energy barrier to movement of the hydrogen between the heavy atoms in the bond is reduced, and the bond becomes largely covalent in character. Proponents of the theory that LBHBs are important in enzyme mechanisms have suggested that a single such bond could contribute 10–20 kcal/mol to the stabilisation of the transition state (Cleland & Kreevoy, 1994).

The existence of LBHBs in the transition states of certain enzymes has been advanced on the basis of evidence including the NMR shift that is observed for the proton in some hydrogen bonds in enzymes, and the short distances observed between the heavy atoms involved in hydrogen bonds in some high-resolution structures of transition state analogue complexes (Cleland *et al.*, 1998).

However, whether LBHBs are important in enzymatic catalysis remains controversial; the theory has been disputed on theoretical (Warshel & Papazyan, 1996) and experimental (Guthrie, 1996) grounds. It has also been proposed that stong hydrogen bonds in enzymes could be ionic rather than covalent in character; it is difficult to distinguish covalent LBHBs from ionic hydrogen bonds using experimental evidence (Schutz & Warshel, 2004;

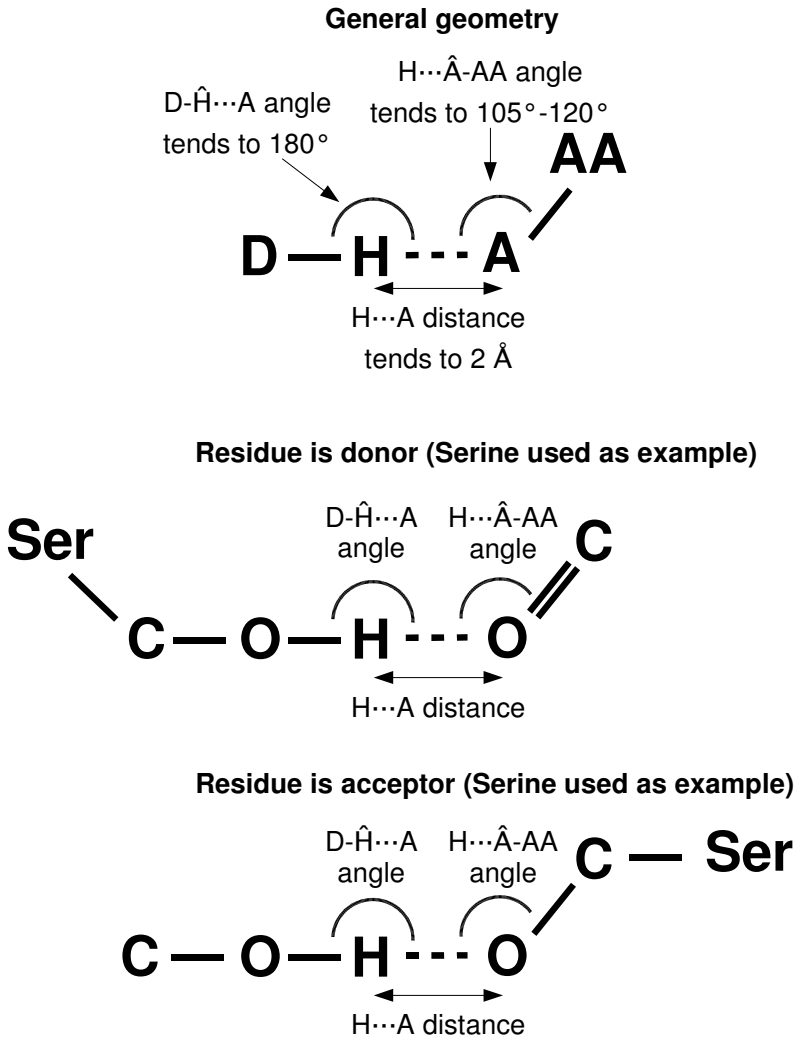


Figure 5.1: Geometry of hydrogen bonding.

Fuhrmann *et al.*, 2006).

Whilst the present chapter describes hydrogen bond geometry in relation to catalysis, it is not possible to address the question of the importance of LBHBs here because the complexes employed in this chapter are not transition state analogues.

This chapter describes an analysis of patterns in the geometry of catalytic residues with regard to bound substrates/ products/ analogues. This analysis makes use of the MACiE database (Holliday *et al.*, 2007a). Like the CSA, MACiE is a database which manually annotates enzymes based on the scientific literature. Whereas the CSA is concerned with catalytic residues, MACiE is concerned with a more detailed chemical description

of catalytic mechanisms for a set of enzymes whose mechanism is well-established. Each reaction described in MACiE is broken down into a series of reaction steps. For each step, the catalytic function performed by each residue is labelled by one or more of a set of functional descriptors: nucleophile, proton donor, charge stabiliser, and so on.

This chapter also includes an equivalent analysis of the geometry of catalytic residues which act upon other catalytic residues. Because this did not require protein structures to have substrate present, a larger dataset could be used. This analysis is based on the CSA, which has a broader coverage than MACiE.

Despite access to these two databases, considerable manual effort was required to assemble a dataset of enzymes where the mechanism is fully described, where structures with cognate substrates (or products or analogues) are available, and where the interactions between the residues and the substrates are described on an atom-by-atom level. Such detailed information is only available for a small number of enzymes.

This chapter is primarily an empirical study of where catalytic groups occur in protein structures, intended to aid in the assignment of likely catalytic residues from structure when attempting to elucidate an enzyme's mechanism. It can only hint at the geometry which actually occurs during the reaction. This limitation is partly due to the fact that the dataset of structures includes many complexes featuring products and/or substrate analogues as described above. The limitation is also due to the fact that enzymes are flexible systems in constant motion and that, furthermore, enzymes move during the process of catalysis (Kern *et al.*, 2005). The crystal structures analysed are only snapshots of the structure of the enzyme, and it is beyond the scope of this initial study to assess the dynamic aspect.

5.2 Results

5.2.1 Residue-substrate dataset

The analysis of the geometry of residues relative to substrate was based on a dataset of 42 enzyme structures, selected from the PDB in the following manner:

The analysis required enzyme structures where the mechanism was well-understood,

and which had either all substrates or all products present. The MACiE database provides a description of the mechanism of 202 enzymes. These enzymes cover a wide range of functions: they cover 158 different third-level EC numbers (EC sub-subclasses), which constitutes 87% of the 181 sub-subclasses for which structures are available. Each of these 202 enzyme mechanism descriptions is linked to an enzyme structure in the PDB. Only a small number of these particular structures had substrate or product present. Therefore, for each of these structures, structures of related proteins were identified (homologues with the same EC number to the fourth level); there were 3295 such structures. For each one of the 202 original MACiE entries, an attempt was made to identify a single relative which had either all substrates or all products (or close analogues to the substrates or products) present. Because structures of this type are relatively rare, this resulted in a final dataset of 42 protein structures, which are described in Table 5.1. Fourteen of these structures are featured in MACiE; the remaining 28 structures are relatives of MACiE entries, found in the PDB. Further details of the constraints on structures used are provided in the methods section.

Table 5.1 describes the reason why each structure analysed was available as a complex. Structures were only used if there was no reason to suppose that the binding of substrate analogues (or other means of disrupting the reaction) would alter the geometry of the catalytic residues relative to the substrate. The most frequent reason why structures were available was the use of a substrate analogue, accounting for 14 cases in the dataset. There are six product complexes, and seven complexes where one substrate, product or transferred group is missing. In six examples a redox agent is in the wrong oxidation state for the reaction to proceed, and in three cases the protein has been mutated in a manner that prevents the reaction occurring. The remaining six structures cannot be fitted into any of the above categories.

For each of these enzymes, the MACiE database describes the reaction in terms of a series of reaction steps. MACiE describes a residue's actions in a given reaction step with one or more functional descriptors, such as *nucleophile* or *proton donor*. These descriptive labels are termed *residue functions* in this chapter. Note that all the residue functions are descriptions of the action performed *by the protein*: thus, the term *proton donor* means that the protein donates a proton to the substrate. An individual function performed

Enzyme name	PDB entry	Parent PDB entry from MACiE	EC number	Resolution (Å)	Reason complex available
Dethiobiotin synthetase	1a82	1dae	6.3.3.3	1.80	One substrate missing (CO_2)
Purine nucleoside phosphorylase	1a9t	1ula	2.4.2.1	2.00	Product complex
Estrogen sulphotransferase	1aqu	1hy3	2.8.2.4	1.60	Transferred group absent from substrate
D-amino acid oxidase	1c0p	1c0l	1.4.3.3	1.20	One substrate missing (O_2)
Creatine amidinohydrolase	1chm	1chm	3.5.3.3	1.90	Substrate analogue
Glucose dehydrogenase	1cq1	1c9u	1.1.5.2	1.90	Redox agent in wrong oxidation state
Orotidine 5'-phosphate decarboxylase	1dbt	1dbt	4.1.1.23	2.40	One product missing (CO_2)
UDP-glucose dehydrogenase	1dlj	1dli	1.1.1.22	1.80	Mutant
DNase I	1dnk	1dnk	3.1.21.1	2.30	DNase bound to DNA with cleavage-resistant sequence
Phosphotriesterase	1dpm	1pta	3.1.8.1	2.10	Substrate analogue
HMG-CoA reductase	1dqa	1dqa	1.1.1.34	2.00	Redox agent in wrong oxidation state, substrate analogue
Quinone reductase	1dxo	1d4a	1.6.5.2	2.50	Redox agent in wrong oxidation state
Uracil-DNA glycosylase	1emh	1eug	3.2.2.3	1.80	Substrate analogue
Alkaline phosphatase	1ew8	1alk	3.1.3.1	2.20	Substrate analogue
Adenylosuccinate synthetase	1gim	1gim	6.3.4.4	2.50	Substrate analogue; also, GDP present in place of GTP
Glutathione synthetase	1gsa	1gsa	6.3.2.3	2.00	Product analogue
Pyruvate formate-lyase	1h16	2pfl	2.3.1.54	1.53	Radical mechanism requires activation of the enzyme by another enzyme
Methyl-coenzyme M reductase	1hbn	1mro	2.8.4.1	1.16	Substrate analogue (methyl group missing). Also, nickel ion is in inactive Ni(III) form.
DNA topoisomerase III	1i7d	1i7d	5.99.1.2	2.05	Mutant
Methylglyoxal synthase	1ik4	1b93	4.2.3.3	2.00	Substrate analogue
Chalcone isomerase	1jep	1eyq	5.5.1.6	2.10	Product complex
Lactate dehydrogenase	1kbi	1fcb	1.1.2.3	2.30	Product complex
Thymidylate synthase	1kzi	1lcb	2.1.1.45	1.75	Substrate analogue
S-adenosylmethionine synthetase	1p7l	1fug	2.5.1.6	2.50	Substrate analogue
Malate synthase G	1p7t	1d8c	2.3.3.9	1.95	Substrate analogue
Phosphomannomutase	1pcm	1p5d	5.4.2.2	1.90	Zn cofactor substituted by Mg
Malic enzyme	1pj2	1do8	1.1.1.38	2.30	Redox agent in wrong oxidation state
2,4-dienoyl-CoA reductase	1ps9	1ps9	1.3.1.34	2.20	Redox agent in wrong oxidation state
Group II chaperonin	1q3q	1a6d	3.6.4.9	2.30	Substrate analogue
Fumarylacetoacetate hydrolase	1qco	1qco	3.7.1.2	1.90	Product complex
Fumarate reductase	1qlb	1qjd	1.3.99.1	2.33	Redox agent in wrong oxidation state
Quinolinic acid phosphoribosyltransferase	1qpr	1qpr	2.4.2.19	2.45	Substrate analogue
Cyclophilin A	1rmh	1m9c	5.2.1.8	2.40	Cis-trans peptidyl prolyl isomerase that can only bind to the cis-conformation
Sulphite oxidase	1sox	1sox	1.8.3.1	1.90	Product complex, missing one product (H_2O_2)
Phosphoglycerate kinase	1vpe	13pk	2.7.2.3	2.00	Product analogue
Cytokinin dehydrogenase	1w1s	1w1o	1.5.99.12	2.00	Substrate analogue
Nucleotide diphosphate kinase	1wkl	1kdn	2.7.4.6	2.20	Unclear— publication not available
UDP-galactose 4-epimerase	1xel	1xel	5.1.3.2	1.80	Redox agent in wrong oxidation state
Propionyl-CoA carboxylase	1xny	1xny	6.4.1.3	2.20	Transferred group absent from substrate
Isoaspartyl dipeptidase	1ybq	1ybq	3.4.19.-	2.00	Mutant
L-arginine-glycine amidinotransferase	3jdw	9jdw	2.1.4.1	2.40	Transferred group absent from substrate
Methylmalonyl-CoA mutase	4req	1req	5.4.99.2	2.20	1:1 mix of substrate and product in crystal

Table 5.1: Protein structures used in the residue-substrate analysis.

by a given residue upon a given atom in a given reaction step is described as a *residue operation* in this chapter. Residue operations were only included in this analysis if they act upon a specific substrate, product, intermediate or transition state atom. Residues which do not act at a specific substrate atom (generally because they act sterically) were not included in this analysis.

Note that it is possible for a residue to act on the same substrate atom in different ways at different steps of a reaction; a residue might be a proton donor in the first step, then act to stabilise a charge in the second step. Where this was the case, both residue operations were included in this analysis. This means that the data for different functions are not independent of one another. This was necessary given the relatively small size of the dataset. Note also that because residue operations from multiple reaction steps were used, a given residue operation may act on a substrate, intermediate, transition state, or product (or an analogue), although the geometry information was always derived from a structure with substrate or product (or an analogue). The term "substrate" is used below for convenience to refer to all of the above species.

A total of 83 residue operations were analysed; these are broken down by residue function in Table 5.2. The low number of residue operations per protein structure (slightly less than two) is a consequence of the requirement that residue operations can only be analysed if the geometry of the residue operation as it occurs in the available complex is likely to resemble that which occurs at the appropriate reaction stage. This requirement is often broken by the use of substrate analogues, or by the chemical changes that occur through the course of a reaction.

The majority of the residue operations are proton donors/acceptors, or stabilisers of charge on the substrate via a hydrogen bond. Table 5.2 also includes data on distances between the residue atoms directly involved in catalysis and the substrate atoms upon which they act.

The distribution of residue types for each function is shown in Table 5.3. Proton exchanging residues tend to be those with sidechain pK_a values near to 7, facilitating loss and gain of protons. For both proton donors and proton acceptors, His is the most common residue, and Tyr is either second or joint second. Proton donors also include three lysines and small numbers of other residues; proton acceptors also include Asp and Glu

Target	Residue function	Number	Distance (σ)
Substrate	Proton donor	20	3.21 (0.65)
Substrate	Proton acceptor	16	2.89 (0.42)
Substrate	Charge stabiliser via H-bond donation	33	2.92 (0.29)
Substrate	Charge stabiliser via H-bond acceptance	4	2.79 (0.30)
Substrate	Nucleophile	4	2.52 (0.67)
Substrate	Electrophile	1	1.67 (-)
Substrate	Radical donor	1	5.16 (-)
Substrate	Radical acceptor	1	2.56 (-)
Substrate	Radical stabiliser	3	4.06 (1.38)
Residue	Proton donor	24	3.04 (0.54)
Residue	Charge stabiliser via H-bond donation	68	3.10 (0.88)

Table 5.2: Distances between residues and their targets.
The distance given is the distance between the residue atom and the substrate (or secondary residue) atom (excluding hydrogen atoms for both residue and substrate) that are involved in the residue function. Note that the average distance between donor and acceptor atoms for hydrogen bonds (using the dataset shown in the line plots in Figures 5.2, 5.3, 5.4, and 5.5) is 2.89 Å.

and a smaller number of lysines. There are a few residues which appear counter-intuitive: the two lysines that accept protons, and the two glutamates that donate protons. Three of these four residues display this behaviour because they both gain and lose protons during the course of the reaction. The remaining case (Glu 559 from PDB entry 1dqa) has an environment likely to raise its pK_a, and thus it is probable that it is protonated and can act as a proton donor (Istvan *et al.*, 2000).

The distribution of residue types is slightly different for charge stabilising residues, where charged residues predominate regardless of pK_a. Residues that stabilise positive charges by hydrogen bond acceptance tend to be Asp or Glu. Residues that stabilise negative charges by hydrogen bond donation tend to be His, Lys or Arg.

5.2.2 Residue-substrate geometry

The majority of catalytic residue functions in this dataset (73 out of 83) involve hydrogens. These include proton donation, proton abstraction, and charge stabilisation through hydrogen bonding. For these residue functions, this section of the analysis ex-

Target	Function	A	C	D	E	G	H	K	M	N	Q	R	S	T	W	Y	Σ
Substrate	Proton donor	0	1	0	2	0	8	3	0	0	0	1	0	0	0	5	20
Substrate	Proton acceptor	0	0	3	3	0	5	2	0	0	0	0	0	0	0	3	16
Substrate	Charge stabiliser via H-bond (D)	1	0	1	1	4	6	8	0	1	0	9	0	2	0	2	33
Substrate	Charge stabiliser via H-bond (A)	0	0	1	2	0	0	0	0	0	0	0	1	0	0	0	4
Substrate	Nucleophile	0	2	0	0	0	1	0	0	0	0	0	1	0	0	0	4
Substrate	Electrophile	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Substrate	Radical donor	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Substrate	Radical acceptor	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Substrate	Radical stabiliser	0	0	0	0	0	1	0	0	0	0	0	0	0	0	2	3
Residue	Proton donor	0	4	3	1	0	1	1	0	0	0	1	12	0	0	1	24
Residue	Proton acceptor	0	0	5	3	0	11	3	0	0	0	1	0	0	0	1	24
Residue	Charge stabiliser via H-bond (D)	0	1	0	1	1	33	11	0	4	1	6	3	2	1	4	68
Residue	Charge stabiliser via H-bond (A)	1	6	26	14	0	1	3	1	2	1	2	4	0	0	7	68

Table 5.3: Residue type distribution for each residue function.
Residue types are specified by their one-letter identifiers. Cases identified as "Charge stabiliser via H-bond (D)" stabilise charge by (D)onating a hydrogen bond, whilst cases identified as "Charge stabiliser via H-bond (A)" stabilise charge by (A)ccepting a hydrogen bond.

amined whether the geometry of residue operations conforms to the standard geometry of hydrogen-bonding, or whether there are notable differences that might be used to differentiate a catalytic residue from a non-catalytic residue based on structure alone. Note that all the structures in this dataset are based on X-ray crystallography, and therefore do not include hydrogen positions. All hydrogen locations are modelled, making the assumption that hydrogens have standard covalent bond angles and lengths and that they are positioned in such as way as to minimise the hydrogen-acceptor distance.

In order to be able to compare the geometry of these residue operations against typical hydrogen bonding geometry between proteins and ligands, a a dataset of non-catalytic protein-ligand hydrogen bonds was prepared. This dataset consisted of all non-catalytic hydrogen bonds between protein and non-water ligands in the dataset of 42 protein structures described above, giving a total of 1092 hydrogen bonds. Because this dataset of hydrogen bonds was based on the same set of protein structures as the residue operation dataset, it should reflect the degree of apparent variation in hydrogen bonding geometry that occurs at the level of resolution found in the residue operation dataset.

