Computational techniques for the study of enzyme active sites

Andrew Campbell Wallace

A thesis submitted to the University of London

in the Faculty of Science

for the degree of Doctor of Philosophy

November 1996

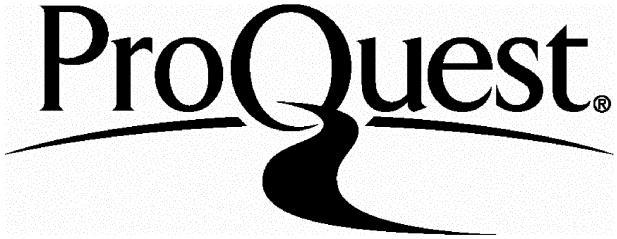
Biomolecular Structure & Modelling Unit,

Department of Biochemistry & Molecular Biology,

University College,

Gower Street,

London WC1E 6BT



Abstract

The aim of this thesis, which is based on the coordinate data from known X–ray crystal and nuclear magnetic resonance (NMR) structures, is to further our understanding of the ways in which enzymes catalyse their reactions.

Many enzyme structures are solved with inhibitors or substrate analogs bound to their active sites. To enable evaluation of the protein–ligand interactions that occur in these complexes, a computational tool called LIGPLOT has been developed which allows swift pictorial evaluation of the ligand and its interactions with the enzyme. These LIGPLOT diagrams, along with other relevent information has been compiled into an enzyme database that is available over the World Wide Web.

The rest of the thesis is devoted to studying structural organisation of the amino acid residues that are directly involved in chemical catalysis. A detailed analysis of the geometry of the Ser–His–Asp catalytic triad found in the serine proteinases and lipase X–ray and NMR structures showed that it is possible to define a 3D consensus template that identifies all catalytic Ser–His–Asp triads with the exclusion of all other interactions. To create 3D consensus templates describing other enzyme active sites we needed a generalised search method. To do this, a computer program called 'TESS' has been developed which is based on the geometric hashing paradigm. Using this program, a database of enzyme active site templates has been created which enables swift evaluation of the function of a new protein structure as it is solved and aid protein design and engineering experiments.

Acknowledgments

I would especially like thank my supervisor Janet M. Thornton for her help and encouragement over the last three years. Also to Roman Laskowski for repeatedly rescuing me from my confusion (especially over John Mitchell's jokes) and the myriad of mistakes I've managed on the computer front. Thanks also to my industrial supervisor, Neera Borkakoti.

Also, all those people who have helped in various ways along the way: Alex PRV Michie, Tezza Attwood, Malcolm McArthur, Andrew Martin, David Jones, Frances Richardson, Christine Orengo, Martin Jones, Tim Slidel and apologies to the ones I've forgotten. The pub philosophers: Mark McAlister, Lorraine Finney, Bernard O'Hara, Tony Guzzler Headley. Also David Moynagh for helping with the crossword puzzles.

Of course, a special thanks to Francoise.

To my parents

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Proteins have an integral role in all living organisms; they have diverse structure and function such as division, motility, immune response and enzymatic activity. Fibrous proteins generally have a structural role, for example hair, bones and nails. Globular proteins have a compact folded structure and perform all other roles and, with the exception of membrane proteins, tend to be soluble in water and are therefore easiest to isolate and study. By far the most structural information is on globular proteins. Proteins are large molecules; advances in X–ray and nuclear magnetic resonance techniques (NMR) has enabled the high resolution structures of around 400 unique protein folds to be identified. These structures are deposited in the Brookhaven Protein Databank (PDB, Bernstein *et al.*, 1977).

In this thesis we will concentrate on the role proteins have as enzymes; these are biological catalysts that determine the rate and type of chemical reactions that occur in every living cell. Specifically, computational techniques have been developed that enable us to analyse enzyme active sites from a structural perspective.

1.1 A brief history of enzyme research

In the late 19^{th} century Emil Fischer used maltase and emulsin to establish the stereochemistry of anomeric derivatives of sugars; this work enabled him to suggest the lock and key hypothesis to described the enzyme–substrate complex. In 1913 Leonor Michaelis and Maud Menton proposed that an enzyme and substrate first combine to form an enzyme–substrate complex which then breaks down to form product and free enzyme. These investigators assumed that substrate and enzyme are in continuous equilibrium; Briggs and Haldane reformulated the reaction process as steady state in 1925. In the same year Lowry suggested that acid/base catalysis could strongly accelerate the reaction by studying mutorotation of glucose (Lowry & Faulkner, 1925).

In the 1930's J.H.Northrop and M. Kunitz crystallised pepsin, trypsin and chymotrypsin. This provided the material to finally prove that enzymes were proteins and allowed the development of the techniques of modern protein chemistry: the sequencing of the protein insulin by Fred Sanger (Sanger & Tuppy, 1951*a*,*b*); the determination of 3D structure of the protein hemoglobin and myoglobin by John Kendrew and Max Perutz (Perutz *et al.*, 1960 and Kendrew *et al.*, 1960 respectively); and the use of rapid–reaction kinetics, which had been initiated by F.J.W. Roughton in 1923.

Linus Pauling suggested in the late 1940's that enzymes catalyse reactions by strongly binding the transition state of the substrate. Koshland's work on hexokinase led to the 'induced fit' hypothesis to explain enzyme specificity (Koshland, 1954). By this time the role of coenzymes was also understood, for example, pyridoxal phosphate (Braunstein, 1960), nicotinamide–adenine dinucleotides (NAD^+ and $NADP^+$, Westheimer *et al.*, 1951) and thiamin pyrophosphate (Breslow, 1958).

In the mid 1960's the X–ray structure of the enzyme lysozyme from hen egg white was determined (Blake *et al.*, 1965, 1967 *a*,*b*). The structure was refined

over the following years (*e.g* Imoto *et al.*, 1972). The development of more powerful computers and software has enabled protein structure determination and refinement to become both faster and easier.

Finally, the discovery that RNA molecules have catalytic power (Zaug & Cech, 1986) showed that proteins are not the only biological molecules with catalytic activity.

1.2 The role of structural biology in understanding enzyme action

Solving the X–ray or NMR structure of enzymes has proved to be the single most important factor in our understanding of enzyme mechanism. Generally speaking, structural information is used in concert with kinetic data to hypothesise as to the enzyme's mechanism and mode of action. If the residues involved in catalysis and ligand binding can be elucidated, it gives the protein engineer a good starting point for mutagenesis experiments and enables inhibitors to be designed that provide a starting point for structure–based rational drug design.

1.2.1 X–ray diffraction methods

Structure determination by X–ray diffraction (*e.g.* Blundell & Johnson, 1976) requires the protein to be purified in sufficient quantities so that it can be crystallised. The process is far from straightforward; the main problems can be expressing and isolating enough quantity of protein to be crystallised. Crystallising the protein itself is not easy and it is not unusual for a protein's structure determination to take several years.

However, if the target protein is sufficiently similar to the structure of a protein in the PDB, the technique of molecular replacement can be used to solve its

structure in only a few days.

Structure determination

When the beam of X–rays strikes the regular lattice of a protein crystal the structure may be calculated from the resultant diffraction pattern by Fourier transformation. This requires knowledge of the intensity, direction and phases of the diffracted rays. Determination of the phases is the most difficult problem and this held up protein crystallography until, in 1954, Perutz and his coworkers applied the method of multiple isomorphous replacement (MIR). Here, heavy metals are bound at specific sites in the protein without disturbing its structure. The metal scatters X–rays more than the protein and information about phases can be deduced from changes in intensity of the diffraction map.

Once the phases and amplitude of every diffraction ray has been calculated, the electron density of the protein may be calculated. Nowadays computer graphics and software have made this procedure much easier.

Accuracy and resolution

The accuracy of the protein model depends on several factors (*e.g.* Branden & Jones, 1990). Firstly, determination of the phases involves calculating small differences between large numbers and this depends on the accuracy of the diffraction measurements and usefulness of the heavy–atom derivatives.

Resolution of the diffraction data depends on how well ordered the crystals are and this directly influences the image that can be produced. Native crystals are usually better quality than the derivatives. At 4–6Å resolution, the electron density map shows little more than the overall topology of the molecule. At 3.5Å it is possible to follow the course of the polypeptide backbone and at 3.0Å the amino acid sidechains can be deciphered. At 2.5Å the atoms can be fitted with

an accuracy of ±0.4Å. At 1.9Å resolution, the atoms can be located to ±0.2Å resolution. In a typical MIR map the phases are determined to as low as 2Å resolution.

From the MIR electron density map, a model of the protein is built. The crystallographer has to decide how the polypeptide chain weaves its way through the map. This is yet to be automated but computer graphics facilitate this process; skeletonised representations of the protein and contour nets of the electron density are the most common. In addition, the crystallographer uses other information to fit the structure such as position of the heavy metal, active site residues, and the distinct density formed by α–helices and β–sheets.