In line with previous studies, this analysis uses geometric thresholds to distinguish hydrogen bonds from other interaction types (McDonald & Thornton, 1994). In this study, interactions were only treated as hydrogen bonds if the D$\cdots$A distance was less than 3.9 Å, the H$\cdots$A distance less than 2.5 Å, and the H$\cdots$Â-AA and D-Ĥ$\cdots$A angles

greater than 90°. These are the same limits used by McDonald (McDonald & Thornton, 1994; McDonald, 1995), and are similar to those used in other studies of hydrogen bonding (Baker & Hubbard, 1984; Gorbitz, 1989).

The geometry of non-catalytic hydrogen bonds between protein and ligand determined in this way is shown in the line histograms in Figures 5.2, 5.3, 5.4, and 5.5. These line histograms are the same in all four figures, although their vertical scale differs because the frequencies were scaled separately for each graph such that the sum of all frequencies matches the sum of all frequencies for the equivalent catalytic residue parameter. D···A distance peaks at 2.8 Å, with the remaining distances closely distributed within 0.5 Å either side. H···A distance peaks at 2.0-2.1 Å, with almost all of the remaining distances falling between 1.7 Å and 2.3 Å. D-Ĥ···A angle peaks between 160° and 170°, with fewer angles above 170°. The reason that the peak is not at 180° is that, unlike previous studies, this analysis has not corrected for the volume of space occupied by each angle band; when the volume correction employed by Kortemme *et al.* (2003) is used, the peak is at 180°. The H···Â-AA angle peaks between 110° and 120°, with observations spread over about 30° either side. These distributions are comparable with those obtained for sidechain-sidechain hydrogen bonds by Kortemme *et al.* (2003) when counts are corrected for the volume of each angle or distance band using the method described in that study.

The geometry of the catalytic residue operations whose functions involve hydrogens is shown by the bar histograms in Figures 5.2 (proton accepting residues), 5.3 (proton donating residues), 5.4 (residues stabilising charge by accepting hydrogen bonds), and 5.5 (residues stabilising charge by donating hydrogen bonds). These data show that the geometry of the hydrogens on the catalytic residues largely resembles the geometry of non-catalytic hydrogen bonds. It is difficult to confirm this by statistical means, for two reasons. Firstly, the dataset used here is small in size, because there were few enzymes available with adequate information. Secondly, any statistical comparison between the non-catalytic hydrogen bonds and the catalytic residues is unequal, because the non-catalytic hydrogen bonds are (by definition) within the parameter thresholds described above, whereas the catalytic residue geometry is not constrained in this way.

Despite these limitations, a statistical comparison was carried out of the catalytic residue geometry with the geometry of non-catalytic residues using a Mann-Whitney

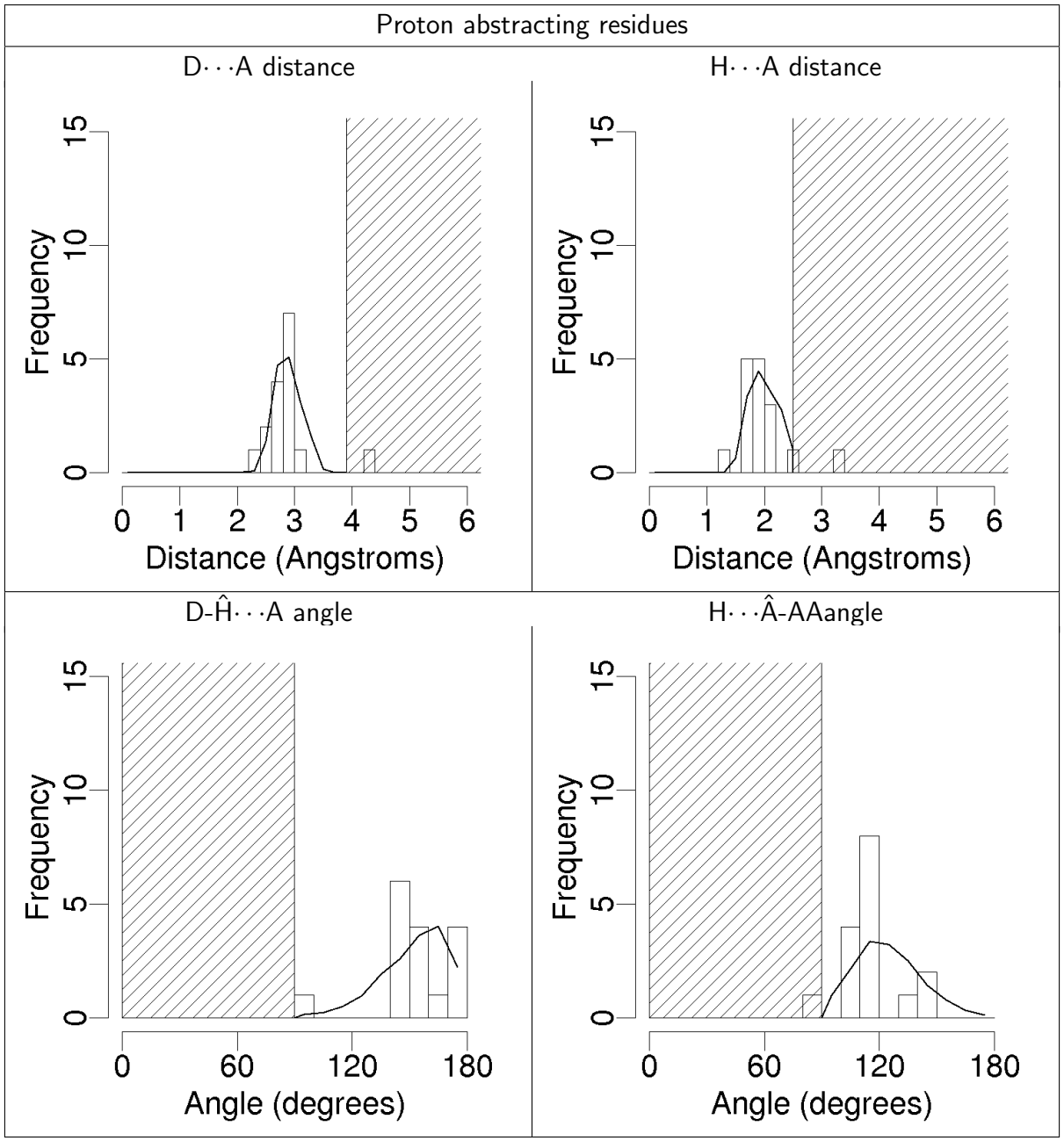


Figure 5.2: Geometry of proton abstracting residues (acting on substrate) compared to hydrogen bonding.
Bars show distribution of geometric parameters for catalytic residues. Lines show distribution of the same geometric parameters for non-catalytic hydrogen bonds. The frequencies for the distribution of non-catalytic hydrogen bond parameters have been adjusted separately for each graph such that the sum of all frequencies matches the sum of all frequencies for the equivalent catalytic residue parameter. Angles and distances are defined in Figure 5.1. Values for parameters which are outside the criteria for recognition as a hydrogen bond are marked by shaded areas.

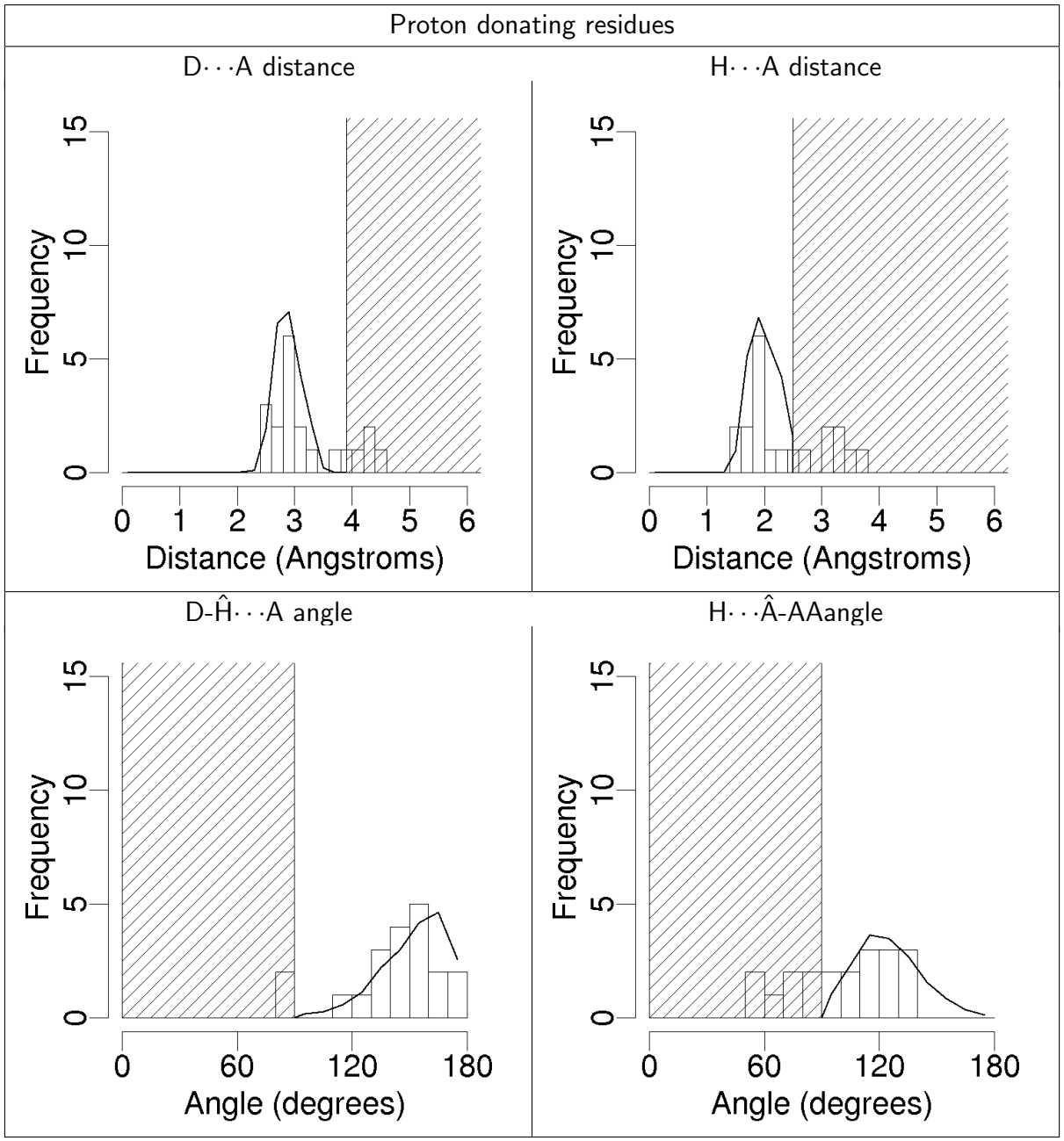


Figure 5.3: Geometry of proton donating residues (acting on substrate) compared to hydrogen bonding.
Bars show distribution of geometric parameters for catalytic residues. Lines show distribution of the same geometric parameters for non-catalytic hydrogen bonds. The frequencies for the distribution of non-catalytic hydrogen bond parameters have been adjusted separately for each graph such that the sum of all frequencies matches the sum of all frequencies for the equivalent catalytic residue parameter. Angles and distances are defined in Figure 5.1. Values for parameters which are outside the criteria for recognition as a hydrogen bond are marked by shaded areas.

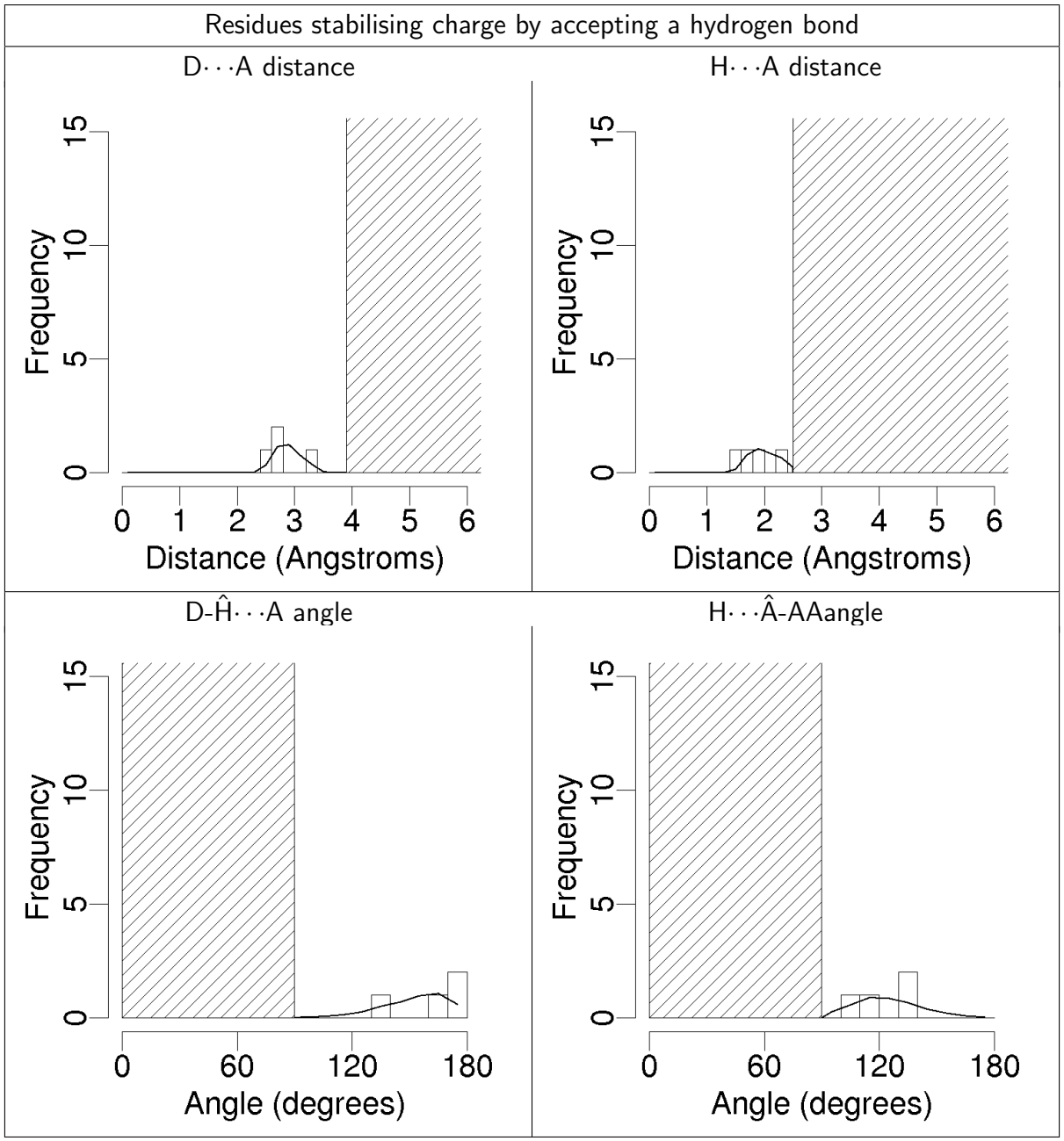


Figure 5.4: Geometry of residues stabilising charge by accepting a hydrogen bond (acting on substrate) compared to hydrogen bonding.
Bars show distribution of geometric parameters for catalytic residues. Lines show distribution of the same geometric parameters for non-catalytic hydrogen bonds. The frequencies for the distribution of non-catalytic hydrogen bond parameters have been adjusted separately for each graph such that the sum of all frequencies matches the sum of all frequencies for the equivalent catalytic residue parameter. Angles and distances are defined in Figure 5.1. Values for parameters which are outside the criteria for recognition as a hydrogen bond are marked by shaded areas.

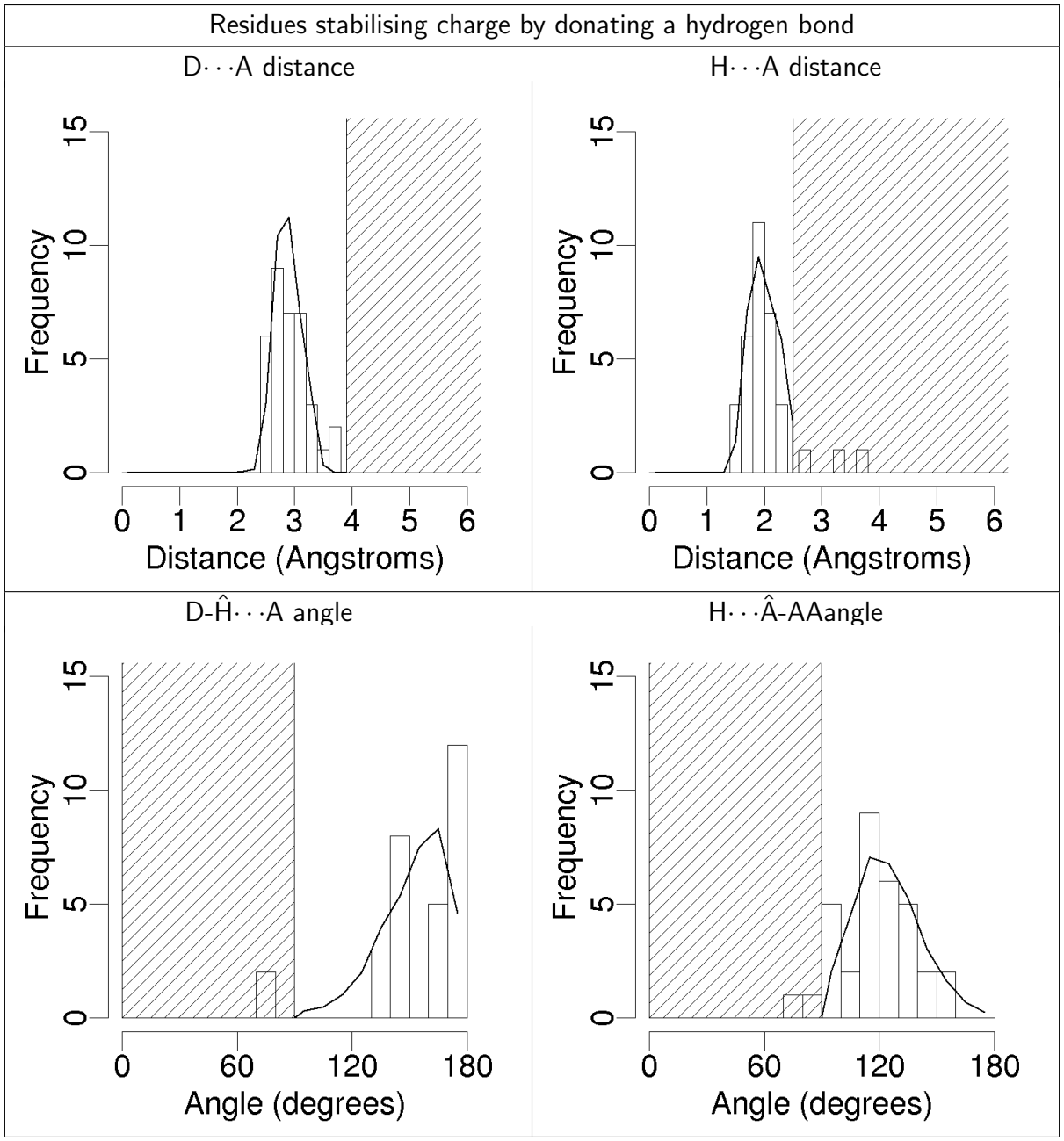


Figure 5.5: Geometry of residues stabilising charge by donating a hydrogen bond (acting on substrate) compared to hydrogen bonding.
Bars show distribution of geometric parameters for catalytic residues. Lines show distribution of the same geometric parameters for non-catalytic hydrogen bonds. The frequencies for the distribution of non-catalytic hydrogen bond parameters have been adjusted separately for each graph such that the sum of all frequencies matches the sum of all frequencies for the equivalent catalytic residue parameter. Angles and distances are defined in Figure 5.1. Values for parameters which are outside the criteria for recognition as a hydrogen bond are marked by shaded areas.

nonparametric test (Mann & Whitney, 1947). In most cases, this comparison found no significant difference (statistical significance threshold $\alpha = 0.05$; applying Bonferroni correction (Bonferroni, 1936) for the 24 Mann-Whitney tests in this chapter gives $\alpha = 2.08 \times 10^{-3}$) between the distribution of the distance and angle parameters for non-catalytic hydrogen bonds and the hydrogen geometry for the catalytic residues. Note that the fact that the non-catalytic data is constrained by the parameter thresholds (whereas the catalytic data is not) biases any test *in favour* of a significant result, suggesting that the similarity is genuine. The only case where there is a significant difference at the $\alpha = 2.08 \times 10^{-3}$ level is the H$\cdots$Â-AA angle for proton donors. The proton donors in this dataset tend to have lower H$\cdots$Â-AA angles than occur for non-catalytic hydrogen bonds.