The final model from the MIR will contain many errors which can usually be removed by the refinement process. The model is changed so that the structure amplitudes calculated from the model fit the observed amplitudes. The goodness of this fit is expressed in terms of the R–factor which is a measure of the difference between the observed and calculated data. Obviously, the better the MIR map, the better the initial model and the more reliable is the structure; unfortunately, not all crystals diffract well. There are various computer programs available for refinement such as XPLOR (Brunger *et al.*, 1987). XPLOR stands for exploration of conformational space of macromolecules confined to regions by experimental data and error estimates. The program is based on an energy function approach: arbitrary combinations of empirical and effective energy terms describing experimental data may be used. The combined energy function can be minimized by a variety of gradient descent, simulated annealing, and conformational search procedures. XPLOR evolved from the CHARMM program (Brooks *et al.*, 1983) and was the first program to combine X-ray crystallographic diffraction data and molecular dynamics for refinement (Brunger *et al*, 1987).

When the structure is published, the quoted R–factor needs to be treated

with caution for several reasons. Firstly, it is affected by the removal of weak reflections and is insensitive to errors in mainchain connectivity. The number of water molecules added during the refinement should also be realistic as this can artificially reduce the R–factor. Brunger (1992) advocated the use of the free R–factor which is an unbiased indicator of the accuracy of the protein models. This calculates the difference between the observed and computed diffraction data for a 'test' set of data that is omitted from the modelling and refinement procedure.

1.2.2 Case study: the serine proteinases and acetylcholinesterase

In this section the serine proteinases and acetylcholinesterase are used as examples to illustrate the relevance of structural biology to our understanding of enzymes.

Research into serine proteinases' mechanism of action began over 60 years ago, when, in 1932 two German investigators, Lange & Krueger, synthesised diethyl fluorophosphate or nerve gas. During World War II, research on this compound was carrried out for military purposes and Adrian and his coworkers first noted the similarity between physiological action of the flourophosphonates and that of reversible inhibitors of acetylcholine esterases (Adrian *et al.*, 1947). This led to a number of investigations of the action of nerve gases on esterases. In 1946 Mazur & Bodansky found that diisopropyl fluorophosphate (DFP) irreversibly inhibits acetylcholinesterase and in 1949 Jansen and coworkers demonstrated the 1:1 stoichiometric reaction of DFP with chymotrypsin Ser 195. The reactivity of Ser 195 is highlighted by the fact that the other 27 Ser residues in chymotrypsin are untouched by DFP.

The significance of the catalytic serine in the serine esterases' mechanism was identified when hydrolysis of the diisopropylphosphoryl derivative of acetylcholinesterase yielded a serine–phosphate covalent bond (Schaffer *et al.*, 1953).

In 1950, Wilson, Bergman & Nachmansohn published a two step mechanism for the action of acetylcholinesterase (Wilson *et al.*, 1950); stop–flow kinetics (Gutfreund & Sturtevant, 1956) supported the hypothesis. In Brian Hartley's laboratory (Hartley & Kilby, 1954) it was found that chymotrypsin catalyses the hydrolysis of *p*–nitrophenylacetate; during the reaction there was initially a rapid liberation of *p*–nitrophenol followed by a slow hydrolysis. This indicated that there were two phases to the catalytic reaction; the 'burst phase' in which the *p*–nitrophenylacetate reacts to form *p*–nitrophenylate and a covalent acyl–enzyme intermediate and, secondly, the 'steady state' phase wherby the intermediate is slowly hydrolysed releasing acetate.

The identification of a His residue as an essential basic group was made by Whitaker & Jandorf (1956); they treated chymotrypsin with 2,4–dinitrophenylbenzene and found that the enzyme was inactivated with destruction of the His residue. This was preceded by Schoellman & Shaw (1953) who demonstrated that N–tosylphenylalanyl chloromethyl ketone (TPCK) reacts irreversibly with the enzyme. The molecule is a strong electrophile and occupies the active site of the enzyme, ireversibly inhibiting it; this led to the concept of 'affinity labelling' (Schoellman & Shaw, 1963) and many other examples of such compounds have subsequently been identified.

David Blow solved the high resolution X–ray structure of chymotrypsin in the late 1960's (Matthews *et al.*, 1967; Blow *et al.*, 1969; Blow, 1976). This work enabled, along with many spectroscopic and kinetic experiments, three structural features to be identified that facilitate the proteolysis by serine–proteinases; these are summarised in Figure 1.1.

Firstly, the enzyme has a catalytic triad consisting of Ser, His and Asp residues. The His and Asp act in concert to form an acid/base catalyst which accepts a proton from the nucleophilic Ser. The Ser attacks the carbonyl group

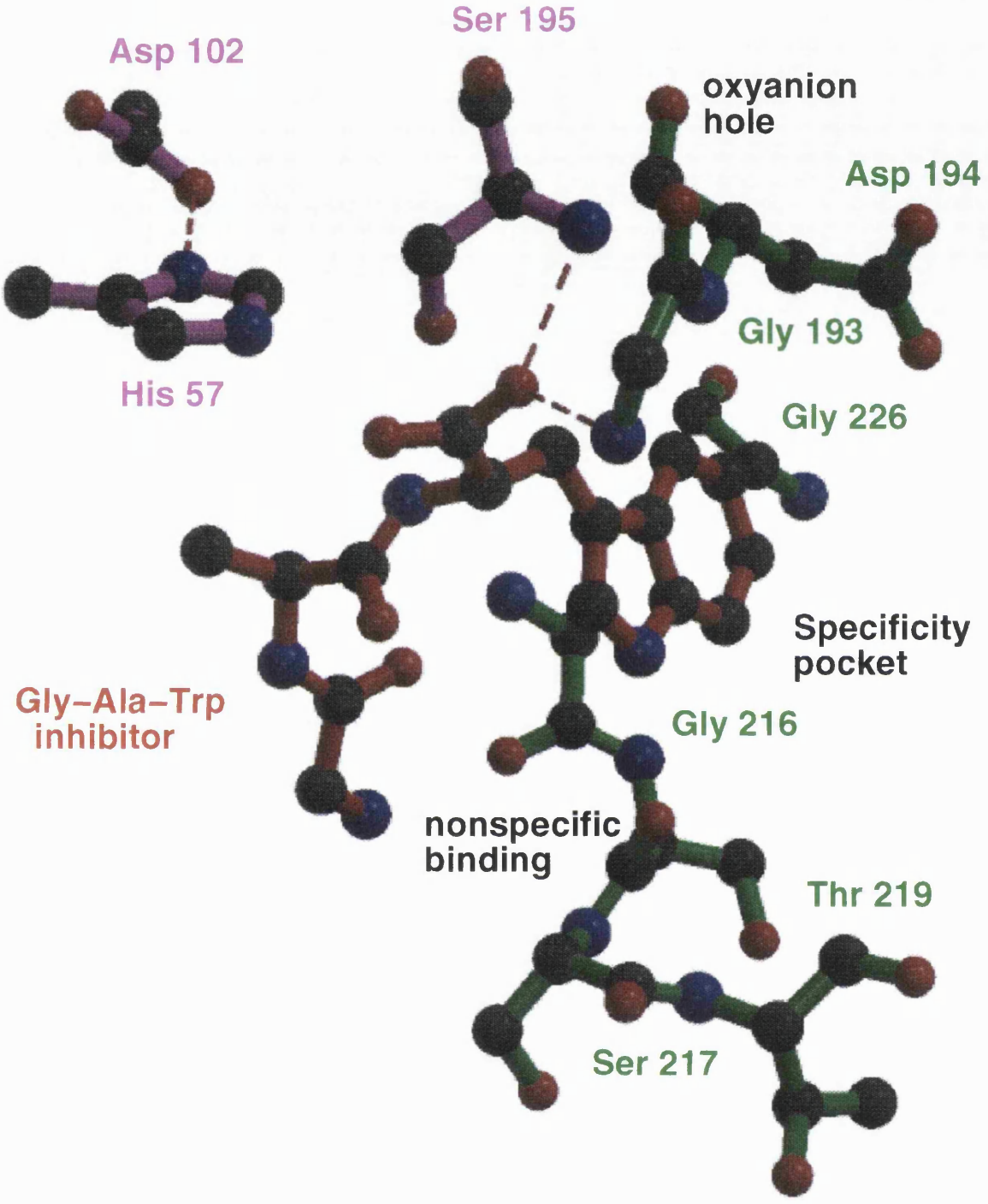


Figure 1.1: A diagrammatic representation of the active site of the serine proteinase chymotrypsin taken from the X–ray structure of chymotrypsin (Harel *et al.*, 1991)

of the peptide bond of the substrate forming a acyl–enzyme intermediate which is subsequently hydrolysed by water to form product.

Secondly, there is tight binding by residues of the tetrahedral transition state intermediate in the so called 'oxyanion hole'. The tetrahedral intermediate is negatively charged and this is stabilised by hydrogen bonding from residues in the oxyanion hole. For example, Figure 1.1 shows Gly 193 to be hydrogen bonded to the inhibitor in a similar manner to the transition state.

Thirdly, serine proteinases have a specificity pocket which binds the amino acid next to the scissile bond. In Figure 1.1 Gly 226 from the enzyme occupies this hole and so a large residue fits into the pocket; in this case it is a Trp from the inhibitor.

1.2.3 Classification of enzymes

The E.C. number (Bielka *et al.*, 1992) classifies enzymes according to their function - both in terms of the reaction they catalyse and the substrate on which they operate. The E.C. number consists of four component numbers. The first defines the six main classes of enzymes: the oxidoreductases, transferases, hydrolases, lyases, isomerases and ligases. The other three components vary among the six groups but in general describe the type of substrate and any cofactors or prosthetic groups that are involved in the enzyme reaction. Each unique enzyme has a unique E.C. number. For example, the serine proteinases are identified by E.C. numbers 3.4.21.n, where n in this case groups the enzymes according to specificity for substrate; for example, in chymotrypsin $n = 1$ and in thrombin $n = 5$.

1.2.4 Enzyme structures in the PDB

A recent paper by Hooft *et al*, 1996 estimated that there are over a million errors in the PDB. These are mainly typographical and format errors in the files

deposited in the PDB, although inconsistences in bond lengths and angles were also identified. Due to these general errors in the PDB files, one cannot rely on a given enzyme structure in this database having its E.C. number quoted. Therefore another system has been developed. This relies on aligning the amino acid sequence of each protein structure in the PDB with the SWISS–PROT (Bairoch A. & Boeckmann B., 1994; March 1995 release) sequence database using the automatic sequence alignment program BLAST (Altschul *et al.*, 1990). This database has an accurate record of E.C. numbers so the enzymes structures in the PDB can be identified on this basis. In addition, as an extra checking system, the name of the enzyme can be located in the enzyme databank (Bairoch, 1996).

A summary of the enzyme structures in the January 1995 release of the PDB is listed in Table 1.1 There are 1581 enzyme structures, many of which are different complexes of the same proteins, giving us 214 unique enzymes by E.C. number. Figure 1.2 is a histogram of the number of structures against E.C. number for all the enzymes in our datasets and shows both the total and number of unique enzyme structures. The hydrolase (E.C.3) enzymes dominate this dataset with 935 structures or 96 unique enzymes by E.C. number and this group is itself dominated by proteinase structures; these are one of the best understood group of enzymes. There are 24 unique serine proteinases (E.C.3.4.21.x), 9 unique aspartic proteinases (E.C.3.4.23.x), 3 unique cysteine proteinases (E.C.3.4.22.x) and 5 unique metallo–proteinases (E.C.3.4.24.x).

Figure 1.3 shows the number of structures for each of the 214 unique enzymes in the PDB. Clearly, the majority of enzymes have between 1 and 25 representative structures in the PDB. The exception to this is lysozyme which has 248 structures; they can be divided into 2 groups, the mammalian *e.g.* Imoto *et al.*, 1972 and bacteriophage T4 lysozymes *e.g.* Weaver & Matthews, 1987.

Oxidoreductases Total 226 Unique 47

E.C. number	Name	Number of structures in PDB
E.C.1.1.1.1	ALCOHOL DEHYDROGENASE	13
E.C.1.1.1.14	L-IDITOL 2-DEHYDROGENASE	1
E.C.1.1.1.21	ALDEHYDE REDUCTASE	7
E.C.1.1.1.27	L-LACTATE DEHYDROGENASE	13
E.C.1.1.1.29	GLYCERATE DEHYDROGENASE	1
E.C.1.1.1.37	MALATE DEHYDROGENASE	6
E.C.1.1.1.42	ISOCITRATE DEHYDROGENASE (NADP+)	9
E.C.1.1.1.44	PHOSPHOGLUCONATE DEHYDROGENASE (DECARBOXYLATING)	1
E.C.1.1.1.50	3-ALPHA-HYDROXYSTEROID DEHYDROGENASE (B-SPECIFIC)	1
E.C.1.1.1.53	3-ALPHA(OR 20-BETA)-HYDROXYSTEROID DEHYDROGENASE	1
E.C.1.1.1.85	3-ISOPROPYLMALATE DEHYDROGENASE	2
E.C.1.1.1.86	KETOL-ACID REDUCTOISOMERASE	4
E.C.1.1.2.3	L-LACTATE DEHYDROGENASE (CYTOCHROME)	2
E.C.1.1.3.4	GLUCOSE OXIDASE	1
E.C.1.1.3.6	CHOLESTEROL OXIDASE	2
E.C.1.1.3.9	GALACTOSE OXIDASE	3
E.C.1.1.3.15	(S)-2-HYDROXY-ACID OXIDASE	1
E.C.1.2.1.12	GLYCERALDEHYDE 3-PHOSPHATE DEHYDROGENASE (PHOSPHORYLATING)	5
E.C.1.2.3.3	PYRUVATE OXIDASE	2
E.C.1.4.99.3	AMINE DEHYDROGENASE	6
E.C.1.5.1.3	DIHYDROFOLATE REDUCTASE	22
E.C.1.5.99.7	TRIMETHYLAMINE DEHYDROGENASE	1
E.C.1.6.4.2	GLUTATHIONE REDUCTASE (NADPH)	12
E.C.1.6.4.5	THIOREDOXIN REDUCTASE (NADPH)	3
E.C.1.6.4.8	TRYPANOTHIONE REDUCTASE	4
E.C.1.6.6.1	NITRATE REDUCTASE (NADH)	1
E.C.1.6.99.1	NADPH DEHYDROGENASE	2
E.C.1.6.99.7	DIHYDROPTERIDINE REDUCTASE	3
E.C.1.7.99.3	NITRITE REDUCTASE	2
E.C.1.8.1.4	DIHYDROLIPOAMIDE DEHYDROGENASE	3
E.C.1.10.3.3	L-ASCORBATE OXIDASE	4
E.C.1.11.1.0	LIGNIN PEROXIDASE	1
E.C.1.11.1.1	NADH PEROXIDASE	2
E.C.1.11.1.5	CYTOCHROME-C PEROXIDASE	23
E.C.1.11.1.6	CATALASE	2
E.C.1.11.1.7	PEROXIDASE	2
E.C.1.11.1.8	IODIDE PEROXIDASE	3
E.C.1.11.1.9	GLUTATHIONE PEROXIDASE	1
E.C.1.13.11.3	PROTOCATECHUATE 3,4-DIOXYGENASE	1
E.C.1.14.13.2	4-HYDROXYBENZOATE 3-MONOOXYGENASE	11
E.C.1.14.14.1	UNSPECIFIC MONOOXYGENASE	2
E.C.1.14.15.1	CAMPHOR 5-MONOOXYGENASE	18
E.C.1.15.1.1	SUPEROXIDE DISMUTASE	18
E.C.1.18.1.2	FERREDOXIN–NADP(+) REDUCTASE	2
E.C.1.18.6.1	NITROGENASE	1

Transferases Total 183 Unique 34

E.C.2.1.1.45	THYMIDYLATE SYNTHASE	12
E.C.2.1.1.73	SITE-SPECIFIC DNA-METHYLTRANSFERASE (CYTOSINE-SPECIFIC)	1
E.C.2.1.2.2	PHOSPHORIBOSYLGLYCINAMIDE FORMYLTRANSFERASE	3
E.C.2.1.3.2	ASPARTATE CARBAMOYLTRANSFERASE	21
E.C.2.2.1.1	TRANSKETOLASE	4
E.C.2.3.1.12	DIHYDROLIPOAMIDE S-ACETYLTRANSFERASE	10
E.C.2.3.1.16	ACETYL-COA C-ACYLTRANSFERASE	1
E.C.2.3.1.28	CHLORAMPHENICOL O-ACETYLTRANSFERASE	5
E.C.2.3.1.61	DIHYDROLIPOAMIDE S-SUCCINYLTRANSFERASE	2
E.C.2.4.1.1	PHOSPHORYLASE	13
E.C.2.4.1.22	LACTOSE SYNTHASE	1
E.C.2.4.1.27	DNA BETA-GLUCOSYLTRANSFERASE	2
E.C.2.4.2.1	PURINE-NUCLEOSIDE PHOSPHORYLASE	2
E.C.2.4.2.4	THYMIDINE PHOSPHORYLASE	1
E.C.2.4.2.10	OROTATE PHOSPHORIBOSYLTRANSFERASE	1
E.C.2.4.2.14	AMIDOPHOSPHORIBOSYLTRANSFERASE	1
E.C.2.4.2.36	NAD(+)-DIPHTHAMIDE ADP-RIBOSYLTRANSFERASE	3
E.C.2.5.1.18	GLUTATHIONE TRANSFERASE	17
E.C.2.5.1.19	3-PHOSPHOSHIKIMATE 1-CARBOXYVINYLTRANSFERASE	1
E.C.2.6.1.1	ASPARTATE AMINOTRANSFERASE	25
E.C.2.7.1.1	HEXOKINASE	3
E.C.2.7.1.11	6-PHOSPHOFRUCTOKINASE	5
E.C.2.7.1.69	PROTEIN-N(PI)-PHOSPHOHISTIDINE-SUGAR PHOSPHOTRANSFERASE	7
E.C.2.7.1.112	PROTEIN-TYROSINE KINASE	13
E.C.2.7.1.117	[MYOSIN LIGHT-CHAIN] KINASE	1
E.C.2.7.2.3	PHOSPHOGLYCERATE KINASE	2
E.C.2.7.4.3	ADENYLATE KINASE	3
E.C.2.7.4.6	NUCLEOSIDE-DIPHOSPHATE KINASE	7
E.C.2.7.4.8	GUANYLATE KINASE	1
E.C.2.7.4.10	NUCLEOSIDE-TRIPHOSPHATE–ADENYLATE KINASE	1
E.C.2.7.7.0	KANAMYCIN NUCLEOTIDYLTRANSFERASE	1
E.C.2.7.7.7	DNA-DIRECTED DNA POLYMERASE	10
E.C.2.7.7.48	RNA-DIRECTED RNA POLYMERASE	2
E.C.2.8.1.1	THIOSULFATE SULFURTRANSFERASE	1