The H$\cdots$A distance for catalytic residue operations was compared with the angle parameters in order to determine whether there was any relationship between them. These comparisons are shown in Figures 5.6 (for proton donors and acceptors) and 5.7 (for residues stabilising charge through hydrogen bonding). There is little correlation between the H$\cdots$A distance and either the D-Ĥ$\cdots$A or H$\cdots$Â-AA angles, for any function. Furthermore, there is no clear relationship between these parameters and the resolution of the protein structure.

5.2.3 Residue-substrate operations with unusual geometry

Given that the overall distribution of geometries for catalytic residue operations involving hydrogens is similar to that observed for non-catalytic hydrogen bonds, it is interesting to look more closely at those individual residue operations which have a geometry distinct from that observed for non-catalytic hydrogen bonds. This section of the analysis examines those outlying residue operations which have parameters outside those permitted for the hydrogen bonds in the non-catalytic dataset. These cases are marked on the histograms in Figures 5.2, 5.3, 5.4, and 5.5, and the scatter plots in Figures 5.6 and 5.7 by shaded areas.

These residue operations which fail to meet the geometric definition of hydrogen bonds specified above include some whose function is to stabilise charge via hydrogen bonding. There is not necessarily a contradiction here: it is possible that the conformation in the

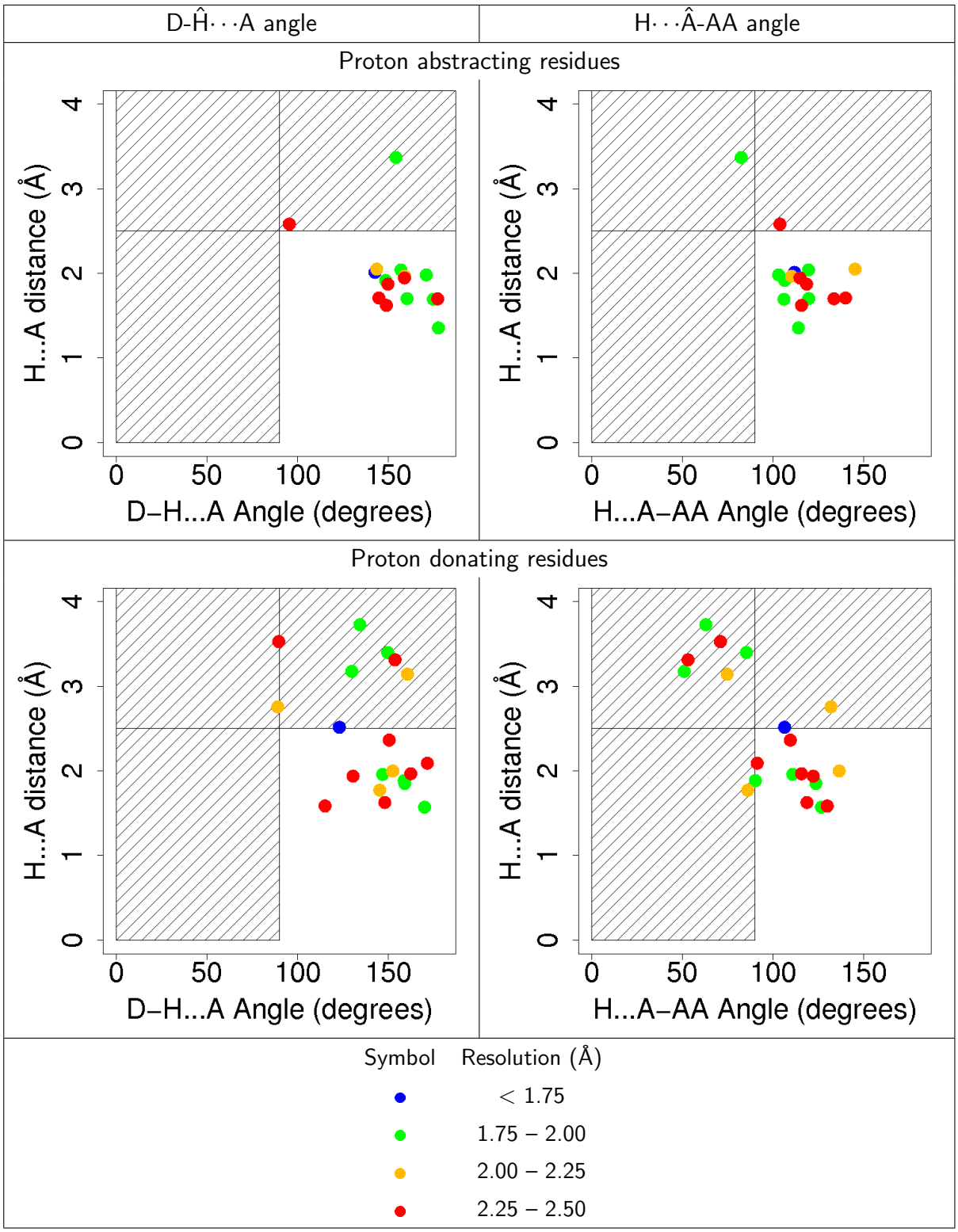


Figure 5.6: Relationship of angles to distances for geometry (relative to substrate) of proton transferring residues.

Angles and distances are defined in Figure 5.1. Values for parameters which are outside the criteria for recognition as a hydrogen bond are marked by shaded areas. Points are coloured according to the resolution of the structure they are derived from, as described in the key.

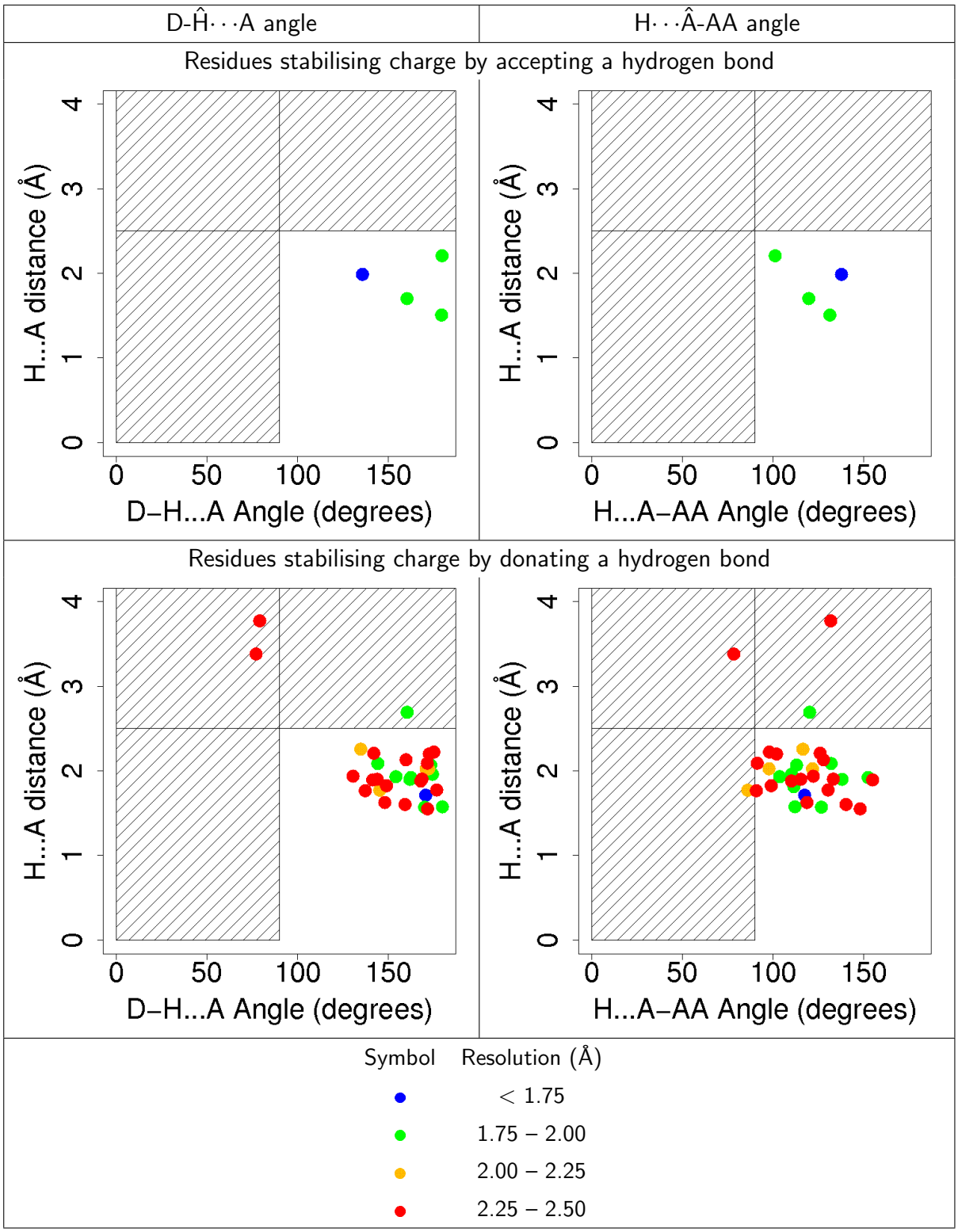


Figure 5.7: Relationship of angles to distances for geometry (relative to substrate) of charge stabilising residues.

Angles and distances are defined in Figure 5.1. Values for parameters which are outside the criteria for recognition as a hydrogen bond are marked by shaded areas. Points are coloured according to the resolution of the structure they are derived from, as described in the key.

transition state differs from that in the complex analysed here, and that a hydrogen bond is formed in the transition state. Alternatively, it is possible that some of these residue operations do not in fact stabilise charge via hydrogen bonding, but rather via purely electrostatic interactions.

There are a total of 14 residue operations where the geometry is unusual on one or more parameters. For any given case, it is difficult to assign a definite reason why the residue geometry should be unusual, but it is possible to speculate on the most likely cause.

There are four cases (from PDB entries 1pj2, 1ps9, 1qco and 1qlb) where a proton donor donates a proton to a carbon which is engaged in a double bond prior to proton donation. In three of these cases, the distance between the proton donating residue is positioned closer to the double bond than to the carbon, in the sense that the distance between the donor atom and the closest point along the line of the double bond is smaller than the distance between the donor atom and the acceptor atom (Figure 5.8). In the one of the above four cases where the proton donating residue was not closer to the double bond than to the acceptor atom (that from PDB entry 1pj2) there was some uncertainty in the literature as to whether the residue in question was truly catalytic.

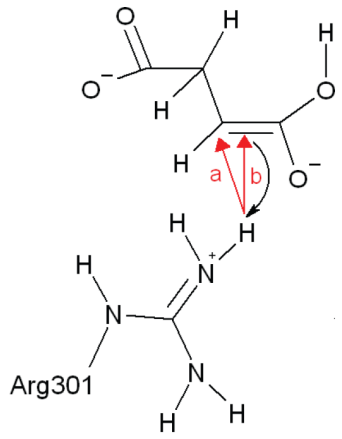


Figure 5.8: Geometry of residues acting on double bonds.
Distance (a) is the distance from the hydrogen to the carbon to which the hydrogen is donated. Distance (b) is the distance from the hydrogen to the closest point on the hydrogen bond (where this is treated as a straight line between the bonded atoms). In some instances (described in the text) distance (a) is larger than distance (b). The example given is from PDB entry 1qlb (Lancaster *et al.*, 1999).

There is one case (from PDB entry 1emh) where the cause appears to be that a competitive inhibitor was present instead of the true substrate. Although this inhibitor differs only slightly from the true substrate, this slight difference appears to be sufficient to distort the position of the catalytic residue.

There are two residue operations (both from PDB entry 1qlb) where close inspection of the literature suggests that the protein departs slightly from its catalytic conformation (Lancaster *et al.*, 1999; Reid *et al.*, 2000). This enzyme (fumarate reductase) undergoes a conformational change from an open to a closed form upon substrate binding. The structure 1qlb was determined by taking a crystal of the enzyme in the open form and diffusing in substrate, which appears to have resulted in a structure which is in a slightly open conformation, despite the presence of substrate. One of these two residue operations involves proton donation to a carbon involved in a double bond, and was mentioned in the previous paragraph; one or both factors may contribute to its outlying geometry.

There is one charge stabilisation residue operation (from PDB entry 1vpe) where the residue is a lysine, and it is possible that the charge stabilisation is primarily electrostatic rather than hydrogen-bond mediated.

There is one case where the reason the residue geometry was unusual appears to be that there was structural change during the reaction. This is *Escherichia coli* topoisomerase III (PDB identifier 1i7d): the residue Glu7 has an unusual conformation. Glu7 protonates a leaving group (Figure 5.9), and it is possible that when the protonation actually occurs, the departure of the group brings it to the correct orientation. However, a nearby residue is mutated (Y328F). Hence an alternative explanation for the unusual conformation for Glu7 with regard to the substrate would be that the mutated residue is altering the normal conformation of the active site.

Residue operations with unusual angles almost always have unusually large distances. There are only two residue operations which have an unusual angle without an unusual distance: these are both associated with the same residue, His115 from *Thermus thermophilus* nucleoside diphosphate kinase (PDB entry 1wkl). This residue acts as a proton donor and charge stabiliser at one atom on the substrate, but it also acts as a nucleophile on a second substrate atom (Hutter & Helms, 2002). This is illustrated in Figure 5.10. The fact that this residue has two target atoms might account for its unusual geometry.

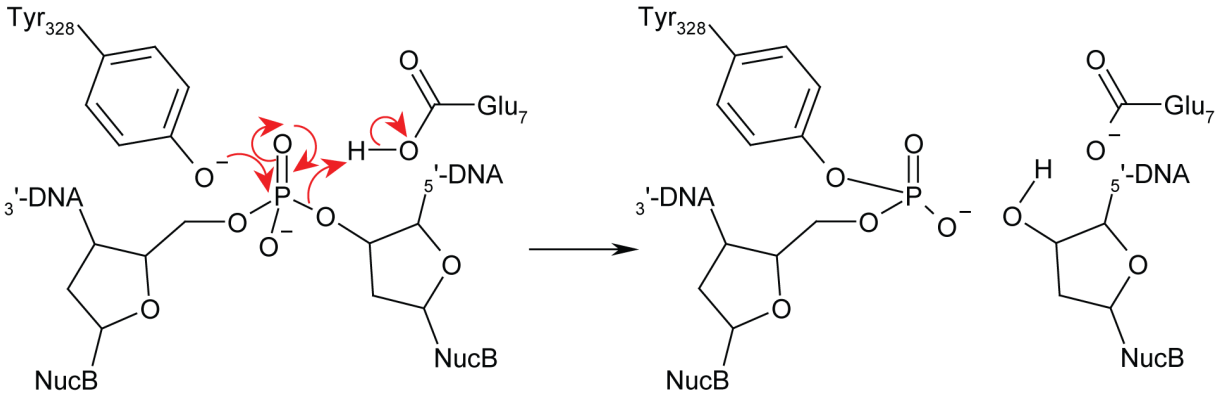


Figure 5.9: Role of Glu7 in *Escherichia coli* topoisomerase III. Glu7 protonates a leaving group (Changela *et al.*, 2001).

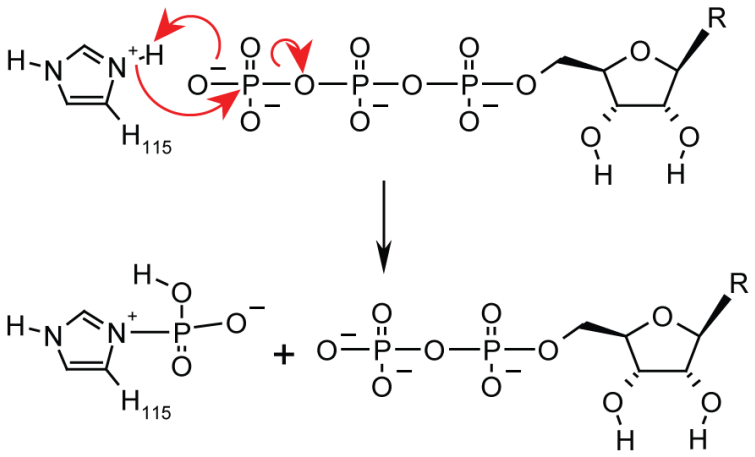


Figure 5.10: Role of His115 in *Thermus thermophilus* nucleoside diphosphate kinase. His115 acts as a proton donor and charge stabiliser at a substrate oxygen atom, but also acts as a nucleophile on a substrate phosphorus atom (Hutter & Helms, 2002).

This leaves a further four cases (less than a third of the outliers) where there is no clear reason why the geometry should be unusual.

5.2.4 Residue-residue dataset

An analysis was also carried out of the geometry of catalytic residues which act on other catalytic residues. This used a dataset of 60 enzyme structures, selected from the high annotation entries in the CSA; as described in Chapter two, these entries record which residues act on other residues, and the nature of the catalytic function they perform. The CSA was used as a data source for this section of the analysis rather than MACiE because the CSA has a broader coverage, and because the more detailed mechanistic information in MACiE was useful in the previous section primarily for determining which residue operations could be analysed based on a given complex of an enzyme with substrate/ product/ analogue, and because it specified which substrate atom(s) the residue acted upon. The structures used in this section of the analysis are not required to include substrate, product or analogue (since this will generally only have a small impact on residue-residue geometry), and the atom which a residue acts upon in another residue is usually easily identified, so the information in a high annotation CSA entry is sufficient for this part of the analysis. Further details of the constraints on the structures used in this residue-residue analysis are provided in the methods section. The structures employed are described in Table 5.4.

The dataset included a total of 92 residue operations; these are broken down by residue function in Table 5.2. All residue functions involve proton sharing or transfer. Since each interaction involves a pair of residues, it is arbitrary whether a given operation is defined as a proton donation, or a proton abstraction with the residue identities reversed. For the purposes of analysing geometry, this analysis treated all operations as cases of proton *donation* or charge stabilisation by hydrogen bond *donation*. However, for the purposes of analysing the distribution of residue types for each function (described in the next paragraph) each residue operation is counted twice: one of the two interacting residues is counted among the proton (or hydrogen bond) donors, and the other residue is counted among the proton (or hydrogen bond) acceptors.