Hydrolases Total 935 Unique 96

E.C. number	Name	Number of structures in PDB
E.C.3.1.1.0	CUTINASE	2
E.C.3.1.1.3	TRIACYLGLYCEROL LIPASE	9
E.C.3.1.1.4	PHOSPHOLIPASE A2	11
E.C.3.1.1.7	ACETYLCHOLINESTERASE	4
E.C.3.1.3.1	ALKALINE PHOSPHATASE	1
E.C.3.1.3.2	ACID PHOSPHATASE	2
E.C.3.1.3.11	FRUCTOSE-BISPHOSPHATASE	13
E.C.3.1.3.25	MYO-INOSITOL-1(OR 4)-MONOPHOSPHATASE	1
E.C.3.1.3.48	PROTEIN-TYROSINE-PHOSPHATASE	7
E.C.3.1.4.11	1-PHOSPHATIDYLINOSITOL-4,5-BISPHOSPHATE PHOSPHODIESTERASE	2
E.C.3.1.21.1	DEOXYRIBONUCLEASE I	3
E.C.3.1.21.4	TYPE II SITE-SPECIFIC DEOXYRIBONUCLEASE	3
E.C.3.1.25.1	DEOXYRIBONUCLEASE (PYRIMIDINE DIMER)	4
E.C.3.1.26.4	RIBONUCLEASE H	17
E.C.3.1.27.0	BARNASE	14
E.C.3.1.27.3	RIBONUCLEASE T1	32
E.C.3.1.27.5	PANCREATIC RIBONUCLEASE	47
E.C.3.1.31.1	MICROCOCCAL NUCLEASE	19
E.C.3.2.1.1	ALPHA-AMYLASE	5
E.C.3.2.1.2	BETA-AMYLASE	5
E.C.3.2.1.3	GLUCAN 1,4-ALPHA-GLUCOSIDASE	4
E.C.3.2.1.4	CELLULASE	2
E.C.3.2.1.8	ENDO-1,4-BETA-XYLANASE	4
E.C.3.2.1.10	OLIGO-1,6-GLUCOSIDASE	7
E.C.3.2.1.14	CHITINASE	1
E.C.3.2.1.17	LYSOZYME	248
E.C.3.2.1.18	EXO-ALPHA-SIALIDASE	18
E.C.3.2.1.20	ALPHA-GLUCOSIDASE	1
E.C.3.2.1.26	BETA-FRUCTOFURANOSIDASE	2
E.C.3.2.1.39	GLUCAN ENDO-1,3-BETA-D-GLUCOSIDASE	1
E.C.3.2.1.45	GLUCOSYLCERAMIDASE	1
E.C.3.2.1.73	LICHENINASE	5
E.C.3.2.1.91	CELLULOSE 1,4-BETA-CELLOBIOSIDASE	4
E.C.3.2.2.22	RRNA N-GLYCOSIDASE	11
E.C.3.4.11.1	LEUCYL AMINOPEPTIDASE	4
E.C.3.4.11.10	BACTERIAL LEUCYL AMINOPEPTIDASE	1
E.C.3.4.11.18	METHIONYL AMINOPEPTIDASE	1
E.C.3.4.14.1	DIPEPTIDYL-PEPTIDASE I	1
E.C.3.4.16.4	SERINE-TYPE D-ALA-D-ALA CARBOXYPEPTIDASE	1
E.C.3.4.16.5	CARBOXYPEPTIDASE C	1
E.C.3.4.17.1	CARBOXYPEPTIDASE A	11
E.C.3.4.17.2	CARBOXYPEPTIDASE B	2
E.C.3.4.21.0	RAT PROTEASE	2
E.C.3.4.21.1	CHYMOTRYPSIN	27
E.C.3.4.21.4	TRYPSIN	43
E.C.3.4.21.5	THROMBIN	42
E.C.3.4.21.6	COAGULATION FACTOR XA	2
E.C.3.4.21.7	PLASMIN	4
E.C.3.4.21.12	ALPHA-LYTIC PROTEASE	22
E.C.3.4.21.22	COAGULATION FACTOR IXA	1
E.C.3.4.21.35	TISSUE KALLIKREIN	3
E.C.3.4.21.36	PANCREATIC ELASTASE	16
E.C.3.4.21.37	LEUKOCYTE ELASTASE	3
E.C.3.4.21.50	LYSYL ENDOPEPTIDASE	2
E.C.3.4.21.59	TRYPTASE	1
E.C.3.4.21.62	SUBTILISIN	25
E.C.3.4.21.64	ENDOPEPTIDASE K	5
E.C.3.4.21.66	THERMITASE	4
E.C.3.4.21.68	T-PLASMINOGEN ACTIVATOR	5
E.C.3.4.21.69	PROTEIN C (ACTIVATED)	2
E.C.3.4.21.73	U-PLASMINOGEN ACTIVATOR	1
E.C.3.4.21.78	CYTOTOXIC T-LYMPHOCYTE PROTEINASE 1	1
E.C.3.4.21.79	CYTOTOXIC T-LYMPHOCYTE PROTEINASE 2	1
E.C.3.4.21.80	STREPTOGRISIN A	5
E.C.3.4.21.81	STREPTOGRISIN B	2
E.C.3.4.21.88	REPRESSOR LEXA	2
E.C.3.4.22.2	PAPAIN	13
E.C.3.4.22.14	ACTINIDAIN	2
E.C.3.4.22.28	PICORNAIN 3C	19
E.C.3.4.22.30	CARICAIN	1
E.C.3.4.23.0	HIV PROTEASE	44
E.C.3.4.23.1	PEPSIN A	6
E.C.3.4.23.4	CHYMOSIN	3
E.C.3.4.23.5	CATHEPSIN D	2
E.C.3.4.23.15	RENIN	3
E.C.3.4.23.20	PENICILLOPEPSIN	8
E.C.3.4.23.21	RHIZOPUSPEPSIN	5
E.C.3.4.23.22	ENDOTHIAPEPSIN	21
E.C.3.4.23.23	MUCOROPEPSIN	2
E.C.3.4.24.18	MEPRIN A	1
E.C.3.4.24.21	ASTACIN	6