Enzyme name	PDB entry	EC	Resolution (Å)
Bromoperoxidase A1	1a8q	1.11.1.10	1.75
Cutinase	1agy	3.1.1.74	1.15
Exonuclease III	1ako	3.1.11.2	1.70
Pyroglutamyl peptidase-1	1aug	3.4.19.3	2.00
Aminopeptidase	1b65	3.4.11.19	1.82
N-(1-D-carboxylethyl)-L-norvaline dehydrogenase	1bg6	1.5.1.28	1.80
Protein tyrosine phosphatase	1bzc	3.1.3.48	2.35
Carboxylesterase	1ci8	3.1.1.1	2.00
Cytosolic phospholipase A2	1cjy	3.1.1.4	2.50
Ubiquitin C-terminal hydrolase	1cmx	3.4.19.12	2.25
Cystathionine gamma-synthase	1cs1	2.5.1.48	1.50
Haloalkane dehalogenase	1cv2	3.8.1.5	1.58
DNA mismatch endonuclease	1cw0	3.1.-.-	2.30
Intron-encoded homing endonuclease I-PpoI	1cz0	3.1.-.-	2.10
Acid phosphatase	1d2t	3.1.3.2	1.90
Proteinase B	1ds2	3.4.21.81	1.70
Diisopropylfluorophosphatase	1e1a	3.1.8.2	1.80
GDP-fucose synthetase	1e7q	1.1.1.271	1.60
D-aminopeptidase	1ei5	3.4.11.9	1.90
Sialidase	1euu	3.2.1.18	2.50
Phosphoribosylglycinamide formyltransferase 2	1ez1	2.1.2.-	1.75
Phosphotriesterase	1ez2	3.1.8.1	1.90
Hyaluronoglucosaminidase	1fcq	3.2.1.35	1.60
N-carbamoyl-D-amino-acid amidohydrolase	1fo6	3.5.1.77	1.95
Beta-1,4-galactanase	1fob	3.2.1.89	1.80
Aspartyl dipeptidase	1fy2	3.4.13.21	1.20
Estradiol 17 beta-dehydrogenase 4	1gz6	1.1.1.35	2.38
Alpha-chymotrypsin	1hja	3.4.21.1	2.30
Pancreatic lipase	1hpl	3.1.1.3	2.30
Riboflavin synthase	1i8d	2.5.1.9	2.00
Pyrazinamidase	1im5	3.5.1.19	1.65
Cyclic phosphodiesterase	1jh6	3.1.4.-	1.80
Serine endopeptidase	1jhf	3.4.21.88	1.80
Pyruvate dehydrogenase kinase	1jm6	2.7.1.99	2.50
Glycerol-3-phosphate acyltransferase	1k30	2.3.1.15	1.90
Tricorn protease	1k32	3.4.21.-	2.00
3,4-dihydroxy-2-butanone 4-phosphate synthase	1k4l	5.4.99.-	1.60
Protein-disulphide reductase	1l6p	1.8.1.8	1.65
Alpha-amino acid ester hydrolase	1mpx	3.1.1.43	1.90
Myrosinase	1myr	3.2.3.1	1.64
Serine-carboxyl peptidase	1nlu	3.4.21.100	1.30
Limonene-1,2-epoxide hydrolase	1nu3	3.3.2.8	1.75
Limonene-1,2-epoxide hydrolase	1nww	3.3.2.8	1.20
Deoxyribose-phosphate aldolase	1p1x	4.1.2.4	0.99
Protein-tyrosine phosphatase YopH	1pa9	3.1.3.48	2.00
Alginate lyase A1-III	1qaz	3.5.1.45	1.78
UDP-sulphoquinovose synthase	1qrr	3.13.1.1	1.60
Caspase-8	1qtn	3.4.22.-	1.20
Lipase	1r4z	3.1.1.3	1.80
Phosphonoacetaldehyde hydrolase	1rql	3.11.1.1	2.40
Tissue plasminogen activator	1rtf	3.4.21.68	2.30
Alpha-lytic protease	1ssx	3.4.21.12	0.83
Uridine phosphorylase	1t0u	2.4.2.3	2.20
2-hydroxy-6-oxo-7-methylocta-2,4-dienoate	1uk7	3.7.1.9	1.70
5'-nucleotidase	1ush	3.1.3.5	1.73
Phospholipase D	1v0y	3.1.4.4	1.71
7,8-dihydroneopterin aldolase	2dhn	4.1.2.25	2.20
Alpha-lytic protease	2lpr	3.4.21.12	2.25
Peroxidase	7atj	1.11.1.7	1.47

Table 5.4: Protein structures used in the residue-residue analysis.

Table 5.3 describes the distribution of residue types for each function. Proton donors are dominated by Ser and Cys; these cases are all due to systems where Ser (or Cys) is being primed to act as a nucleophile on the substrate, the majority of which are Ser-His-Asp or Cys-His-Asp triads. The high frequency of such systems in this dataset is in turn a result of the fact that such systems have evolved independently many times (Dodson & Wlodawer, 1998). Proton acceptors are dominated by His and the negatively charged Asp and Glu. Cases of charge stabilisation by hydrogen bond donation and acceptance are dominated by positively charged and negatively charged residues respectively.

5.2.5 Residue-residue geometry

The geometry of catalytic residues acting upon other residues was compared with the geometry of hydrogen bonding. The comparison set of non-catalytic hydrogen bonds consisted of all sidechain-sidechain hydrogen bonds between non-catalytic residues from the 60 proteins in the residue-residue dataset. The distributions for the geometric parameters for these hydrogen bonds are shown in the line histograms in Figures 5.11 and 5.12, using the terms defined in Figure 5.1. These line histograms are the same in all both figures, although their vertical scale differs because the frequencies were scaled separately for each graph such that the sum of all frequencies matches the sum of all frequencies for the equivalent catalytic residue parameter. These distributions are similar to those for the hydrogen bond dataset used for the residue-substrate analysis.

The geometry of the catalytic residue operations is shown by the bar histograms in Figures 5.11 (residues which donate protons to other residues) and 5.12 (residues which stabilise charge on other residues by donating hydrogen bonds). The geometry is generally similar both to that of non-catalytic hydrogen bonds and to that of catalytic residues acting directly on substrate. Using a Mann-Whitney nonparametric test found no significant difference ($\alpha = 0.05$; applying Bonferroni correction for the 24 Mann-Whitney tests in this chapter gives $\alpha = 2.08 \times 10^{-3}$) between the distribution of the distance and angle parameters for non-catalytic hydrogen bonds and the hydrogen geometry for the catalytic residues.

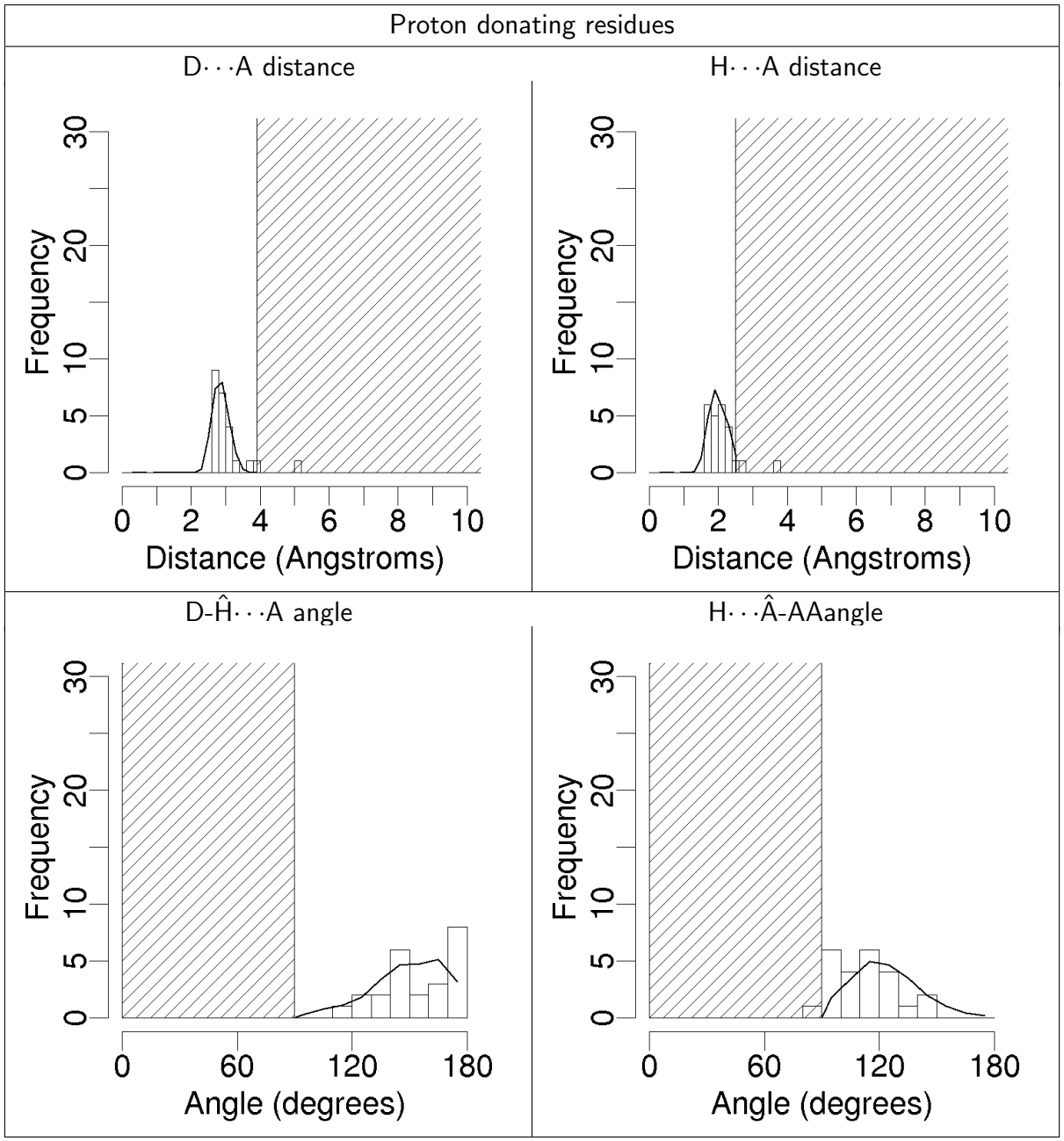


Figure 5.11: Geometry of proton donating residues (acting on other residues) compared to hydrogen bonding.
Bars show distribution of geometric parameters for catalytic residues. Lines show distribution of the same geometric parameters for non-catalytic hydrogen bonds. The frequencies for the distribution of non-catalytic hydrogen bond parameters have been adjusted separately for each graph such that the sum of all frequencies matches the sum of all frequencies for the equivalent catalytic residue parameter. Angles and distances are defined in Figure 5.1. Values for parameters which are outside the criteria for recognition as a hydrogen bond are marked by shaded areas.

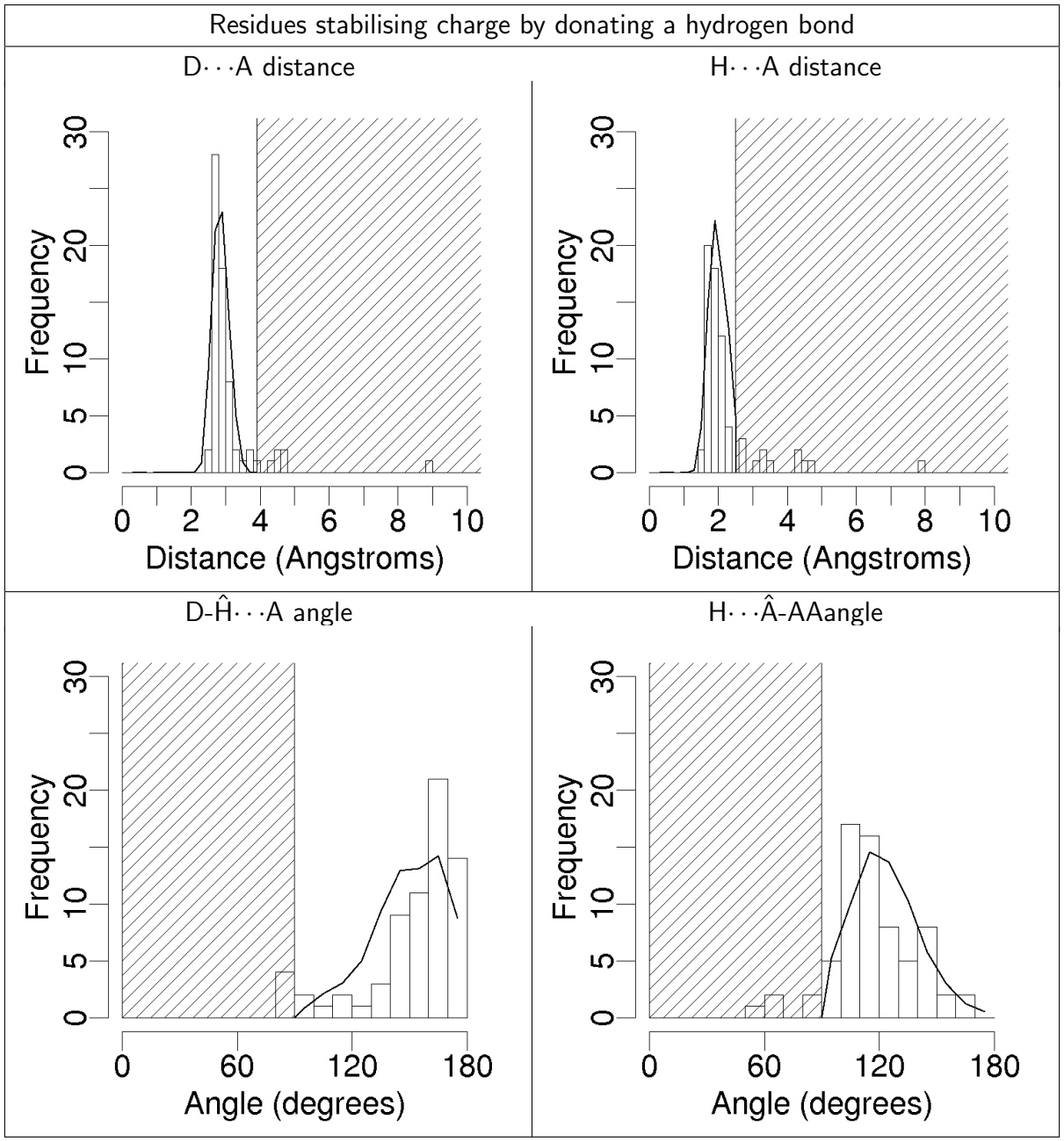


Figure 5.12: Geometry of residues stabilising charge by donating a hydrogen bond (acting on other residues) compared to hydrogen bonding.
Bars show distribution of geometric parameters for catalytic residues. Lines show distribution of the same geometric parameters for non-catalytic hydrogen bonds. The frequencies for the distribution of non-catalytic hydrogen bond parameters have been adjusted separately for each graph such that the sum of all frequencies matches the sum of all frequencies for the equivalent catalytic residue parameter. Angles and distances are defined in Figure 5.1. Values for parameters which are outside the criteria for recognition as a hydrogen bond are marked by shaded areas.

5.2.6 Residue-residue operations with unusual geometry

There are a total of 15 residue-residue operations with geometric parameters which lie outside those permitted for the hydrogen bonds in the non-catalytic dataset.

Four of these cases are proton donor operations. Three of these are Cys-His pairs from proteases where the histidine deprotonates the cysteine in order for the latter to carry out a nucleophilic attack on the substrate. In all three of these cases, the H···A distance is slightly above 2.5 Å, although other parameters are within normal limits. These three proteins are not related to one another, and are the only such Cys-His proton-exchanging pairs in the dataset. This larger distance is almost certainly due to the large van der Waals radius of the sulphur in cysteine (1.80 Å) compared to other elements acting as hydrogen bond donors in this study (nitrogen at 1.55 Å and oxygen at 1.52 Å) (Bondi, 1964).

The other proton donor operation with an unusual geometry is a Ser-His pair where the histidine deprotonates the serine to prepare the latter for a nucleophilic attack. This residue pair has a H···Â-AA geometry slightly below 90°, although other parameters are within normal limits. There are six other Ser-His proton exchanging pairs in the dataset, all of which also act to prepare the serine to act as a nucleophile. None of these have geometric parameters outside the normal limits for hydrogen bonds, although their H···Â-AA angles tend to be slightly low (mean 98.6°, standard deviation 11.9°). There is nothing unusual about their H···A distances (mean 1.92 Å, standard deviation 0.29 Å).

The other 11 residue-residue operations with unusual geometries are charge stabilisers through hydrogen bond donation. One of these appears to be caused by a conformational change occurring during the reaction (Scharff *et al.*, 2001); three are cases where the charge-stabilising function of one of the residues may not be its primary catalytic role. For the other seven examples, there is no clear explanation. These cases without clear explanation include three instances where residues are apparently acting over long distances, with H···A distances exceeding 4 Å. In all these long-distance cases there is mutagenesis evidence for the importance of these residues in catalysis, and there is no indication in the literature that the active site is not in its catalytic conformation. Two of the three cases are interactions between pairs of charged residues, and may be chiefly

electrostatic interactions rather than specific hydrogen bonds.

All the residue operations which have unusual angles also have unusual distances, with the single exception of the Ser-His proton transfer mentioned above.

5.3 Discussion

Most catalytic residues involved in accepting or donating protons or hydrogen bonds correspond closely to a hydrogen-bonding geometry. There is no evidence that residues involved in proton transfer are more restricted in their geometry than residues engaged in hydrogen bonding, either in terms of very short distances or highly constrained angles. Since the dataset used here is relatively small, it is not possible to say definitively that the geometry distribution for all such residues is identical to that for hydrogen bonds. However, any difference must be subtle.

There are a small number of residue operations with a more unusual geometry. These unusual residue geometries occur for a variety of reasons; many have no obvious explanation. Since residue operations with unusual angles nearly always have unusually large distances, it is possible that these residues are not in their catalytic conformation. They may have conventional hydrogen-bonding angles when they are at a conventional distance from their target. It is possible that in some cases there are slight conformational changes associated with catalysis that bring these residues to bear.

There is only one catalytic residue parameter for which the overall distribution differs significantly from the geometry of non-catalytic hydrogen bonds. This is the H$\cdots$Â-AA angle for proton donating residues. Four of the seven residue operations where this angle is lower than the 90° threshold represent donation of a proton to a carbon that is involved in a double bond. If these four cases are removed, the distribution for the proton donor H$\cdots$Â-AA angles becomes similar to the H$\cdots$Â-AA angle distribution for non-catalytic hydrogen bonds. Most of the proton donors in the dataset used here donate a proton to an oxygen atom or a nitrogen atom, and these residues are therefore positioned to align the proton with the lone pair on the acceptor atom— a conventional hydrogen bonding geometry. However, for these cases which donate protons to carbon, there are no lone pairs on the acceptor atom. It appears that this leads to the proton donor being oriented

towards the double bond which provides the electrons for the covalent bond between the carbon and the hydrogen, resulting in an unusually low H···Â-AA angle.

The residue operations in this dataset involving charge stabilisation by hydrogen bonding have often (but not always) been assigned in the literature on the basis of their geometry alone. Thus, it is perhaps not surprising that they do not deviate from hydrogen bonding geometry. Nevertheless, these charge-stabilising residues might in principle have occupied only a part of the spectrum of allowed geometric parameters— yet this is not observed.

In analysing charge-stabilisation interactions, the assumption was made that this charge stabilisation occurs through hydrogen bonding. In the great majority of cases, the geometry of the interaction supports this interpretation. However, there is a small minority of cases where the charge stabilisation involves residues with an overall charge, and whose geometry suggests that the interaction is purely electrostatic, rather than being based on hydrogen bond formation.