E.C. number	Name	Number of structures in PDB
Hydrolases cont'd		
E.C.3.4.24.26	PSEUDOLYSIN	1
E.C.3.4.24.27	THERMOLYSIN	14
E.C.3.4.24.46	ADAMALYSIN	1
E.C.3.5.1.1	ASPARAGINASE	1
E.C.3.5.1.28	N-ACETYLMURAMOYL-L-ALANINE AMIDASE	1
E.C.3.5.1.38	GLUTAMINASE-(ASPARAGIN-)ASE	2
E.C.3.5.1.52	PEPTIDE-N4-(N-ACETYL-BETA-GLUCOSAMINYL)ASPARAGINE AMIDASE	1
E.C.3.5.1.59	N-CARBAMOYLSARCOSINE AMIDASE	1
E.C.3.5.2.6	BETA-LACTAMASE	7
E.C.3.5.3.3	CREATINASE	1
E.C.3.5.4.4	ADENOSINE DEAMINASE	2
E.C.3.6.1.1	INORGANIC PYROPHOSPHATASE	3
E.C.3.6.1.7	ACYLPHOSPHATASE	1
E.C.3.6.1.34	H(+)-TRANSPORTING ATP SYNTHASE	1
E.C.3.8.1.5	HALOALKANE DEHALOGENASE	9
Lyases Total 114 Unique 16		
E.C.4.1.1.1	PYRUVATE DECARBOXYLASE	2
E.C.4.1.1.22	HISTIDINE DECARBOXYLASE	1
E.C.4.1.1.39	RIBULOSE-BISPHOSPHATE CARBOXYLASE	6
E.C.4.1.1.48	INDOLE-3-GLYCEROL-PHOSPHATE SYNTHASE	1
E.C.4.1.1.64	2,2-DIALKYLGLYCINE DECARBOXYLASE (PYRUVATE)	4
E.C.4.1.2.13	FRUCTOSE-BISPHOSPHATE ALDOLASE	2
E.C.4.1.3.7	CITRATE (SI)-SYNTHASE	12
E.C.4.1.3.18	ACETOLACTATE SYNTHASE	1
E.C.4.1.99.2	TYROSINE PHENOL-LYASE	1
E.C.4.2.1.1	CARBONATE DEHYDRATASE	66
E.C.4.2.1.3	ACONITATE HYDRATASE	7
E.C.4.2.1.11	PHOSPHOPYRUVATE HYDRATASE	7
E.C.4.2.1.20	TRYPTOPHAN SYNTHASE	1
E.C.4.2.2.2	PECTATE LYASE	1
E.C.4.2.99.18	DNA-(APURINIC OR APYRIMIDINIC SITE) LYASE	1
E.C.4.3.1.8	HYDROXYMETHYLBILANE SYNTHASE	1
Isomerases Total 99 Unique 12		
E.C.5.1.1.7	DIAMINOPIMELATE EPIMERASE	1
E.C.5.1.2.2	MANDELATE RACEMASE	3
E.C.5.1.3.2	UDP-GLUCOSE 4-EPIMERASE	1
E.C.5.2.1.8	PEPTIDYLPROLYL ISOMERASE	15
E.C.5.3.1.1	TRIOSEPHOSPHATE ISOMERASE	19
E.C.5.3.1.5	XYLOSE ISOMERASE	52
E.C.5.3.3.4	MUCONOLACTONE DELTA-ISOMERASE	1
E.C.5.4.2.1	PHOSPHOGLYCERATE MUTASE	1
E.C.5.4.2.2	PHOSPHOGLUCOMUTASE	1
E.C.5.4.99.5	CHORISMATE MUTASE	3
E.C.5.5.1.1	MUCONATE CYCLOISOMERASE	1
E.C.5.5.1.7	CHLOROMUCONATE CYCLOISOMERASE	1
Ligases Total 24 Unique 9		
E.C.6.1.1.1	TYROSINE-TRNA LIGASE	7
E.C.6.1.1.10	METHIONINE-TRNA LIGASE	2
E.C.6.1.1.11	SERINE-TRNA LIGASE	4
E.C.6.1.1.18	GLUTAMINE-TRNA LIGASE	1
E.C.6.2.1.1	ACETATE-COA LIGASE	1
E.C.6.3.1.2	GLUTAMATE-AMMONIA LIGASE	3
E.C.6.3.2.3	GLUTATHIONE SYNTHASE	2
E.C.6.3.2.19	UBIQUITIN-PROTEIN LIGASE	2
E.C.6.3.4.15	BIOTIN-[ACETYL-COA-CARBOXYLASE] LIGASE	2

Table 1.1: A list of all the enzymes present in the January 1995 release of the PDB. At the top of each class are the total number of structures and the number of unique enzymes within that class.

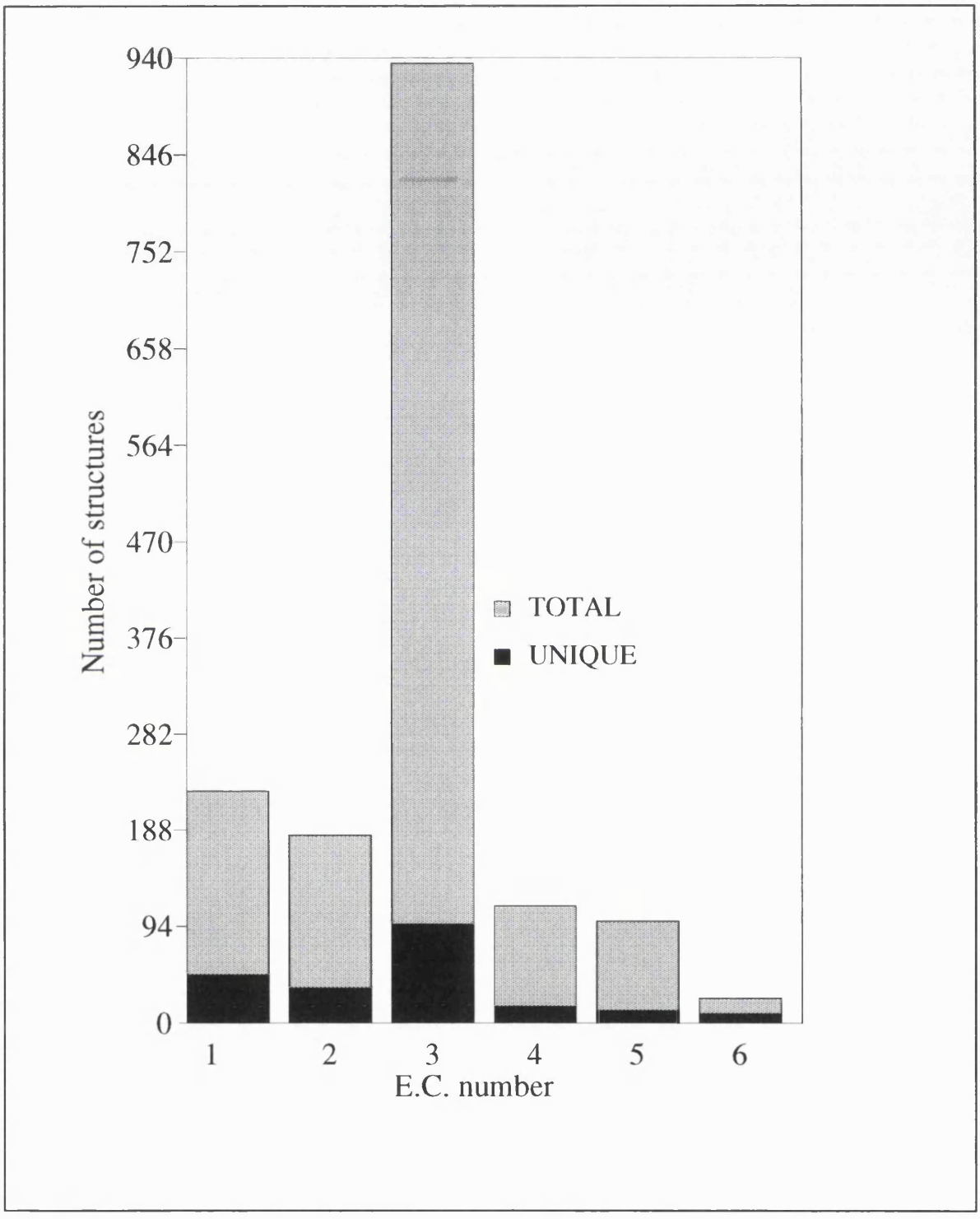


Figure 1.2: A histogram of the number of structures for each of the 6 E.C. numbers. The bars in grey are the total number of structures (*i.e* 935 for the E.C.3) whereas the black bars give the total number of unique enzymes.