The modelling of hydrogen positions assumes that the covalent bond to the hydrogen has its conventional length. Since this analysis considers hydrogens which are involved in hydrogen bonding, and some of which are transferred in the process of catalysis, this assumption may not hold. It is best to view the values given here for H···A distance, D-Ĥ···A angle, and H···Â-AA angle as essentially descriptions of the heavy atom geometries which are cast in terms of assumed hydrogen locations in order to make these descriptions comparable with previous descriptions of hydrogen bonding geometry.

It has been previously noted that a poor hydrogen bonding geometry is often seen between the Ser-His pair in Ser-His-Asp catalytic triads (Lamba *et al.*, 1996; Dodson & Wlodawer, 1998). There is one case in the dataset used here where the H···Â-AA angle for the Ser-His pair in a serine protease is unusual. However, most Ser-His pairs in this dataset do not display an unusual hydrogen bonding geometry, although it is conceivable that they have a weak hydrogen bonding strength despite being within the normal geometric parameters limits for hydrogen bonding.

The dataset analysed here includes a range of residue functions that do not involve hydrogens. Unfortunately, there are only a few instances of each of these functions, so it is difficult to extract useful trends.

Enzymatic catalysis does not appear to depend on individual catalytic residues with special geometry; the geometry of catalytic residues with respect to their targets resembles that of a non-catalytic hydrogen bond. This raises the question of what causes these catalytic residues to engage in catalytic operations where binding residues do not. This is discussed as part of the general conclusions in Chapter six.

5.4 Methods

5.4.1 Residue-substrate dataset selection

The residue-substrate dataset consisted of enzyme structures with all substrates or products present, where the enzyme had a relative whose mechanism was known and which was in the MACiE database version 2.0. These relatives were selected, and equivalent catalytic residues identified, using the method employed by the Catalytic Site Atlas database for annotating relatives, as described in Chapter two. The protein structures used in this dataset were required to perform the same reaction as their relative in MACiE; this was regarded as satisfied if they had an identical EC classification to the fourth level.

As stated above, structures were only used if all the substrates or products were present. The only exception to this rule was where the missing substrate or product was a very small molecule: water, oxygen, CO_2 or H_2O_2. Cofactors such as NAD were regarded as substrates. Substrate/product analogues were permitted where it appeared that any atoms directly interacting with the catalytic residues would be in the same positions as if the true substrate/product were present.

In some cases, the chemical changes occurring in one reaction step implied a structural change so large that it was not possible to use a substrate complex as a guide to the geometry involved in later reaction steps. Where this was the case, the analysis only considered those reaction steps between the substrate complex (or the product complex, if that was the structure used) and the step where the structural change occurred. A substantial structural change was viewed as having occurred if the substrate atom involved was transferred as part of a group (such as a phosphate group), or if there was any large-scale conformational change in the substrate, such as the opening of a ring or a whole

compound being rotated in the active site. Residue operations occurring in later steps that involved affected substrate atoms were not included in the dataset.

5.4.2 Residue-residue dataset selection

The residue-residue dataset was drawn from those enzyme structures in the Catalytic Site Atlas database for which residue-residue interaction data was available.

5.4.3 Redundancy and quality constraints on both datasets

In order to ensure a non-redundant dataset, all structures within each of the two datasets were required to have a unique combination of catalytic domain (as classified in the CATH database (Pearl *et al.*, 2005)) and third-level EC classification. Additionally, where two structures had the same catalytic domain classification in CATH, but had *different* third-level EC classifications, the reactions were checked manually to establish that there was no apparent homology between them.

All structures were determined using X-ray crystallography and had a resolution of 2.5 Å or better, and an R-value of 0.25 or better. All atoms in protein residues analysed had an occupancy of 1, indicating that their position was unambiguous. The occupancy of the substrate/product/analogue was 1 in almost all cases; there were only three exceptions. These cases had the PDB identifiers 1chm, 1ew8, and 4req. For the last two of these, the reason was that a mixture of substrate and product was present and both were represented superposed with occupancy below 1.

5.4.4 Non-catalytic hydrogen bond geometry

The geometry of non-catalytic hydrogen bonds was determined using the program HBPLUS (McDonald & Thornton, 1994), with parameters D$\cdots$A distance < 3.9 Å, H$\cdots$A distance < 2.5 Å, H$\cdots\hat{\text{A}}$-AA angle and D-$\hat{\text{H}}\cdots$A angle both $> 90°$. Hydrogen bonds were excluded from this dataset if they involved residues identified as catalytic based in the MACiE database (for the dataset used in the residue-substrate analysis) or the CSA database (for the dataset used in the residue-residue analysis), even if these residues were not included in the dataset of residue operations analysed in this study.

5.4.5 Catalytic hydrogen placement

In order to determine the geometry of hydrogens for those catalytic residue operations whose function involved hydrogens, it was necessary to estimate the position of these hydrogens. All hydrogens were positioned using an algorithm implemented by the present author which was largely identical to that employed by the program HBPLUS (McDonald & Thornton, 1994; McDonald, 1995). This algorithm positions the hydrogen in such a manner as to minimise its distance from the acceptor atom, given the known distance and angle parameters for the covalent bond between the hydrogen and the donor atom. The algorithm imposes no constraints on the D-$\hat{\text{H}}\cdots$A and H$\cdots\hat{\text{A}}$-AA angle parameters; it attempts to minimise the H$\cdots$A distance. The algorithm used here differed from that in HBPLUS in that the acceptor atom of the hydrogen bond was determined by the nature of the residue operation, and not by any geometric consideration.

Chapter 6

Conclusions

This final chapter discusses the limitations imposed on the work presented in this thesis by the data employed. It also compares the results of the different chapters in terms of the general themes of evolution and protein function prediction, along with a discussion of possible directions for future work in each of these areas.

6.1 Data employed

6.1.1 The necessity of small datasets

Several portions of the work described in this thesis used relatively small datasets. This is not a unique circumstance; in the early days of the collection of protein sequence data, protein structure data, and many other forms of biochemical data, analyses were carried out on relatively small datasets because of the paucity of data available.

In the case of the analysis of the geometry of catalytic residues relative to their targets (Chapter five) the use of a relatively small dataset (42 structures for the residue-substrate analysis) was unavoidable. MACiE is the largest database of catalytic mechanisms available. The size of this database is limited by the need to manually extract information from the literature. Assembling a single MACiE entry takes a PhD student roughly a week. The dataset which was used in Chapter five is smaller than the whole of MACiE (respectively) because of the necessary additional constraints on the dataset which were described in that chapter, principally the need to obtain structures with a bound

substrate, product, or analogue.

The same types of constraint apply to the data employed in the analysis of catalytic sites drawn from the CSA in Chapter three. However, the CSA is considerably more wide-ranging than MACiE because individual entries require less detail. In consequence, the size of the dataset (147 CSA families) was not generally a limitation on that work, although the ability to reach general conclusions about the geometry of sites with more than four catalytic residues was limited by the number of cases available.

With hindsight, the size of the metal binding site dataset used in Chapter four (11 calcium binding sites and 7 zinc binding sites) could perhaps have been increased. It is possible to derive sets of metal-binding residues from structure alone in an automated manner, unlike catalytic residues. However, these metal-binding residues will be unreliable in several respects: the metal binding site may not be biologically meaningful; for biologically meaningful metal binding sites, the metal seen in the structure may not be the cognate metal; occasionally uncertainties in structure determination may mean that a residue appears to be involved in binding the metal when this is not actually the case. Furthermore, an automated method will not reveal the function of the metal. These pieces of information were available for the dataset used in this thesis because of Dr MacArthur's careful inspection of high-resolution structures. Nevertheless, it would be possible to assemble a larger dataset of metal binding sites in a semi-automated manner, accepting a small amount of error on the issues just mentioned.

6.1.2 Difficulties arising from the use of small datasets

There are several difficulties associated with small datasets: they may be unrepresentative, they may necessitate grouping together potentially dissimilar entities for analysis, and their small size will limit the power of any statistical test carried out upon them. However, as described below, these factors are only likely to have had a small impact on the results presented in this thesis.

The metal binding site analysis is focused on zinc and calcium binding sites serving structural roles, and its conclusions may not generalise to other types of metal binding site. By contrast, the annotation efforts of both the CSA and MACiE are geared towards

obtaining as broad a coverage as possible in terms of EC number/CATH domain combinations. Whilst this cannot eliminate the possibility that the datasets in this thesis which are derived from these databases are in some way unrepresentative of enzymes in general, it reduces the likelihood. Furthermore, the work described in this thesis employs nonredundant datasets, which should reduce the danger that the datasets are biased towards one particular group of relatives.

In the analysis of catalytic residue geometry presented in Chapter five, a range of different residue types and substrate types were grouped together for analysis purposes. It is possible that further geometric patterns might be visible if it were practical to break the dataset down by catalytic residue type and using some chemical classification of substrate types; however, since there were often only a handful of cases in the dataset of a particular residue performing a particular function, this was not possible.

In the case of the comparison of catalytic residue geometry with hydrogen bonding geometry in Chapter five, no significant difference was detected for most parameters. However, the relatively small size of the dataset meant that any subtle differences in geometry distribution could not be detected; a larger dataset would permit greater statistical power. The analyses of possible causes of structural variations in catalytic sites and metal binding sites were also limited in their statistical power by the sizes of the datasets involved. However, in all these cases any patterns that could not be detected by the statistical tests employed are likely to be very weak.

6.1.3 Annotating enzymes and metal binding sites

The annotations of catalytic residues in the CSA inevitably include a small number of uncertain cases. Any assignment of a residue as "catalytic" in the scientific literature is a hypothesis based on more or less experimental evidence. In some cases this experimental evidence may be ambiguous; it may also be the case that different research groups have reached different conclusions, and the annotator must then choose between them.

Sometimes a residue may be proposed as catalytic on the basis of weak experimental evidence, such as its positioning and conservation alone. Whether this is judged sufficient to annotate a residue as catalytic will vary: this would be weak evidence for a residue

playing a key role in catalysis, but for a peripheral residue (such as one stabilising the charge of another residue) it may be sufficient, particularly since such peripheral residues are less likely to have their catalytic status confirmed by other methods such as site-directed mutagenesis.

There is inevitably a subjective element in making the decisions described above, for a minority of difficult cases. It is important that the approach taken should be consistent between enzymes and between annotators. Since the CSA has been assembled by 26 individuals over a period of six years, perfect consistency is not possible. However, over the last few years, consistency has been promoted not only by laying down explicit guidelines for annotation, but also by using past annotations as models, by ensuring that annotators cross-check a proportion of one another's entries, and by encouraging annotators to consult with one another and with other members of the research group regarding difficult cases.

There are also subjective issues in the design of the CSA, in terms of which residues are designated as catalytic and the functional labels that are applied to them. Evolution uses any chemical means for carrying out catalysis which come to hand, and these do not always fit neatly into the categories which we use. Are residues which act as an electron tunneling medium catalytic? Are residues that sterically hinder the formation of undesired products catalytic, or should they be considered binding residues, despite the fact that they may only come into contact with the substrate in later stages of the reaction? In both cases the CSA does label these residues catalytic, but it is possible to argue otherwise.

It should be stressed that the above caveats apply only to a minority of difficult cases, some of which were discussed in more detail in the analysis of residue-function combinations in the CSA in Chapter two. The majority of residues in the CSA are unambiguously catalytic and have a clearly defined function.

The annotations in MACiE have always been assembled by two annotators working closely together and cross-checking one another's work (earlier entries were annotated by Gemma Holiday working with Gail Bartlett, later entries by Gemma Holliday working with Daniel Almonacid). As a result, MACiE is highly self-consistent, and there is less possibility of human error than in the CSA. However, the issues described above still apply to a lesser degree. Moreover, it is sometimes the case that multiple catalytic mech-

anisms are compatible with the available experimental evidence, and there is an element of subjectivity in choosing which to record in MACiE.

The metal binding site dataset was produced solely by Malcolm MacArthur, and this mitigates against inconsistency. Metal binding sites are often a secondary consideration when a structure is determined, and the authors for the structure may not be concerned with the function of a metal binding site, or with whether the metal that is bound is the cognate metal. This must often be inferred by comparison with structures and literature descriptions of homologous proteins. However, the subset of metal binding proteins which was used for this work was confined to those where Dr MacArthur considered that the function and cognate metal type were unambiguous.

6.1.4 Small structural variations and experimental uncertainty

Much of the work in this thesis is concerned with small differences between structures. This raises the question of the extent to which experimental uncertainties in protein structure are responsible for these differences, and whether these uncertainties drown out any meaningful signal.

Unfortunately, this question is difficult to answer rigorously. As mentioned in the introduction, it is not straightforward to cast uncertainties in protein structure determination in terms of coordinate standard uncertainties, and these standard uncertainties are not reported for all structures. Furthermore, the variables which are used to quantify structural differences in this thesis are compound measures such as distance, angle, and RMSD. It is mathematically difficult to translate a given level of coordinate uncertainty into a projected uncertainty in these compound measures. In any case, any attempt to do so would need to assume that the coordinates each individual atom are independent of one another; this is not the case, because some atoms are closely connected to others by covalent bonds, and because the geometry between less closely-connected atoms may have been taken into account in the process of building the structural model. The implications of superposing two structures are also difficult to take into account.

It is possible to obtain a very crude ballpark figure for uncertainty of distances. As noted in Chapter one, the median standard coordinate uncertainty reported in PDB files

is 0.28 Å(Laskowski, 2003) (although standard uncertainties for individual structures vary considerably). If one considers only a single axis and assumes that the coordinates of the two atoms are independent, this implies a standard uncertainty for a distance between two atoms of 0.4 Å (because the variance of a difference between two variables is equal to the sum of the variances of the two variables).

This very rough figure is similar in magnitude to the standard deviation of the distances observed between catalytic residues and their target atoms in Chapter five. This suggests that it is possible that the observed variation in catalytic residue geometry between individual cases is largely due to structural uncertainty. It also suggests that it is conceivable that there are subtle differences between the geometry of catalytic residues relative to their substrates and the geometry of hydrogen bonding, but that these differences are undetectable due to experimental uncertainty. Nevertheless, this degree of uncertainty is relatively small compared to the overall distances between the catalytic residues and their target atoms in the substrate.

For the RMSD figures employed in Chapters three and four, it is not possible to obtain even such a crude approximation for the likely effects of experimental uncertainty. However, some gauge may be obtained by looking at those cases where different structures of the same enzyme or metal binding site (or structures of very closely related proteins) are compared. Other than in cases where the structures differ due to factors such as binding of an allosteric effector, these structural differences are usually small— generally below 0.2 Å RMSD for C_α/C_β templates.

6.2 Evolution of functional sites

The work described in Chapter three demonstrates that most catalytic sites are highly structurally conserved; most differ by less than 1 Å RMSD even in distantly related enzymes. The same is true of the structural calcium and zinc binding sites surveyed in Chapter four. Furthermore, the analysis described in Chapter five suggests that the locations of catalytic residues relative to their substrates fall within a limited range of possibilities, with the standard deviation of the distances between catalytic residues and their targets being well below 1 Å for most functions, and angles of approach for residues

whose functions involve the sharing or transfer of hydrogens falling in the same limited range observed for residues forming hydrogen bonds. It appears that the constraints which chemistry and physics impose on protein structure at these functional sites are quite rigid.

6.2.1 Divergent evolution

Structural variation was very similar for catalytic sites and zinc binding sites with the same number of residues, using C_α/C_β atoms for comparison. (Four-residue catalytic sites have a mean RMSD of 0.29 Å, compared to 0.25 Å for the four-residue zinc binding sites.) Because catalytic sites (as defined in this thesis) tend to be smaller than metal binding sites, four-residue sites are the only case where there is extensive data available for both catalytic and metal binding sites. It is difficult to say to what extent this similarity in behaviour can be generalised to other metal types and sizes of site. The very slightly greater structural variability of the catalytic sites might be explained by the fact that the catalytic site dataset included structures both with and without substrate bound, whereas all the metal binding sites had metal bound.

The structural conservation of catalytic sites and metal binding sites is so strong that that in most cases there was little or no discernible relationship between the structural similarity of the sites in a pair of homologues and their sequence similarity. C_α and C_β atoms are structurally more conserved than functional atoms for catalytic sites. For metal binding sites, the structural variability of C_α/C_β atoms and metal binding atoms appear to be similar.

Despite this structural conservation of sites with similar function, where function changes, structure changes also. The analysis of structural variation in catalytic sites described in Chapter three included cases such as the catechol 2,3-dioxygenase CSA family, where a relative (homoprotocatechuate 2,3-dioxygenase) differed in function and also had a subtly different catalytic site structure. A more thorough investigation of how catalytic site structure varies between relatives with different functions might shed more light on the connection between catalytic site geometry and function. However, there are myriad ways in which catalytic sites can differ in function, and in many cases it might be difficult to disentangle the structural effects of a difference in catalytic chemistry from

the (in some ways more straightforward) changes imposed on catalytic site geometry by binding to a different substrate. Thus, it is not obvious how this interesting topic could be studied systematically.

The loss of metal binding sites illustrates an extreme case of the relationship between change in the purpose of a functional site and change in its structure. Where a relative of a calcium binding protein does not bind that calcium, the equivalent site in the protein typically differs by a few point mutations; for similar cases involving zinc, the lack of metal binding involves loss of entire secondary structural elements and loops. One explanation, as described in Chapter four, is that the zincs are more fundamental to the protein's structural stability than the calciums, and that their loss therefore requires more restructuring of the protein to maintain its stability. It would be interesting to see how this applies to sites binding other metals, although it should be noted that the approach used in this thesis involved a significant amount of manual effort, and it is not easy to see how it could be automated to cover a larger dataset. The structural consequences of this kind of loss of a functional site can also be investigated for enzymes; Bartlett *et al.* (2003) examined the differences between the catalytic sites of three enzymes and the equivalent sites in their non-enzyme relatives. They found that in each case, some features of the catalytic site were retained in the non-enzyme relative, including catalytic residues and catalytic metals. This small sample suggests that, as one might expect, the loss of a catalytic site has fewer structural consequences than the loss of a site which serves to maintain the protein's stability.

The analyses of structural change in catalytic sites and metal binding sites presented in Chapters three and four only examined relatives where the residue identity was conserved. It is conceivable that catalytic sites and metal binding sites show more structural variation between relatives where these residue types differ. An investigation of the frequency and structural and functional consequences of changes in the identity of residues in these sites might prove interesting.

6.2.2 Convergent evolution

If proteins are constrained in the way that they solve a given functional problem, does this mean that it is common for unrelated proteins to convergently evolve structurally similar sites to accomplish similar functions? The analysis of convergent evolution of metal binding sites showed that, whilst many unrelated structural zinc binding sites feature the same arrangement of residues, this is less common for calcium binding sites. Whilst these calcium binding sites typically involve similar conformations of oxygen atoms, this seldom equates to similar arrangements of residues. It appears that once evolution has selected a given structural solution to a functional problem, this remains closely structurally conserved, but that there are many different solutions which might be employed.

Convergent evolution of overall enzyme function (as described by EC number) is a common phenomenon, as shown by the comparison of EC numbers and CATH domain codes in Chapter two. This raises the question of how often convergent evolution in enzymes, where it occurs, is reminiscent of convergent evolution of zinc binding sites ("strict" convergent evolution, with similar residues in a similar geometry playing similar functional roles), how often it is reminiscent of convergent evolution of calcium binding sites ("mechanistic" convergent evolution, with different residues presenting similar functional groups in a similar geometry to play similar functional roles), and how often it involves entirely different catalytic mechanisms ("functional" convergent evolution). This question is complicated by the fact that there is a continuum between these options, with certain elements of geometry and mechanism recurring in otherwise dissimilar enzymes.