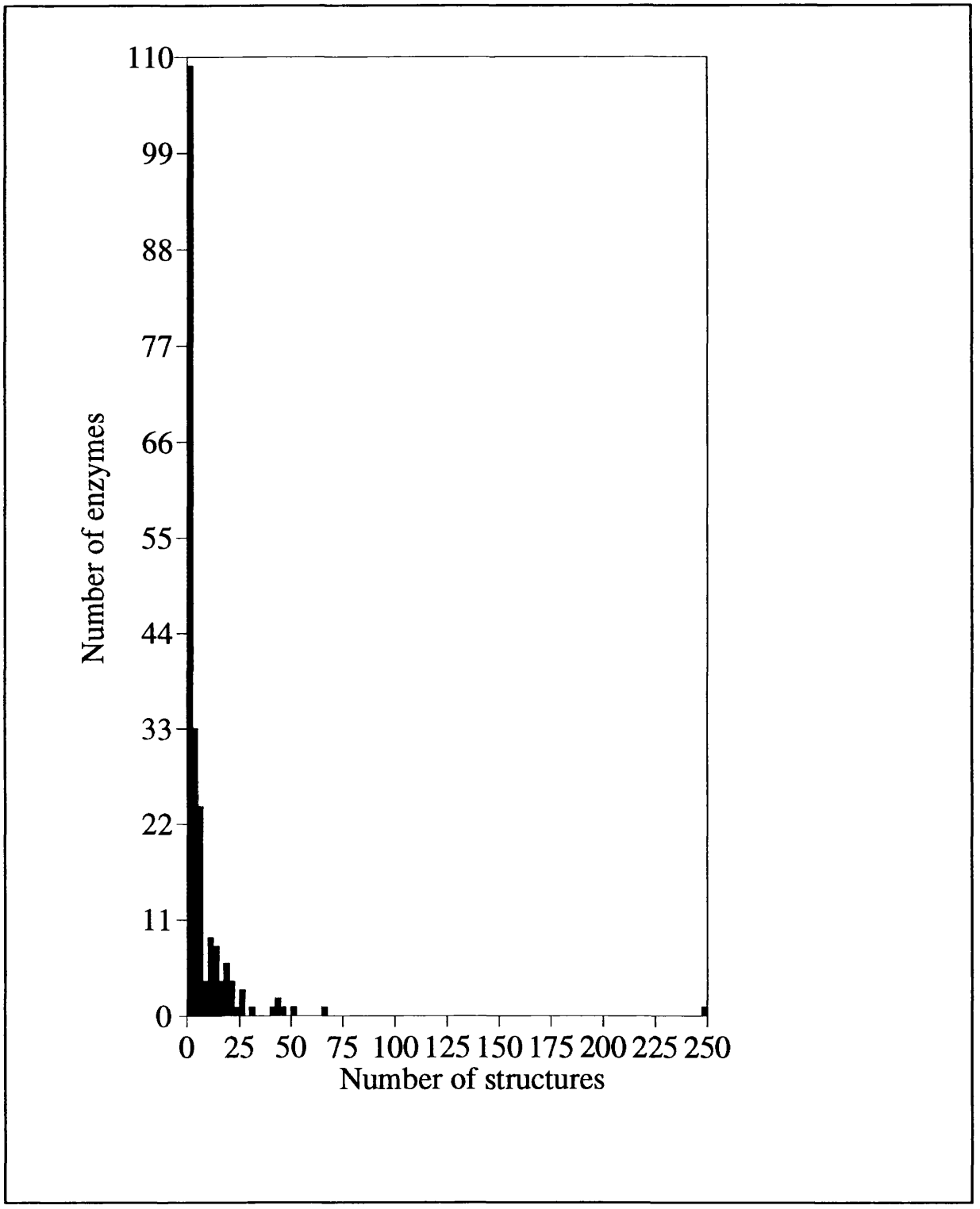


Figure 1.3: A histogram giving the total number of structures for each of the 214 unique enzymes in the January 1995 release of the PDB.

1.3 Organisation of this thesis

This thesis is divided up in the following manner. Chapter 2 describes the algorithm LIGPLOT developed to visualise protein–ligand interactions as derived from the coordinates in the PDB. Chapter 3 describes a detailed analysis of the Ser–His–Asp catalytic triad of the serine proteinases and lipases. This work led to the concept of a '3D template' that is able to describe the conformation of the active site residues of these enzymes.

A computer program called TESS was developed and is described in detail in chapter 4 which enables the construction of 3D templates for any constellation of residues in the PDB. Chapter 5 uses this software to compare and contrast the conformation of catalytic triads found in the α/β hydrolase enzymes and the serine proteinases and lipases. This analysis illustrates how convergent evolution has enabled nature to use similar catalytic machinery to catalyse different reactions. Chapter 6 investigates the orientation of ligand binding sites relative to the catalytic triad.

In chapter 7, the TESS program is used to investigate metal binding sites in proteins. Specifically, we look at triads formed in these sites such as metal–His–Asp. There are many different types of metal sites in the PDB and the structure of this triad is dependent on more than one factor. We also note similarities in the structure of the metal–His–Asp triad and the Ser–His–Asp triad.

Chapter 8 describes the problems involved in trying to automatically produce a database of '3D templates' describing all enzyme active sites in the PDB; these problems are illustrated by using two enzymes, lysozyme and ribonuclease, as examples.

1.4 References

Altschul S.F., Warren G., Webb M., Eugene W.M. & Lipman D.J (1990) Basic local alignment tool *J. Mol. Biol.* **215** 403–410

Bairoch A. The ENZYME data bank in 1995 (1996) *Nucleic Acids Res.* **24** 221-222

Bairoch A. & Boeckmann B. (1994) The SWISS–PROT protein sequence data bank: current status *Nucleic Acid Research* **22** 3578–3580

Bernstein F.C., Koetzle T.F., Williams G.J.B., Meyer E.F.Jr., Brice M.D., Rogers J.R., Kennard O., Shimanouchi T. & Tasumi M. (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures *J. Mol. Biol.* **112** 535–542

Bielka H., Dixon H. B. F., Karlson P., Liebecq C., Sharon N., Van Lenten E. J., Velick S. F., Vliegenthart J. F. G. & Webb E.C. (1992) *E. C. Enzyme nomenclature 1992: recommendations of the Nomenclature Committee of the International Union Of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes. Nomenclature Committee of the International Union of Biochemistry* Academic Press, Inc., (London) Ltd.

Blake C.C.F., Koenig D.F., Mair G.A., North A.C.T., Phillips D.C. & Sarma V.R. (1965) Structure of hen egg white lysozyme. A three dimensional Fourier synthesis at 2.0Å resolution *Nature* **206** 757–761

Blake C.C.F., Mair G.A., North A.C., Phillips D.C. & Sarma V.R. (1967*a*) On the conformation of hen egg white lysozyme *Proc. Roy. Soc. Lond. Series B: Biol. Sci.* **167(9)** 365–377

Blake C.C.F., Johnson L.N., Mair G.A., North A.C.T., Phillips D.C. & Sarma V.R. (1967*b*) Crystallographic studies of the activity of hen egg white lysozyme *Proc. Roy. Soc. Lond. Series B: Biol. Sci.* **167(9)** 378–388

Blow D. M., Birktoft J. J. & Hartley B. S. (1969) Role of a buried acid group in the mechanism of action of chymotrypsin *Nature* **221** 337–340

Blow D. M. (1976) Structure and mechanism of chymotrypsin *Acc. Chem. Res.* **9** 145–152

Blundell T.L. & Johnson L.N. (1976) *Protein Crystallography* Academic Press, London.

Branden C-I & Jones A.J. (1990) Between objectivity and subjectivity *Nature* **343** 687–689

Braunstein A.E. (1960) *The Enzymes*, 2^{nd} edition (Boyer P.D., Lardy H., & Myrback K., eds.) **2** Academic Press, New York.

Breslow D.S. (1958) Mechanism of thiamine action. Evidence from studies on model systems *J. Am. Chem. Soc.* **80** 3719–3729

Brooks B., Bruccoleri R., Olafson B., States D., Swaminathan S. & Karplus, M. (1983). CHARMM: A Program for Macromolecular Energy, Minimization, and Molecular Dynamics Calculations *J. Comp. Chem.* **4** 187–217

Brunger A.T., Kuriyan J., & Karplus, M. (1987) Crystallographic R Factor Refinement by Molecular Dynamics *Science* **235** 458–460

Brunger A.T. (1992) Free R value: a novel statistical quantity for assessing the accuracy of crystal structures *Nature* **355** 472–475

Gutfriend H. & Sturtvevant J.M. (1956) The mechanism of the chymotrypsin catalysed reaction *Proc. Natl. Acad. Sci. USA* **42** 719–728

Harel M., Sussman J.L. & Silman I. (1991) γ–chymotrypsin is a complex of α–chymotrypsin with its own autolysis products *Biochemistry* **30** 5217–5225

Hartley B.S. & Kilby B.A. (1954) The reaction of *p*–nitrophenyl esters with chymotrypsin and insulin *Biochem J.* **56** 288–297

Hooft R.W.W., Vriend G., Sander C. & Abola E.E. (1996) Errors in protein structures *Nature* **381** 272

Kendrew J.C., Dickerson R.E., Strandberg B.E., Hoit R.G., Davies D.R., Phillips D.C. & Shore V.C. (1960) Structure of myoglobin *Nature* **185** 422–427

Koshland D.E. Jr (1954) *The Mechanism of enzyme action* (McElroy W.D. and Glass B. eds.) Academic Press, New York

Laskowski R.A., MacArthur M.W., Moss D.. & Thornton J.M. (1993) PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.* **126** 283–291.

Lowry T.M. & Faulkner I.J. (1925) Studies of dynamic isomerase. Amphoteric solvents as catalysts for the mutorotation of the sugars *J. Chem. Soc.* **127** 2883–2887

Matthews B.W., Sigler P.B., Henderson R. & Blow D.M. (1967) Three–dimensional structure of tosyl–α–chymotrypsin *Nature* **214** 652–656

Orengo C. (1994) Classification of protein folds *Curr. Opin. Stuct. Biol.* **4** 429–440

Pauling L. & Corey R.B. (1951) Configurations of polypeptide chains with favored orientations around single bonds: two new pleated sheets *Proc. Natl. Acad. Sci.* **37** 729–740

Pauling L., Corey R.B., & Branson H.R. (1951) The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain *Proc. Natl. Acad. Sci., USA* **37** 205–211

Perutz M.F., Rossman M.G., Cullis A.F., Muirhead H., Will G. & North A.C.T. (1960) Structure of hemoglobin *Nature* **185** 416–422

Sanger F. & Tuppy H. (1951*a*) The amino acid sequence in the phenylalanyl domain of insulin: the identification of lower peptides from partial hydrosylates *Biochem J.* **49** 463–481

Sanger F. & Tuppy H. (1951*b*) The amino acid sequence in the phenylalanyl domain of insulin: the investigation of peptides from enzymatic hydrolysis *Biochem J.* **49** 481–490

Schoellman G. & Shaw E. (1953) Direct evidence for the role of Histidine in the active site of chymotrypsin *Biochemistry* **2** 252–255