The classic case of strict convergent evolution of enzymes is the hydrolases using a Ser-His-Asp triad. There are at least six different folds where a Ser-His-Asp triad has independently evolved, employing the same residues in the same geometry to perform the same chemical functions. There are also cases of mechanistic convergent evolution with this triad, which use different residues in a similar geometry to perform the same chemical roles. Some use cysteine or threonine in place of the serine; some use aspartate in place of the glutamate (Dodson & Wlodawer, 1998). However, there also cases of functional convergent evolution such as the β-lactamases where unrelated enzymes have evolved entirely different catalytic mechanisms for carrying out the same overall reaction. It is difficult

to determine by searching the literature how prevalent these different types of convergent evolution are; it is difficult to carry out a comprehensive search of the literature for this type of topic, particularly since different authors use the term "convergent evolution" to describe very different phenomena. There is no general review publication that assembles cases of convergent evolution of catalytic sites. Furthermore, it is possible that there are pairs of unrelated enzymes whose mechanism is known but whose similarity has not yet been realised; naturally, these could not be discovered by searching the literature. Structural bioinformatics methods could be applied to search the PDB for cases of convergent evolution; structural templates are one obvious approach. This is discussed further below.

Using structural templates to search for cases of strict convergent evolution can succeed, as shown by a number of publications that use template matching to identify cases of the Ser-His-Asp triad. It is an open question whether other cases of convergent evolution are structurally similar enough to one another for structural templates to be able to reliably discriminate them from random matches. As stressed above, convergent evolution of both catalytic sites and metal binding sites can occur on many levels. It might be useful to gather together a dataset of known cases of convergent evolution, (using a variety of definitions of the term "convergent evolution") and to observe the ways in which residue types and geometries are similar and different. This would be of interest in itself, and it would also allow one to ask how one could design structural templates using the more flexible structural template matching programs (such as Jess) to capture this diversity. One could then ask whether such structural templates are capable of discriminating meaningful matches from random ones. However, it must be borne in mind that assembling such a dataset would be difficult, for the reasons discussed in the previous paragraph.

6.3 Function prediction

As described in Chapter one, prediction of protein function purely on the basis of structural features is a difficult task. Most methods for protein function prediction make some use of homology. Catalytic templates have the potential to recognise both homologues and cases of convergent evolution. As explained in the introduction to Chapter four, metal binding site templates are less likely to be relevant to function prediction.

6.3.1 Function prediction using templates to identify homologues

Many methods of structural template matching have been developed, but most tests of their usefulness for function prediction have examined just a few individual cases. In Chapter three, a library of structural templates was produced which can discriminate between matches to related proteins and random matches with over 85% sensitivity and predictive accuracy. The usefulness of statistical significance measures for scoring template matches was investigated; these were found to be superior to RMSD for scoring matches to five-residue templates, although there was little benefit to the use of statistical significance measures for smaller templates. Templates which represented residues using their C_α/C_β atoms were found to be better able to discriminate between relatives and random matches than templates composed of functional atoms.

This library of structural templates has been made publically available as part of the CSA, including performance statistics for each template (sensitivity, predictive accuracy and optimal threshold RMSD). A webserver is also available which permits structures to be uploaded and searched against the set of representative templates (www.ebi.ac.uk/thornton-srv/databases/CSS). It supplies RMSDs and statistical significance values for all matches.

Structural templates can be useful in understanding proteins with sequence homologues of known function, for the reasons mentioned in the introduction to Chapter three: resolving ambiguous alignments, dealing with sites spread over multiple chains, and helping confirm or reject a hypothesis of similar function to the homologue. Nonetheless, structural templates are potentially most useful for detecting very distant homologues or instances of convergent evolution. The work in Chapter three defined families using sequence relatives because this was the most practical way to carry out a large-scale study. In principle, homologues that were detectable on the basis of structure alone could be used as a dataset, but it would then be a more difficult and uncertain task to assign equivalent catalytic residues, both because of possible alignment problems and because the enzymes would be less likely to share a function. This problem could be avoided by manually checking structural alignments and enzyme functions, or by referring to the literature to check the catalytic residues for every homologue. However, this manual element would

probably limit such a study to a smaller dataset than the one used in Chapter three.

6.3.2 Function prediction using templates to identify cases of convergent evolution

Establishing the ability of structural template methods to detect cases of convergent evolution is even more difficult than studying their usefulness for looking at distant homologues, because (as described above) it is difficult to obtain a dataset of cases of convergent evolution. Smaller-scale assessments of well-defined enzyme families, such as the study of Ser-His-Asp triad hydrolases by Wallace *et al.* (1996) and the study of the enolase superfamily by Meng *et al.* (2004), may prove a useful way to investigate the effectiveness of templates in detecting convergence and extreme divergence. The template library created as part of the work described in Chapter three may prove a useful resource for such studies.

As discussed above, it is possible that structural templates with more flexibility in terms of residue types or geometry might be more effective in detecting cases of convergent evolution, although these might also be less able to distinguish genuine cases from random matches.

6.3.3 Comparisons of structural templates with other methods

If one constructed a dataset of cases of convergently evolved catalytic sites, or of catalytic sites in proteins that have retained similar function despite substantial divergence of the overall sequence and fold, then it would be interesting to carry out a thorough comparison of the ability of structural template methods to predict the function of these proteins relative to the effectiveness of other methods such as the cleft or surface patch comparison, or *ab initio* methods described in the introduction. It would also be interesting to compare the effectiveness of structural templates of the type employed in this thesis (which only describe the catalytic residues) with an approach like that used in the ProFunc server, which also compares the residue context of the template matches (Laskowski *et al.*, 2005). One might expect that methods which take context into account would work better on close relatives, and worse on cases of convergent evolution; it would be interesting to see

how the two types of approach compare at different levels of evolutionary divergence.

Structural templates are not simply a competitor to other methods of function prediction, but can be complementary. The library of structural templates representing catalytic sites described in Chapter three forms part of the meta-server ProFunc (Laskowski *et al.*, 2005) which collates the results of many separate function prediction methods. It might be possible to devise function prediction methods that integrate template matching more thoroughly with other function prediction approaches.

6.3.4 Predicting enzyme mechanisms

The analysis presented in Chapter five described the geometry of catalytic residues relative to their substrates. This is typically indistinguishable from the geometry of a hydrogen bond. What causes these catalytic residues to engage in catalytic operations where binding residues do not? An important part of the answer must be that individual catalytic residues do not act alone. All the enzymes in the dataset analysed in Chapter five employed multiple catalytic residues or cofactors, acting on the substrate or one another. The results presented in this thesis thus emphasise the cooperativity involved in catalysis: enzyme function does not require a single residue with finely-tuned geometry, but rather a whole cascade of interactions.

What does this imply for function prediction? The description of catalytic residue geometry presented here may prove useful to those making speculations about the mechanism of a enzyme based on its structure. Given the similarity of catalytic residue geometry to hydrogen bonding geometry, this information about typical geometry is mainly likely to be helpful in ruling out residues with an unusual location rather than separating catalytic residues from binding residues. Catalytic residues must also be distinguished according to whether the residue type is a common one for the proposed catalytic operation (statistics on this point are also reported in Chapter five), according to the chemistry of the substrate itself, according to the nature of the other residues surrounding the proposed catalytic residue, and in the context of an overall hypothetical reaction mechanism.

Might it be possible to produce a complete automated method for predicting the catalytic mechanism of an enzyme from structure? At best, any such prediction could only

hope to be a speculation about mechanism with an associated probability, which would need to be confirmed by experiment. Even this would probably require the substrate or a close analogue to be bound to the structure, which would limit its applicability. Nevertheless, it is not inconceivable that a system could be constructed that would be of use in providing such speculations to enzymologists working with a structure, but a considerable other number of aspects of the relationship between enzyme mechanism and protein structure would need to be researched first. These aspects would include the typical geometry between common cofactors and the residues and substrate they interact with; whether residues with a given catalytic function are distinctive from merely binding residues in terms of the residues which surround them, and their predicted pK_a; and a study of the different mechanisms (in terms of catalytic operations on the substrate) which enzymes employ to carry out a given overall chemical change from substrate to product. All of these research topics would be of interest in their own right. The dataset of protein structures with bound substrates, products, or analogues which was assembled for the analysis described in Chapter five might prove a useful starting point for any such analysis.

Publications arising from this work

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

- Torrance, J., Holliday, G., Mitchell J. & Thornton J. (2007) The geometry of interactions between catalytic residues and their substrates. *J Mol Biol*, **369**, 1140–52.

- Torrance, J., MacArthur, M. & Thornton, J. Evolution of binding sites for zinc and calcium ions playing structural roles. *Proteins*, in press.

- Najmanovich, R., Torrance, J. & Thornton, J. (2005) Prediction of protein function from structure: insights from methods for the detection of local structural similarities. *Biotechniques*, **38**, 847–51.

- Holliday, G., Almonacid, D., Bartlett, G., O'Boyle N., Torrance J., Murray-Rust P., Mitchell J. & Thornton J. (2007) MACiE (Mechanism, Annotation and Classification in Enzymes): novel tools for searching catalytic mechanisms. *Nucleic Acids Res*, **35**, D515–20.

References

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

Altschul, S., Gish, W., Miller, W., Myers, E. & Lipman, D. (1990). Basic local alignment search tool. *J Mol Biol*, **215**, 403–10.

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

Amitai, G., Shemesh, A., Sitbon, E., Shklar, M., Netanely, D., Venger, I. & Pietrokovski, S. (2004). Network analysis of protein structures identifies functional residues. *J Mol Biol*, **344**, 1135–46.

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

Arakaki, A., Zhang, Y. & Skolnick, J. (2004). Large-scale assessment of the utility of low-resolution protein structures for biochemical function assignment. *Bioinformatics*, **20**, 1087–96.

Artymiuk, P., Poirrette, A., Grindley, H., Rice, D. & Willett, P. (1994). A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. *J Mol Biol*, **243**, 327–44.

Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., Harris, M., Hill, D., Issel-Tarver, L., Kasarskis, A.,

Lewis, S., Matese, J., Richardson, J., Ringwald, M., Rubin, G. & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, **25**, 25–9.

Attwood, T., Bradley, P., Flower, D., Gaulton, A., Maudling, N., Mitchell, A., Moulton, G., Nordle, A., Paine, K., Taylor, P., Uddin, A. & Zygouri, C. (2003). PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res*, **31**, 400–2.

Auld, D. (2004). *Handbook of metalloproteins*, vol. 3, chap. Structural zinc sites, 403–415. Wiley.

Ausiello, G., Via, A. & Helmer-Citterich, M. (2005a). Query3d: a new method for high-throughput analysis of functional residues in protein structures. *BMC Bioinformatics*, **6 Suppl 4**, S5.

Ausiello, G., Zanzoni, A., Peluso, D., Via, A. & Helmer-Citterich, M. (2005b). pdbFun: mass selection and fast comparison of annotated PDB residues. *Nucleic Acids Res*, **33**, W133–7.

Babbitt, P., Hasson, M., Wedekind, J., Palmer, D., Barrett, W., Reed, G., Rayment, I., Ringe, D., Kenyon, G. & Gerlt, J. (1996). The enolase superfamily: a general strategy for enzyme-catalyzed abstraction of the alpha-protons of carboxylic acids. *Biochemistry*, **35**, 16489–501.

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

Baker, E. & Hubbard, R. (1984). Hydrogen bonding in globular proteins. *Prog Biophys Mol Biol*, **44**, 97–179.

Balaji, S., Sujatha, S., Kumar, S. & Srinivasan, N. (2001). PALI-a database of Phylogeny and ALIgnment of homologous protein structures. *Nucleic Acids Res*, **29**, 61–5.

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

Barrett, A., Rawlings, N. & Woessner, J. (1998). *Handbook of proteolytic enzymes*. Academic Press, San Diego.

Bartlett, G. (2004). *An analysis of enzyme active sites, catalytic residues and mechanisms from a structural and evolutionary perspective.*. Ph.D. thesis, University College London.

Bartlett, G., Porter, C., Borkakoti, N. & Thornton, J. (2002). Analysis of catalytic residues in enzyme active sites. *J Mol Biol*, **324**, 105–21.

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

Belrhali, H., Yaremchuk, A., Tukalo, M., Larsen, K., Berthet-Colominas, C., Leberman, R., Beijer, B., Sproat, B., Als-Nielsen, J. & Grubel, G. (1994). Crystal structures at 2.5 angstrom resolution of seryl-tRNA synthetase complexed with two analogs of seryl adenylate. *Science*, **263**, 1432–6.

Ben-Shimon, A. & Eisenstein, M. (2005). Looking at enzymes from the inside out: the proximity of catalytic residues to the molecular centroid can be used for detection of active sites and enzyme-ligand interfaces. *J Mol Biol*, **351**, 309–26.

Benning, M., Haller, T., Gerlt, J. & Holden, H. (2000). New reactions in the crotonase superfamily: structure of methylmalonyl CoA decarboxylase from Escherichia coli. *Biochemistry*, **39**, 4630–9.

Berman, H., Henrick, K., Nakamura, H. & Markley, J. (2007). The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res*, **35**, D301–3.

Bhaduri, A., Pugalenthi, G. & Sowdhamini, R. (2004). PASS2: an automated database of protein alignments organised as structural superfamilies. *BMC Bioinformatics*, **5**, 35.

Binkowski, T., Adamian, L. & Liang, J. (2003). Inferring functional relationships of proteins from local sequence and spatial surface patterns. *J Mol Biol*, **332**, 505–26.

Blow, D. (2002). *Outline of Crystallography for Biologists*. Oxford University Press, Oxford.

Blow, D., Birktoft, J. & Hartley, B. (1969). Role of a buried acid group in the mechanism of action of chymotrypsin. *Nature*, **221**, 337–40.

Bondi, A. (1964). van der Waals Volumes and Radii. *J Phys Chem*, **68**, 441–451.

Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, **8**, 3–62.

Bradley, P., Kim, P. & Berger, B. (2002). TRILOGY: Discovery of sequence-structure patterns across diverse proteins. *Proc Natl Acad Sci U S A*, **99**, 8500–5.

Brannigan, J. & Wilkinson, A. (2002). Protein engineering 20 years on. *Nat Rev Mol Cell Biol*, **3**, 964–70.

Brenner, S. (2001). A tour of structural genomics. *Nat Rev Genet*, **2**, 801–9.

Brenner, S., Chothia, C. & Hubbard, T. (1998). Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc Natl Acad Sci U S A*, **95**, 6073–8.

Brown, R. (2005). Zinc finger proteins: getting a grip on RNA. *Curr Opin Struct Biol*, **15**, 94–8.

Buchner, E. (1894). Alkoholische gährung ohne hefezellen. *Ber Dtsch Chem Ges*, **30**, 117–24.

Bugg, T. (1997). *An introduction to enzyme and coeznyme chemistry*. Blackwell Science, Oxford.

Bugg, T. (2001). The development of mechanistic enzymology in the 20th century. *Nat Prod Rep*, **18**, 465–493.

Bürgi, H. & Dunitz, J. (1983). From crystal statics to chemical dynamics. *Acc Chem Res*, **16**, 153–161.

Burmeister, W., Cottaz, S., Driguez, H., Iori, R., Palmieri, S. & Henrissat, B. (1997). The crystal structures of Sinapis alba myrosinase and a covalent glycosyl-enzyme inter-

mediate provide insights into the substrate recognition and active-site machinery of an S-glycosidase. *Structure*, **5**, 663–75.

Campbell, R., Mosimann, S., van De Rijn, I., Tanner, M. & Strynadka, N. (2000). The first structure of UDP-glucose dehydrogenase reveals the catalytic residues necessary for the two-fold oxidation. *Biochemistry*, **39**, 7012–23.

Cane, D. (1990). Enzymatic formation of sesquiterpenes. *Chem Rev*, **90**, 1089–1103.

Changela, A., DiGate, R. & Mondragon, A. (2001). Crystal structure of a complex of a type IA DNA topoisomerase with a single-stranded DNA molecule. *Nature*, **411**, 1077–81.

Choe, J., Fromm, H. & Honzatko, R. (2000). Crystal structures of fructose 1,6-bisphosphatase: mechanism of catalysis and allosteric inhibition revealed in product complexes. *Biochemistry*, **39**, 8565–74.

Chothia, C. & Lesk, A. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J*, **5**, 823–6.

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

Cleland, W., Frey, P. & Gerlt, J. (1998). The low barrier hydrogen bond in enzymatic catalysis. *J Biol Chem*, **273**, 25529–32.

Conklin, D. (1995). Machine discovery of protein motifs. *Machine Learning*, **21**, 125–150.

Cook, P., ed. (1991). *Enzyme Mechanism from Isotope Effects.*. CRC Press, Boca Raton, Florida.

Copeland, R. (2000). *Enzymes: A practical introduction to structure, mechanism, and data analysis*. Wiley, New York, 2nd edn.

Copley, S. (2003). Enzymes with extra talents: moonlighting functions and catalytic promiscuity. *Curr Opin Chem Biol*, **7**, 265–72.

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press, Princeton, NJ.

Cruickshank, D. (1999). Remarks about protein structure precision. *Acta Cryst. D*, **55**, 583–601.

Davies, G. & Henrissat, B. (1995). Structures and mechanisms of glycosyl hydrolases. *Structure*, **3**, 853–9.

Devos, D. & Valencia, A. (2000). Practical limits of function prediction. *Proteins*, **41**, 98–107.

Ding, X., Rasmussen, B., Petsko, G. & Ringe, D. (1994). Direct structural observation of an acyl-enzyme intermediate in the hydrolysis of an ester substrate by elastase. *Biochemistry*, **33**, 9285–93.

Dodson, G. & Wlodawer, A. (1998). Catalytic triads and their relatives. *Trends Biochem Sci*, **23**, 347–52.

Drenth, J., ed. (1999). *Principles of Protein X-Ray Crystallography.*. Springer-Verlag, New York.

Dudev, T. & Lim, C. (2003). Principles governing Mg, Ca, and Zn binding and selectivity in proteins. *Chem Rev*, **103**, 773–88.

Dupont, C., Yang, S., Palenik, B. & Bourne, P. (2006). Modern proteomes contain putative imprints of ancient shifts in trace metal geochemistry. *Proc Natl Acad Sci U S A*, **103**, 17822–7.

Durbin, R., Eddy, S., Krogh, A. & Mitchison, G. (1999). *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge University Press, Cambridge, UK.

Eklund, H., Samama, J. & Jones, T. (1984). Crystallographic investigations of nicotinamide adenine dinucleotide binding to horse liver alcohol dehydrogenase. *Biochemistry*, **23**, 5982–96.

Engelhardt, B., Jordan, M., Muratore, K. & Brenner, S. (2005). Protein molecular function prediction by Bayesian phylogenomics. *PLoS Comput Biol*, **1**, e45.

Esnouf, R., Ren, J., Ross, C., Jones, Y., Stammers, D. & Stuart, D. (1995). Mechanism of inhibition of HIV-1 reverse transcriptase by non-nucleoside inhibitors. *Nat Struct Biol*, **2**, 303–8.

Fersht, A. (1999). *Structure and mechanism in protein science: a guide to enzyme catalysis and protein folding*. W.H. Freeman and Company.

Fetrow, J. & Skolnick, J. (1998). Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. *J Mol Biol*, **281**, 949–68.

Fife, T. (1972). General acid catalysis of acetal, ketal, and ortho ester hydrolysis. *Accts Chem Res*, **5**, 264–72.

Finn, R., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., Eddy, S., Sonnhammer, E. & Bateman, A. (2006). Pfam: clans, web tools and services. *Nucleic Acids Res*, **34**, D247–51.

Fischer, D., Wolfson, H., Lin, S. & Nussinov, R. (1994). Three-dimensional, sequence order-independent structural comparison of a serine protease against the crystallographic database reveals active site similarities: potential implications to evolution and to protein folding. *Protein Sci*, **3**, 769–78.

Fischer, E. (1894). Einfluss der configuration auf die wirkung der enzyme. *Ber Dtsch Chem Ges*, **27**, 2985–93.

Flexner, C. (1998). HIV-protease inhibitors. *N Engl J Med*, **338**, 1281–92.

Frausto da Silva, J. & Williams, R. (2001). *The biological chemistry of the elements: the inorganic chemistry of life*. Oxford University Press, Oxford, UK, 2nd edn.

Friedberg, I. (2006). Automated protein function prediction–the genomic challenge. *Brief Bioinform*, **7**, 225–42.

Friedberg, I., Harder, T. & Godzik, A. (2006). JAFA: a protein function annotation meta-server. *Nucleic Acids Res*, **34**, W379–81.

Fuhrmann, C., Daugherty, M. & Agard, D. (2006). Subangstrom crystallography reveals that short ionic hydrogen bonds, and not a His-Asp low-barrier hydrogen bond, stabilize the transition state in serine protease catalysis. *J Am Chem Soc*, **128**, 9086–102.

Galperin, M., Walker, D. & Koonin, E. (1998). Analogous enzymes: independent inventions in enzyme evolution. *Genome Res*, **8**, 779–90.

Gamblin, S., Davies, G., Grimes, J., Jackson, R., Littlechild, J. & Watson, H. (1991). Activity and specificity of human aldolases. *J Mol Biol*, **219**, 573–6.

Garcia-Viloca, M., Gao, J., Karplus, M. & Truhlar, D. (2004). How enzymes work: analysis by modern rate theory and computer simulations. *Science*, **303**, 186–95.

George, R., Spriggs, R., Thornton, J., Al-Lazikani, B. & Swindells, M. (2004). SCOPEC: a database of protein catalytic domains. *Bioinformatics*, **20 Suppl 1**, I130–I136.

George, R., Spriggs, R., Bartlett, G., Gutteridge, A., MacArthur, M., Porter, C., Al-Lazikani, B., Thornton, J. & Swindells, M. (2005). Effective function annotation through catalytic residue conservation. *Proc Natl Acad Sci U S A*, **102**, 12299–304.

Gerlt, J. & Babbitt, P. (2001). Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. *Annu Rev Biochem*, **70**, 209–46.

Gerlt, J. & Gassman, P. (1993). Understanding the rates of certain enzyme-catalyzed reactions: proton abstraction from carbon acids, acyl-transfer reactions, and displacement reactions of phosphodiesters. *Biochemistry*, **32**, 11943–52.

Gerlt, J. & Raushel, F. (2003). Evolution of function in (beta/alpha)8-barrel enzymes. *Curr Opin Chem Biol*, **7**, 252–64.

Gerlt, J., Babbitt, P. & Rayment, I. (2005). Divergent evolution in the enolase superfamily: the interplay of mechanism and specificity. *Arch Biochem Biophys*, **433**, 59–70.

Glusker, A.B., J.P.; Katz (1999). Metal ions in biological systems. *Rigaku J*, **16**, 8–16.

Gorbitz, C. (1989). Hydrogen-bond distances and angles in the structures of amino acids and peptides. *Acta Cryst B*, **45**, 390–395.

Gouaux, J. & Lipscomb, W. (1990). Crystal structures of phosphonoacetamide ligated T and phosphonoacetamide and malonate ligated R states of aspartate carbamoyltransferase at 2.8-A resolution and neutral pH. *Biochemistry*, **29**, 389–402.

Gray, H. & Winkler, J. (1996). Electron transfer in proteins. *Annu Rev Biochem*, **65**, 537–61.

Groth, D., Lehrach, H. & Hennig, S. (2004). GOblet: a platform for Gene Ontology annotation of anonymous sequence data. *Nucleic Acids Res*, **32**, W313–7.

Guex, N. & Peitsch, M. (1997). SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis*, **18**, 2714–23.

Gutfreund, H. (1971). Transients and relaxation kinetics of enzyme reactions. *Annu Rev Biochem*, **40**, 315–44.

Guthrie, J. (1996). Short strong hydrogen bonds: can they explain enzymic catalysis? *Chem Biol*, **3**, 163–70.

Gutteridge, A. & Thornton, J. (2004). Conformational changes observed in enzyme crystal structures upon substrate binding. *J Mol Biol*, **In press**.

Gutteridge, A., Bartlett, G. & Thornton, J. (2003). Using a neural network and spatial clustering to predict the location of active sites in enzymes. *J Mol Biol*, **330**, 719–34.

Hadley, C. & Jones, D. (1999). A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. *Structure*, **7**, 1099–112.

Haft, D., Selengut, J. & White, O. (2003). The TIGRFAMs database of protein families. *Nucleic Acids Res*, **31**, 371–3.

Hamelryck, T. (2003). Efficient identification of side-chain patterns using a multidimensional index tree. *Proteins*, **51**, 96–108.

Hammond, B. & Gutfreund, H. (1955). Two steps in the reaction of chymotrypsin with acetyl-l-phenylalanine ethyl ester. *Biochem J*, **61**, 187–9.

Hanks, S. & Hunter, T. (1995). Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *FASEB J*, **9**, 576–96.

Harding, M. (1999). The geometry of metal-ligand interactions relevant to proteins. *Acta Crystallogr D Biol Crystallogr*, **55 (Pt 8)**, 1432–43.

Harding, M. (2000). The geometry of metal-ligand interactions relevant to proteins. II. Angles at the metal atom, additional weak metal-donor interactions. *Acta Crystallogr D Biol Crystallogr*, **56 (Pt 7)**, 857–67.

Harding, M. (2001). Geometry of metal-ligand interactions in proteins. *Acta Crystallogr D Biol Crystallogr*, **57**, 401–11.

Harding, M. (2004). The architecture of metal coordination groups in proteins. *Acta Crystallogr D Biol Crystallogr*, **60**, 849–59.

Harrison, A., Pearl, F., Sillitoe, I., Slidel, T., Mott, R., Thornton, J. & Orengo, C. (2003). Recognizing the fold of a protein structure. *Bioinformatics*, **19**, 1748–59.

Hartley, B. (1964). Amino-acid sequence of bovine chymotrypsinogen-a. *Nature*, **201**, 1284–7.

Hartridge, H. & Roughton, F. (1923). A method of measuring the velocity of very rapid chemical reactions. *Proc R Soc*, **A104**, 376–94.

Hasson, M., Schlichting, I., Moulai, J., Taylor, K., Barrett, W., Kenyon, G., Babbitt, P., Gerlt, J., Petsko, G. & Ringe, D. (1998). Evolution of an enzyme active site: the structure of a new crystal form of muconate lactonizing enzyme compared with mandelate racemase and enolase. *Proc Natl Acad Sci U S A*, **95**, 10396–401.

Hawkins, T., Luban, S. & Kihara, D. (2006). Enhanced automated function prediction using distantly related sequences and contextual association by PFP. *Protein Sci*, **15**, 1550–6.

Henikoff, J., Greene, E., Pietrokovski, S. & Henikoff, S. (2000). Increased coverage of protein families with the blocks database servers. *Nucleic Acids Res*, **28**, 228–30.

Henikoff, S. & Henikoff, J. (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, **89**, 10915–9.

Henrick, K. & Thornton, J. (1998). PQS: a protein quaternary structure file server. *Trends Biochem Sci*, **23**, 358–61.

Henrissat, B. & Davies, G. (1997). Structural and sequence-based classification of glycoside hydrolases. *Curr Opin Struct Biol*, **7**, 637–44.

Hoersch, S., Leroy, C., Brown, N., Andrade, M. & Sander, C. (2000). The GeneQuiz web server: protein functional analysis through the Web. *Trends Biochem Sci*, **25**, 33–5.

Holliday, G., Bartlett, G., Almonacid, D., O'Boyle, N., Murray-Rust, P., Thornton, J. & Mitchell, J. (2005). MACiE: a database of enzyme reaction mechanisms. *Bioinformatics*, **21**, 4315–6.

Holliday, G., Almonacid, D., Bartlett, G., O'Boyle, N., Torrance, J., Murray-Rust, P., Mitchell, J. & Thornton, J. (2007a). MACiE (Mechanism, Annotation and Classification in Enzymes): novel tools for searching catalytic mechanisms. *Nucleic Acids Res*, **35**, D515–20.

Holliday, G., Almonacid, D., Mitchell, J. & Thornton, J. (2007b). The chemistry of protein catalysis. *J Mol Biol*, **372**, 1261–77.

Holm, L. & Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *J Mol Biol*, **233**, 123–38.

Holm, L., Ouzounis, C., Sander, C., Tuparev, G. & Vriend, G. (1992). A database of protein structure families with common folding motifs. *Protein Sci*, **1**, 1691–8.

Hoppe, A. & Frommel, C. (2003). NeedleHaystack: a program for the rapid recognition of local structures in large sets of atomic coordinates. *J Appl Cryst*, **36**, 1090–1097.

Huber, R. & Bode, W. (1978). Structural basis of the activation and action of trypsin. *Acc Chem Res*, **11**, 114–122.

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

Hutter, M. & Helms, V. (2002). The mechanism of phosphorylation of natural nucleosides and anti-HIV analogues by nucleoside diphosphate kinase is independent of their sugar substituents. *Chembiochem*, **3**, 643–51.

Istvan, E., Palnitkar, M., Buchanan, S. & Deisenhofer, J. (2000). Crystal structure of the catalytic portion of human HMG-CoA reductase: insights into regulation of activity and catalysis. *EMBO J*, **19**, 819–30.

Ivanisenko, V., Pintus, S., Grigorovich, D. & Kolchanov, N. (2004). PDBSiteScan: a program for searching for active, binding and posttranslational modification sites in the 3D structures of proteins. *Nucleic Acids Res*, **32**, W549–54.

Ivanisenko, V., Pintus, S., Grigorovich, D. & Kolchanov, N. (2005). PDBSite: a database of the 3D structure of protein functional sites. *Nucleic Acids Res*, **33**, D183–7.

Jambon, M., Imberty, A., Deleage, G. & Geourjon, C. (2003). A new bioinformatic approach to detect common 3D sites in protein structures. *Proteins*, **52**, 137–45.

Jencks, W. (1975). Binding energy, specificity, and enzymic catalysis: the circe effect. *Adv Enzymol Relat Areas Mol Biol*, **43**, 219–410.

Jencks, W. & Gilchrist, M. (1968). Nonlinear structure-reactivity correlations. the reactivity of nucleophilic reagents towards esters. *J Am Chem Soc*, **90**, 2622–37.

Jencks, W. & Page, M. (1974). "Orbital steering", entropy, and rate accelerations. *Biochem Biophys Res Commun*, **57**, 887–92.

Jensen, R. (1976). Enzyme recruitment in evolution of new function. *Annu Rev Microbiol*, **30**, 409–25.

Jonassen, I., Eidhammer, I., Conklin, D. & Taylor, W. (2002). Structure motif discovery and mining the PDB. *Bioinformatics*, **18**, 362–7.

Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G., Wu, G., Matthews, L., Lewis, S., Birney, E. & Stein, L. (2005). Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res*, **33**, D428–32.

Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K., Itoh, M., Kawashima, S., Katayama, T., Araki, M. & Hirakawa, M. (2006). From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res*, **34**, D354–7.

Kaplan, J., Forbush, B. & Hoffman, J. (1978). Rapid photolytic release of adenosine 5'-triphosphate from a protected analogue: utilization by the Na:K pump of human red blood cell ghosts. *Biochemistry*, **17**, 1929–35.

Karlin, S. & Zhu, Z. (1996). Characterizations of diverse residue clusters in protein three-dimensional structures. *Proc Natl Acad Sci U S A*, **93**, 8344–9.

Karpeisky, M. & Ivanov, V. (1966). A molecular mechanism for enzymatic transamination. *Nature*, **210**, 493–6.

Kawabata, T. (2003). MATRAS: A program for protein 3D structure comparison. *Nucleic Acids Res*, **31**, 3367–9.

Keitel, T., Thomsen, K. & Heinemann, U. (1993). Crystallization of barley (1-3,1-4)-beta-glucanase, isoenzyme II. *J Mol Biol*, **232**, 1003–4.

Kern, D., Eisenmesser, E. & Wolf-Watz, M. (2005). Enzyme dynamics during catalysis measured by NMR spectroscopy. *Methods Enzymol*, **394**, 507–24.

Kim, S., Shin, D., Choi, I., Schulze-Gahmen, U., Chen, S. & Kim, R. (2003). Structure-based functional inference in structural genomics. *J Struct Funct Genomics*, **4**, 129–35.

Kirk, O., Borchert, T. & Fugisang, C. (2002). Industrial enzyme applications. *Curr Opin Biotechnol*, **13**, 345–51.

Kita, A., Kita, S., Fujisawa, I., Inaka, K., Ishida, T., Horiike, K., Nozaki, M. & Miki, K. (1999). An archetypical extradiol-cleaving catecholic dioxygenase: the crystal structure of catechol 2,3-dioxygenase (metapyrocatechase) from Pseudomonas putida mt-2. *Structure Fold Des*, **7**, 25–34.

Klaassen, C., Liu, J. & Choudhuri, S. (1999). Metallothionein: an intracellular protein to protect against cadmium toxicity. *Annu Rev Pharmacol Toxicol*, **39**, 267–94.

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

Kobayashi, N. & Go, N. (1997). A method to search for similar protein local structures at ligand binding sites and its application to adenine recognition. *Eur Biophys J*, **26**, 135–44.

Kortemme, T., Morozov, A. & Baker, D. (2003). An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *J Mol Biol*, **326**, 1239–59.

Koshland, D., Jr (1998). Conformational changes: how small is big enough? *Nat Med*, **4**, 1112–4.

Kraut, D., Carroll, K. & Herschlag, D. (2003). Challenges in enzyme mechanism and energetics. *Annu Rev Biochem*, **72**, 517–71.

Kraut, J. (1977). Serine proteases: structure and mechanism of catalysis. *Annu Rev Biochem*, **46**, 331–58.

Krishna, S., Majumdar, I. & Grishin, N. (2003). Structural classification of zinc fingers: survey and summary. *Nucleic Acids Res*, **31**, 532–50.

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

Kumar, V., Dooley, D., Freeman, H., Guss, J., Harvey, I., McGuirl, M., Wilce, M. & Zubak, V. (1996). Crystal structure of a eukaryotic (pea seedling) copper-containing amine oxidase at 2.2 A resolution. *Structure*, **4**, 943–55.

L., M. & Menten, M. (1913). Die kinetic der invertinwirkung. *Biochem Z*, **49**, 333–69.

Lamba, D., Bauer, M., Huber, R., Fischer, S., Rudolph, R., Kohnert, U. & Bode, W. (1996). The 2.3 A crystal structure of the catalytic domain of recombinant two-chain human tissue-type plasminogen activator. *J Mol Biol*, **258**, 117–35.

Lancaster, C., Kroger, A., Auer, M. & Michel, H. (1999). Structure of fumarate reductase from Wolinella succinogenes at 2.2 A resolution. *Nature*, **402**, 377–85.

Landro, J., Gerlt, J., Kozarich, J., Koo, C., Shah, V., Kenyon, G., Neidhart, D., Fujita, S. & Petsko, G. (1994). The role of lysine 166 in the mechanism of mandelate racemase from Pseudomonas putida: mechanistic and crystallographic evidence for stereospecific alkylation by (R)-alpha-phenylglycidate. *Biochemistry*, **33**, 635–43.

Laskowski, R. (2003). *Structural Bioinformatics*, chap. Structural quality assurance, 273–304. Wiley, New Jersey.

Laskowski, R., Luscombe, N., Swindells, M. & Thornton, J. (1996). Protein clefts in molecular recognition and function. *Protein Sci*, **5**, 2438–52.

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

Le Du, M., Stigbrand, T., Taussig, M., Menez, A. & Stura, E. (2001). Crystal structure of alkaline phosphatase from human placenta at 1.8 A resolution. Implication for a substrate specificity. *J Biol Chem*, **276**, 9158–65.

Leppanen, V., Merckel, M., Ollis, D., Wong, K., Kozarich, J. & Goldman, A. (1999). Pyruvate formate lyase is structurally homologous to type I ribonucleotide reductase. *Structure*, **7**, 733–44.

Letunic, I., Copley, R., Pils, B., Pinkert, S., Schultz, J. & Bork, P. (2006). SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res*, **34**, D257–60.

Lewit-Bentley, A. & Rety, S. (2000). EF-hand calcium-binding proteins. *Curr Opin Struct Biol*, **10**, 637–43.

Lippard, S. & Berg, J. (1994). *Principles of bioinorganic chemistry*. University Science Books.

Luzzati, P. (1952). Traitement statistique des erreurs dans la determination des structures cristallines. *Acta Cryst.*, **5**, 802–810.

Madej, T., Gibrat, J. & Bryant, S. (1995). Threading a database of protein cores. *Proteins*, **23**, 356–69.

Madsen, D. & Kleywegt, G. (2002). Interactive motif and fold recognition in protein structures. *J Appl Cryst*, **35**, 137–139.

Mancia, F., Smith, G. & Evans, P. (1999). Crystal structure of substrate complexes of methylmalonyl-CoA mutase. *Biochemistry*, **38**, 7999–8005.

Mann, H. & Whitney, D. (1947). On a test of whether one of 2 random variables is stochastically larger than the other. *Annals Math Stat*, **18**, 50–60.

Marti-Renom, M., Rossi, A., Al-Shahrour, F., Davis, F., Pieper, U., Dopazo, J. & Sali, A. (2007). The AnnoLite and AnnoLyze programs for comparative annotation of protein structures. *BMC Bioinformatics*, **8 Suppl 4**, S4.

Martin, D., Berriman, M. & Barton, G. (2004). GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes. *BMC Bioinformatics*, **5**, 178.

Matthews, B. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta*, **405**, 442–51.

McDonald, I. (1995). *Computational analysis of intramolecular interactions in proteins*. Ph.D. thesis, University of London.

McDonald, I. & Thornton, J. (1994). Satisfying hydrogen bonding potential in proteins. *J Mol Biol*, **238**, 777–93.

McLachlan, A. (1982). Rapid comparison of protein structures. *Acta Cryst A*, **38**, 871–873.

McPhalen, C., Strynadka, N. & James, M. (1991). Calcium-binding sites in proteins: a structural perspective. *Adv Protein Chem*, **42**, 77–144.

Meng, E., Polacco, B. & Babbitt, P. (2004). Superfamily active site templates. *Proteins*, **55**, 962–76.

Mesecar, A., Stoddard, B. & Koshland, D., Jr (1997). Orbital steering in the catalytic power of enzymes: small structural changes with large catalytic consequences. *Science*, **277**, 202–6.

Mi, H., Lazareva-Ulitsky, B., Loo, R., Kejariwal, A., Vandergriff, J., Rabkin, S., Guo, N., Muruganujan, A., Doremieux, O., Campbell, M., Kitano, H. & Thomas, P. (2005). The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res*, **33**, D284–8.