Schaffer N.K., May S.C. & Summerson W.H. (1953) Serine phosphoric acid from diisopropyl chymotrypsin *J. Biol. Chem.* **202** 67–76

Vriend G. (1990) WHAT IF: A molecular modelling and drug design program *J. Mol. Graphics* **8** 52–56

Westheimer F.H., Fisher H.F., Conn E.E., & Vennesland B. (1951) The enzymmatic transfer of hydrogen from alcohol to DPN *J. Am. Chem. Soc.* **73** 2403

Whitaker J.R. & Jandorf B.J. (1956) Specific reactions of dinitrofluorobenzene with active groups of chymotrypsin *J. Biol. Chem.* **223** 751–764

Wilson I.B., Bergman F. & Nachmansohn D. (1950) Acetylcholinesterase: mechanism of action *J. Biol. Chem.* **186** 781–790

Zaug A.J. & Cech T.R. (1986) The intervening sequence RNA of *Tetrahymena* is an enzyme *Science* **231** 431–475

Chapter 2

LIGPLOT: a program to generate schematic diagrams of protein–ligand interactions

2.1 Introduction

The exact nature of non-covalent interactions in macromolecules, such as those between a protein and a ligand, are often difficult to visualise and comprehend easily without detailed inspection on a graphics terminal. This makes them especially difficult to illustrate in two dimensions, as when presented in a paper, even with the benefit of stereo plots. Schematic diagrams are frequently used in the literature to try to clarify the interactions; to show, for example, which atoms of the ligand are hydrogen-bonded to which residues in the protein. These are often drawn by hand, sometimes with the help of drawing packages and are invariably time-consuming to produce.

Here we describe a program called LIGPLOT that generates schematic diagrams automatically from the 3D coordinates of the protein and its bound ligand.

These diagrams illustrate the pattern of interactions between the two molecules and are particularly useful for comparing different structures or for studying the interactions between different ligands and the same enzyme.

The interactions shown by LIGPLOT are hydrogen bonds and hydrophobic contacts. Hydrogen bonds are indicated by dashed lines between the atoms involved; each hydrogen-bonded residue from the protein is shown in full, although there is an option to include/exclude its main-chain atoms. Hydrophobic contacts are indicated more schematically; residues from the protein involved in these contacts are represented by an arc with spokes radiating towards the ligand atoms they contact. The contacted atoms are shown with spokes radiating back.

Atom accessibility can also be depicted. The ligand atoms can be colour-coded to indicate their accessibility to solvent. Together, all this information provides a schematic representation of the types and locations of the ligand's important non-covalent interactions.

The program is completely general and will work for any ligand. Indeed, it has also been used for segments of proteins to show, for example, interactions between a helix and the residues in its vicinity.

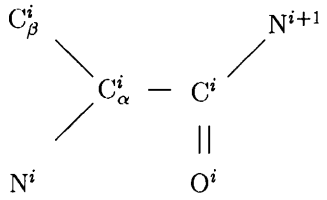
2.2 The algorithm - an overview

The LIGPLOT algorithm consists of many stages but in principle is very simple. It takes the 3D structure of the ligand and 'unrolls' it, flattening it out onto the 2D page. As it does so, it takes with it the hydrogen-bonded sidechains and sidechains involved in hydrophobic contacts, flattening those out too and placing them so that the overlap of atoms and the crossing of bonds in the final diagram is kept to a minumum.

The unrolling is performed about each of the structure's 'rotatable' bonds.

The structure on either side of such a bond can be independently rotated or adjusted; bonds that are part of a ring are *non*-rotatable as moving the structure on one side of them affects the structure on the other side by virtue of the ring connection.

The unrolling process involves rotating the structure on either side of the bond in such a way that the bonds springing directly from its two ends come to lie in the same plane. For example, consider a main-chain Cα–C bond, as shown below.



This is a rotatable bond as it is not part of a ring. Firstly, the structure on the left-hand side of it is rotated about the C^i_α atom until the N^i and C^i_β atoms lie in the same plane as the C^i_α and the C^i atoms. This rotation, which will be described in detail later, is repeated for the structure on the right-hand side of the bond, which is rotated until the N^{i+1} and O^i atoms also lie in the same plane.

The result is that all the atoms connected to the rotatable bond come to be the same plane. Repetition of the procedure on all the rotatable bonds in the structure, in turn, gives a structure that has been completely flattened onto the page. This unrolling usually proceeds from one end of the ligand to the other, though where branching occurs, the branches have to be unrolled in turn.

Note that none of the bond-lengths are distorted in this process, and even some of the bond angles are maintained. Thus the N^i–C^i_α–C^i_β and N^{i+1}–C^i–O^i bond angles are the same as in the 3D structure.

Although completely flat, the structure at this stage will probably include extensive overlap between atoms and bonds, resulting in a very congested and

confusing diagram of the interactions. A 'clean-up' procedure tackles this problem. This involves, once again, cycling through each of the rotatable bonds in turn. This time a test is made to see if a rotation of one side of the structure through 180 degrees about the bond will reduce the number of atom clashes and bond overlaps. This is just a flip of 180 degrees about the bond. If the flip reduces the overlaps, it is retained, otherwise the original structure is kept. The entire cycle of all possible flips is repeated several times until the number of atom and bond overlaps reaches a minimum, and the diagram of the structure is plotted.

Of course, this procedure will introduce some distortions into the structure, as is inevitable when converting a 3D object into a 2D representation. For example, all the torsion angles, being completely flat, will be either 0 or 180 degrees. As an extreme example, a *trans* peptide in the ligand may occasionally appear as *cis*, although this is actually very rare. Many angles will be distorted by the flattening process. Perhaps more significantly, residues in the protein that are close to one another in 3D, might sometimes appear on either side of the ligand, simply because this makes the interactions clearer to see. Such unavoidable side-effects are a consequence of any attempt to present structures in 2D and do not detract from the information the diagrams are aiming to convey.

2.3 Details of the algorithm

2.3.1 Coordinates, hydrogen bonds and connectivity

Figure 2.1 illustrates the principal stages of the LIGPLOT algorithm and Figure 2.2 shows how a structure evolves during these stages, from its starting 3D structure to its final 2D representation.

The first stage involves the reading in of the 3D coordinates of the structure from the specified PDB file and identifying the atoms belonging to the ligand (or

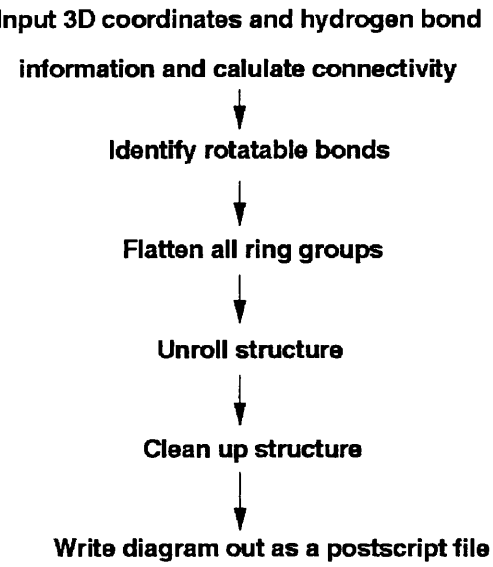


Figure 2.1: Flow diagram illustrating the main stages in the LIGPLOT algorithm.

segment of protein) as defined by the user. The protein residues that are either hydrogen bonded to the ligand, or are involved in hydrophobic interactions with it, are identified from two lists of such interactions. These can be generated by the user and supplied to the LIGPLOT program in the format specified in the operating instructions.

The program we use, for generating both a list of hydrogen bonds and of non-bonded interactions, is HBPLUS (McDonald and Thornton, 1994) which,like LIGPLOT, is available by anonymous ftp. The program computes all possible positions for hydrogen atoms (H), attached to donor atoms (D), which satisfy specified geometrical criteria with acceptor atoms (A) in the vicinity (McDonald and Thornton, 1994). The criteria used here are that the H–A distance < 2.7Å, the D–A distance < 3.3Å, the D–H–A angle > 90 degrees, and that the H–A–AA angle > 90 degrees, where the AA atom is the one atached to the acceptor, usually preceding it along the amino acid chain. These criteria can be altered by the user if so desired. Hydrophobic interactions are defined as any carbon