Michal, G., ed. (1999). *Biochemical Pathways: An Atlas of Biochemistry and Molecular Biology.*. Wiley, New York.

Milik, M., Szalma, S. & Olszewski, K. (2003). Common Structural Cliques: a tool for protein structure and function analysis. *Protein Eng*, **16**, 543–52.

Miller, B., Hassell, A., Wolfenden, R., Milburn, M. & Short, S. (2000). Anatomy of a proficient enzyme: the structure of orotidine 5'-monophosphate decarboxylase in the presence and absence of a potential transition state analog. *Proc Natl Acad Sci U S A*, **97**, 2011–6.

Mitchell, E., Artymiuk, P., Rice, D. & Willett, P. (1990). Use of techniques derived from graph theory to compare secondary structure motifs in proteins. *J Mol Biol*, **212**, 151–66.

Modis, Y. & Wierenga, R. (2000). Crystallographic analysis of the reaction pathway of Zoogloea ramigera biosynthetic thiolase. *J Mol Biol*, **297**, 1171–82.

Moffat, K. & Ren, Z. (1997). Synchrotron radiation applications to macromolecular crystallography. *Curr Opin Struct Biol*, **7**, 689–96.

Momany, F., McGuire, R., Burgess, A. & Scheraga, H. (1975). Energy Parameters in Polypeptides. VII. Geometric Parameters, Partial Atomic Charges, Nonbonded Interactions, Hydrogen Bond Interactions, and Intrinsic Torsional Potentials for the Naturally Occurring Amino Acids. *J Phys Chem*, **79**, 2361–2381.

Mulder, N., Apweiler, R., Attwood, T., Bairoch, A., Bateman, A., Binns, D., Bork, P., Buillard, V., Cerutti, L., Copley, R., Courcelle, E., Das, U., Daugherty, L., Dibley, M., Finn, R., Fleischmann, W., Gough, J., Haft, D., Hulo, N., Hunter, S., Kahn, D., Kanapin, A., Kejariwal, A., Labarga, A., Langendijk-Genevaux, P., Lonsdale, D., Lopez, R., Letunic, I., Madera, M., Maslen, J., McAnulla, C., McDowall, J., Mistry, J., Mitchell, A., Nikolskaya, A., Orchard, S., Orengo, C., Petryszak, R., Selengut, J., Sigrist, C., Thomas, P., Valentin, F., Wilson, D., Wu, C. & Yeats, C. (2007). New developments in the InterPro database. *Nucleic Acids Res*, **35**, D224–8.

Murzin, A., Brenner, S., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, **247**, 536–40.

Nagano, N. (2005). EzCatDB: the Enzyme Catalytic-mechanism Database. *Nucleic Acids Res*, **33**, D407–12.

Najmanovich, R.J., Torrance, J. & Thornton, J. (2005). Prediction of protein function from structure: Insights from methods for the detection of local structural similarities. *BioTechniques*, **38**, 847–851.

Nakatsu, T., Kato, H. & Oda, J. (1998). Crystal structure of asparagine synthetase reveals a close evolutionary relationship to class II aminoacyl-tRNA synthetase. *Nat Struct Biol*, **5**, 15–9.

Nardini, M. & Dijkstra, B. (1999). Alpha/beta hydrolase fold enzymes: the family keeps growing. *Curr Opin Struct Biol*, **9**, 732–7.

Nimrod, G., Glaser, F., Steinberg, D., Ben-Tal, N. & Pupko, T. (2005). In silico identification of functional regions in proteins. *Bioinformatics*, **21 Suppl 1**, i328–37.

Northrop, D. (1975). Steady-state analysis of kinetic isotope effects in enzymic reactions. *Biochemistry*, **14**, 2644–51.

Novotny, M., Madsen, D. & Kleywegt, G. (2004). Evaluation of protein fold comparison servers. *Proteins*, **54**, 260–70.

O'Brien, P. & Herschlag, D. (1998). Sulfatase activity of E. coli alkaline phosphatase demonstrates a functional link to arylsulfatases, an evolutionarily related enzyme family. *J Am Chem Soc*, **120**, 12369–70.

O'Brien, P. & Herschlag, D. (1999). Catalytic promiscuity and the evolution of new enzymatic activities. *Chem Biol*, **6**, R91–R105.

Oldfield, T. (2002). Data mining the protein data bank: residue interactions. *Proteins*, **49**, 510–28.

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

Ordentlich, A., Barak, D., Kronman, C., Ariel, N., Segall, Y., Velan, B. & Shafferman, A. (1995). Contribution of aromatic moieties of tyrosine 133 and of the anionic subsite tryptophan 86 to catalytic efficiency and allosteric modulation of acetylcholinesterase. *J Biol Chem*, **270**, 2082–91.

Orengo, C., Pearl, F. & Thornton, J. (2003). *Structural Bioinformatics*, chap. The CATH structural domain database, 249–272. Wiley, New Jersey.

Oubrie, A., Rozeboom, H., Kalk, K., Olsthoorn, A., Duine, J. & Dijkstra, B. (1999). Structure and mechanism of soluble quinoprotein glucose dehydrogenase. *EMBO J*, **18**, 5187–94.

Page, M. & Jencks, W. (1971). Entropic contributions to rate accelerations in enzymic and intramolecular reactions and the chelate effect. *Proc Natl Acad Sci U S A*, **68**, 1678–83.

Pauling, L. (1946). Molecular architecture and biological reactions. *Chem Eng News*, **24**, 1375–7.

Pavlovsky, A., Williams, M., Ye, Q., Ortwine, D., Purchase, C., 2nd, White, A., Dhanaraj, V., Roth, B., Johnson, L., Hupe, D., Humblet, C. & Blundell, T. (1999). X-ray structure of human stromelysin catalytic domain complexed with nonpeptide inhibitors: implications for inhibitor selectivity. *Protein Sci*, **8**, 1455–62.

Pazos, F. & Bang, J.W. (2006). Computational prediction of functionally important regions in proteins. *Curr Bioinf*, **1**, 15–23.

Pazos, F. & Sternberg, M. (2004). Automated prediction of protein function and detection of functional sites from structure. *Proc Natl Acad Sci U S A*, **101**, 14754–9.

Pearl, F., Bennett, C., Bray, J., Harrison, A., Martin, N., Shepherd, A., Sillitoe, I., Thornton, J. & Orengo, C. (2003). The CATH database: an extended protein family resource for structural and functional genomics. *Nucleic Acids Res*, **31**, 452–5.

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

Pearson, R. (1963). Hard and soft acids and bases. *J Am Chem Soc*, **85**, 3533–3539.

Pennec, X. & Ayache, N. (1998). A geometric algorithm to find small but highly similar 3D substructures in proteins. *Bioinformatics*, **14**, 516–22.

Petrova, N. & Wu, C. (2006). Prediction of catalytic residues using Support Vector Machine with selected protein sequence and structural properties. *BMC Bioinformatics*, **7**, 312.

Petsko, G. & Ringe, D. (2004). *Protein structure and function*. New Science Press Ltd, London.

Plapp, B. (1995). Site-directed mutagenesis: a tool for studying enzyme catalysis. *Methods Enzymol*, **249**, 91–119.

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

Price, N. & Stevens, L. (1999). *Fundamentals of enzymology: the cell and molecular biology of catalytic proteins*. Oxford University Press.

Purich, D. & Allison, R. (2002). *The Enzyme Reference: A Comprehensive Guidebook to Enzyme Nomenclature, Reactions, and Methods.*. Academic Press, San Diego.

Raaijmakers, H., Toro, I., Birkenbihl, R., Kemper, B. & Suck, D. (2001). Conformational flexibility in T4 endonuclease VII revealed by crystallography: implications for substrate binding and cleavage. *J Mol Biol*, **308**, 311–23.

Read, R. (1986). Improved Fourier coefficients for maps using phases from partial structures with errors. *Acta Cryst. A*, **42**, 140–149.

Regni, C., Naught, L., Tipton, P. & Beamer, L. (2004). Structural basis of diverse substrate recognition by the enzyme PMM/PGM from P. aeruginosa. *Structure*, **12**, 55–63.

Reid, G., Miles, C., Moysey, R., Pankhurst, K. & Chapman, S. (2000). Catalysis in fumarate reductase. *Biochim Biophys Acta*, **1459**, 310–5.

Rety, S., Sopkova, J., Renouard, M., Osterloh, D., Gerke, V., Tabaries, S., Russo-Marie, F. & Lewit-Bentley, A. (1999). The crystal structure of a complex of p11 with the annexin II N-terminal peptide. *Nat Struct Biol*, **6**, 89–95.

Rhodes, G. (2000). *Crystallography made crystal clear*. Academic Press, San Diego, 2nd edn.

Rigden, D. & Galperin, M. (2004). The DxDxDG motif for calcium binding: multiple structural contexts and implications for evolution. *J Mol Biol*, **343**, 971–84.

Rost, B. (2002). Enzyme function less conserved than anticipated. *J Mol Biol*, **318**, 595–608.

Roughton, F. (1934). The kinetics of haemoglobin vi– the competition of carbon monoxide and oxygen for haemoglobin. *Proc R Soc*, **B115**, 473–95.

Rubach, J. & Plapp, B. (2003). Amino acid residues in the nicotinamide binding site contribute to catalysis by horse liver alcohol dehydrogenase. *Biochemistry*, **42**, 2907–15.

Russell, R. (1998). Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. *J Mol Biol*, **279**, 1211–27.

Sander, C. & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.

Scharff, E., Koepke, J., Fritzsch, G., Lucke, C. & Ruterjans, H. (2001). Crystal structure of diisopropylfluorophosphatase from Loligo vulgaris. *Structure*, **9**, 493–502.

Schein, C., Zhou, B., Oezguen, N., Mathura, V. & Braun, W. (2005). Molego-based definition of the architecture and specificity of metal-binding sites. *Proteins*, **58**, 200–10.

Schmitt, S., Kuhn, D. & Klebe, G. (2002). A new method to detect related function among proteins independent of sequence and fold homology. *J Mol Biol*, **323**, 387–406.

Schneider, T., Gerhardt, E., Lee, M., Liang, P., Anderson, K. & Schlichting, I. (1998). Loop closure and intersubunit communication in tryptophan synthase. *Biochemistry*, **37**, 5394–406.

Schomburg, I., Chang, A., Hofmann, O., Ebeling, C., Ehrentreich, F. & Schomburg, D. (2002). BRENDA: a resource for enzyme data and metabolic information. *Trends Biochem Sci*, **27**, 54–6.

Schramm, V. (1998). Enzymatic transition states and transition state analog design. *Annu Rev Biochem*, **67**, 693–720.

Schumacher, M., Rivard, A., Bachinger, H. & Adelman, J. (2001). Structure of the gating domain of a Ca2+-activated K+ channel complexed with Ca2+/calmodulin. *Nature*, **410**, 1120–4.

Schutz, C. & Warshel, A. (2004). The low barrier hydrogen bond (LBHB) proposal revisited: the case of the Asp... His pair in serine proteases. *Proteins*, **55**, 711–23.

Schwede, T., Retey, J. & Schulz, G. (1999). Crystal structure of histidine ammonia-lyase revealing a novel polypeptide modification as the catalytic electrophile. *Biochemistry*, **38**, 5355–61.

Seibert, C. & Raushel, F. (2005). Structural and catalytic diversity within the amidohydrolase superfamily. *Biochemistry*, **44**, 6383–91.

Shi, W., Li, C., Tyler, P., Furneaux, R., Grubmeyer, C., Schramm, V. & Almo, S. (1999). The 2.0 A structure of human hypoxanthine-guanine phosphoribosyltransferase in complex with a transition-state analog inhibitor. *Nat Struct Biol*, **6**, 588–93.

Shindyalov, I. & Bourne, P. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng*, **11**, 739–47.

Shortle, D., DiMaio, D. & Nathans, D. (1981). Directed mutagenesis. *Annu Rev Genet*, **15**, 265–94.

Singh, R. & Saha, M. (2003). Identifying structural motifs in proteins. *Pac Symp Biocomput*, 228–39.

Sjolander, K. (2004). Phylogenomic inference of protein molecular function: advances and challenges. *Bioinformatics*, **20**, 170–9.

Sokolovsky, M., Riordan, J. & Vallee, B. (1966). Tetranitromethane. A reagent for the nitration of tyrosyl residues in proteins. *Biochemistry*, **5**, 3582–9.

Spratt, B. (1975). Distinct penicillin binding proteins involved in the division, elongation, and shape of Escherichia coli K12. *Proc Natl Acad Sci U S A*, **72**, 2999–3003.

Spriggs, R., Artymiuk, P. & Willett, P. (2003). Searching for patterns of amino acids in 3D protein structures. *J Chem Inf Comput Sci*, **43**, 412–21.

Stark, A. & Russell, R. (2003). Annotation in three dimensions. PINTS: Patterns in Non-homologous Tertiary Structures. *Nucleic Acids Res*, **31**, 3341–4.

Stark, A., Sunyaev, S. & Russell, R. (2003). A model for statistical significance of local similarities in structure. *J Mol Biol*, **326**, 1307–16.

Starks, C., Back, K., Chappell, J. & Noel, J. (1997). Structural basis for cyclic terpene biosynthesis by tobacco 5-epi-aristolochene synthase. *Science*, **277**, 1815–20.

Storm, C. & Sonnhammer, E. (2002). Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics*, **18**, 92–9.

Storm, D. & Koshland, D. (1970). A Source for the Special Catalytic Power of Enzymes: Orbital Steering. *Proc Natl Acad Sci U S A*, **66**, 445–452.

Tao, H. & Cornish, V. (2002). Milestones in directed enzyme evolution. *Curr Opin Chem Biol*, **6**, 858–64.

Taylor, R. & Kennard, O. (1984). Hydrogen-bond geometry in organic crystals. *Acc Chem Res*, **17**, 320–326.

Thompson, J., Higgins, D. & Gibson, T. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, **22**, 4673–80.

Thorn, J., Barton, J., Dixon, N., Ollis, D. & Edwards, K. (1995). Crystal structure of Escherichia coli QOR quinone oxidoreductase complexed with NADPH. *J Mol Biol*, **249**, 785–99.

Thornton, J., Todd, A., Milburn, D., Borkakoti, N. & Orengo, C. (2000). From structure to function: approaches and limitations. *Nat Struct Biol*, **7 Suppl**, 991–4.

Tian, W. & Skolnick, J. (2003). How well is enzyme function conserved as a function of pairwise sequence identity? *J Mol Biol*, **333**, 863–82.

Todd, A., Orengo, C. & Thornton, J. (2001). Evolution of function in protein superfamilies, from a structural perspective. *J Mol Biol*, **307**, 1113–43.

Todd, A., Orengo, C. & Thornton, J. (2002). Plasticity of enzyme active sites. *Trends Biochem Sci*, **27**, 419–26.

Todd, A., Marsden, R., Thornton, J. & Orengo, C. (2005). Progress of structural genomics initiatives: an analysis of solved target structures. *J Mol Biol*, **348**, 1235–60.

Trentham, D. (1971). Rate-determining processes and the number of simultaneously active sties of D-glyceraldehyde 3-phosphate dehydrogenase. *Biochem J*, **122**, 71–7.

UniProt Consortium (2007). The Universal Protein Resource (UniProt). *Nucleic Acids Res*, **35**, D193–7.

Valegard, K., Terwisscha van Scheltinga, A., Dubus, A., Ranghino, G., Oster, L., Hajdu, J. & Andersson, I. (2004). The structural basis of cephalosporin formation in a mononuclear ferrous enzyme. *Nat Struct Mol Biol*, **11**, 95–101.

Via, A., Ferre, F., Brannetti, B. & Helmer-Citterich, M. (2000). Protein surface similarities: a survey of methods to describe and compare protein surfaces. *Cell Mol Life Sci*, **57**, 1970–7.

Wallace, A., Laskowski, R. & Thornton, J. (1996). 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**, 1001–13.

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

Wang, G. & Dunbrack, R., Jr (2003). PISCES: a protein sequence culling server. *Bioinformatics*, **19**, 1589–91.

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

Wangikar, P., Tendulkar, A., Ramya, S., Mali, D. & Sarawagi, S. (2003). Functional sites in protein families uncovered via an objective and automated graph theoretic approach. *J Mol Biol*, **326**, 955–78.

Warshel, A. (1978). Energetics of enzyme catalysis. *Proc Natl Acad Sci U S A*, **75**, 5250–4.

Warshel, A. & Papazyan, A. (1996). Energy considerations show that low-barrier hydrogen bonds do not offer a catalytic advantage over ordinary hydrogen bonds. *Proc Natl Acad Sci U S A*, **93**, 13665–70.

Webb, E. (1992). *Enzyme Nomenclature*, chap. Recommendations of the nomenclature committee of the International Union of Biochemistry and Molecular Biology. Academic Press, San Diego.

Wei, L. & Altman, R. (2003). Recognizing complex, asymmetric functional sites in protein structures using a Bayesian scoring function. *J Bioinform Comput Biol*, **1**, 119–38.

Whisstock, J. & Lesk, A. (2003). Prediction of protein function from protein sequence and structure. *Q Rev Biophys*, **36**, 307–40.

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

Wistow, G. & Piatigorsky, J. (1987). Recruitment of enzymes as lens structural proteins. *Science*, **236**, 1554–6.

Wofsy, L., Metzger, H. & Singer, S. (1962). Affinity labeling-a general method for labeling the active sites of antibody and enzyme molecules. *Biochemistry*, **1**, 1031–9.

Wolf, F. & Cittadini, A. (2003). Chemistry and biochemistry of magnesium. *Mol Aspects Med*, **24**, 3–9.

Wolfe, S., Grant, R. & Pabo, C. (2003). Structure of a designed dimeric zinc finger protein bound to DNA. *Biochemistry*, **42**, 13401–9.

Yang, G., Liu, R., Taylor, K., Xiang, H., Price, J. & Dunaway-Mariano, D. (1996). Identification of active site residues essential to 4-chlorobenzoyl-coenzyme A dehalogenase catalysis by chemical modification and site directed mutagenesis. *Biochemistry*, **35**, 10879–85.

Ye, Y. & Godzik, A. (2004). FATCAT: a web server for flexible structure comparison and structure similarity searching. *Nucleic Acids Res*, **32**, W582–5.

Ye. Y., Xie, T. & Ding D. (2000). Protein functional-group 3D motif and its applications. *Chinese Sci Bull*, **45**, 2044–2052.

Zdobnov, E. & Apweiler, R. (2001). InterProScan–an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**, 847–8.

Zehetner, G. (2003). OntoBlast function: From sequence similarities directly to potential functional annotations by ontology terms. *Nucleic Acids Res*, **31**, 3799–803.

Zhang, E., Hatada, M., Brewer, J. & Lebioda, L. (1994). Catalytic metal ion binding in enolase: the crystal structure of an enolase-Mn2+-phosphonoacetohydroxamate complex at 2.4-A resolution. *Biochemistry*, **33**, 6295–300.

Zhang, J. (2003). Evolution by gene duplication: an update. *Trends Ecol Evol*, **18**, 292–8.

Zhu, J. & Weng, Z. (2005). FAST: a novel protein structure alignment algorithm. *Proteins*, **58**, 618–27.

Zmasek, C. & Eddy, S. (2002). RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics*, **3**, 14.

Zollner, H. (1999). *Handbook of Enzyme Inhibitors*. Wiley-VCH, Weinheim, 3rd edn.