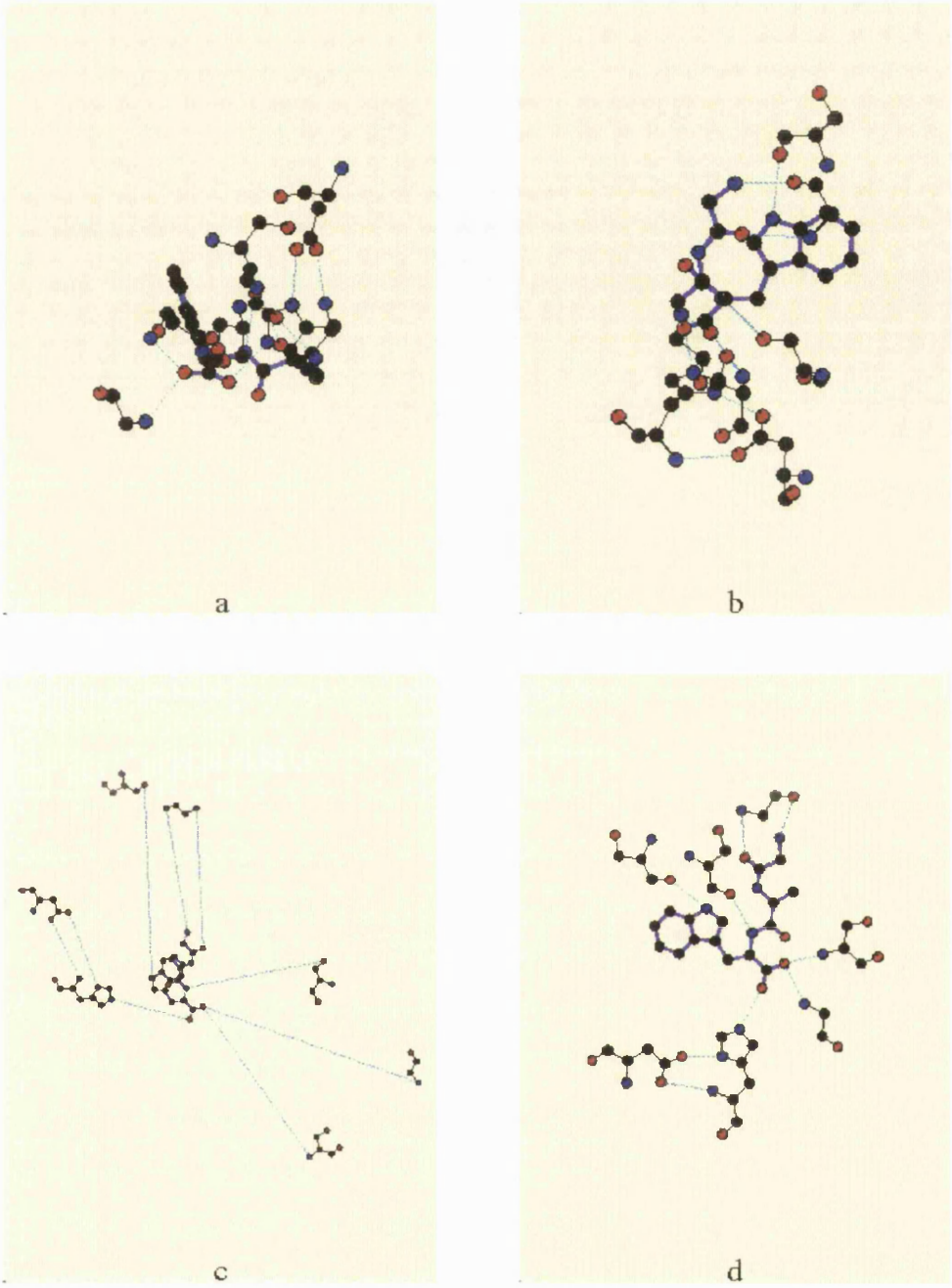


Figure 2.2: An example of how a ligand is converted from its starting 3D structure to a 2D LIGPLOT representation of its interactions. The ligand shown here is Gly-Ala-Trp, complexed with γ-chymotrypsin (PDB code 8gch). This is the same structure as in Figure 2.3, but here only the hydrogen-bonded groups from the protein are included. The bold lines represent the ligand's bonds, the thin lines represent the bonds in the sidechains of the protein, while the dashed lines correspond to the hydrogen bonds. The four stages shown are: *a.* simple orthographic projection of the starting 3D structure; *b.* after the flattening of all rings and the unrolling of the entire structure onto a 2D plane, but with a considerable amount of atom-clashes and bond-overlaps to be got rid of; *c.* explosion of the hydrogen-bonded groups away from the ligand to ease minimisation of atom- and bond-overlaps (in fact, for clarity, the groups shown here have been exploded out to only a quarter of their usual distance); and *d.* final picture after flipping of rotatable bonds to minimise overlaps and swinging and relaxation of hydrogen-bonded groups back toward the ligand.

atom that is within the sum of the van der Waals radius of any other atom plus a distance cut–off. We set the distance cut–off as 1.0Å.

HBPLUS can also list all non-bonded contacts between atoms that are less than a specified distance apart. The cut-off used for LIGPLOT is 3.9 Å, but this can be amended by the user. This list is used by LIGPLOT when extracting all hydrophobic interactions, being just those between pairs of carbon atoms.

In some molecules, the above interactions are not the only ones of interest. For example, in the catalytic triad of chymotrypsin a single layer of hydrogen bonds would fail to include the functionally important His57–Asp102 catalytic pair. Thus LIGPLOT has an option which also allows additional sidechains, not directly bonded to the ligand, to be included. In this case one would specify that His-Asp pairs are to be included where the His is hydrogen-bonded to the ligand and the Asp is hydrogen-bonded to the His. The option is also useful in cases where the hydrogen bonding between the protein and ligand is mediated by one or more water molecules. Thus LIGPLOT can not only show the water molecules that are hydrogen bonded to the ligand, but also the residues from the protein that are hydrogen-bonded to these waters.

Hydrogen-bonded and hydrophobic groups are treated slightly differently. In the hydrogen-bonded groups all sidechain atoms are retained, with there being an option to retain main-chain atoms as well. Hydrophobic groups, on the other hand, are represented by a single position for the residue as a whole. This position is linked to the atoms on the ligand with which it is in contact by 'virtual' bonds. This simplifies the unrolling procedure while making the final picture more informative.

The covalent connectivity of all the retained atoms is calculated using a simple distance cut-off of 1.85Å. Various bonds are then 'cut' to simplify the unrolling and clean-up stages. These include bonds linking adjacent hydrogen-bonded

groups. For example, if residues 195 and 196 in the protein are both hydrogen-bonded to the ligand, the peptide bond joining them will be removed so that the two sidechains can move independently of one another during the unrolling and clean-up of the structure. This reduces the constraints on the minimisation process and increases the chances of a clearer diagram resulting (*ie* one with fewer atom-clashes and bond-overlaps).

Any atoms that cannot be reached by tracing along connected bonds from some starting point on the ligand are deleted. This avoids problems that might be caused by chain-breaks; for the unrolling procedure to be successful, all atoms must be connected by one or more bonds, or they will 'float free' of the structure and interfere with the final diagram.

Also read in at this stage are the atom accessibilities, if required. These give a measure of the solvent accessibility of each atom, and are calculated by the program ACCESS (Hubbard, 1991). The accessibility values are represented on the final plot by different shading of the background of each ligand atom.

2.3.2 Identification of bonds for rotation

The second stage (see Figure 2.1) determines which of the structure's bonds are 'rotatable' - *ie* those at which the unrolling procedure can be applied. As mentioned above, any bonds in ring structures are non-rotatable as the structure on either side of them cannot be moved independently; any movement on one side affects the structure on the other side and so, in general, any attempt at flattening will distort the overall structure.

This applies not just to bonds in recognised ring groups - such as in the aromatic ring of Phe sidechains - but to *any* bond that is part of a closed loop of connected bonds. In other words, if it is possible to track through the structure from one end of a bond to its other end, the bond must be part of a closed loop

within the structure and so cannot be treated as rotatable.

Hydrogen-bonds often create such loops. If a sidechain forms two hydrogen bonds to the ligand - as when a carboxylate group has both its oxygens involved in hydrogen bonds (*eg* Asp102 in Figure 2.3) - it is possible to trace a loop from one hydrogen bond through the ligand to the other hydrogen-bond and then through the sidechain back to the starting point. The existence of the loop makes flattening impossible since rotation about one bond to improve planarity will probably worsen the planarity at the other bond. Such loops are dealt with by making one of the hydrogen bonds 'elastic' and the other a rotatable bond. The two atoms either side of the elastic bond can be moved independently of one another, stretching and distorting the bond whenever the structures either side of it are moved independently. Thus, in Figure 2.3, the O^{δ_1}–N bond from the Asp102 to the His57 might be elastic, while the O^{δ_2}–N^{δ_1} bond is a rotatable bond about which the Asp102 can be flipped this way and that. Alternatively, their roles might be reversed. Which is which depends merely on the order of the atoms in the original PDB file.

Internal hydrogen bonds, between atoms entirely within the specified ligand, also create problematical loops. Thus they, too, are non-rotatable and are treated as elastic bonds, being free to stretch as the ligand is gradually unrolled.

Another type of non-rotatable bond is any end-bond. That is, any bond having links at one end only (*eg* the C^i=O^i bond in the example above). The flattening of these bonds is usually taken care of automatically when the flattening process is applied at the connected end. Thus, for the C^i=O^i bond, the two rotations about the C^i_α–C^i bond described above are sufficient to bring the O^i atom into the same plane as the other 4 atoms shown. There is a special case, however, where the flattening is not automatic, and this is when the end-bond is attached to a ring group. Cases such as this are dealt with in the next stage.

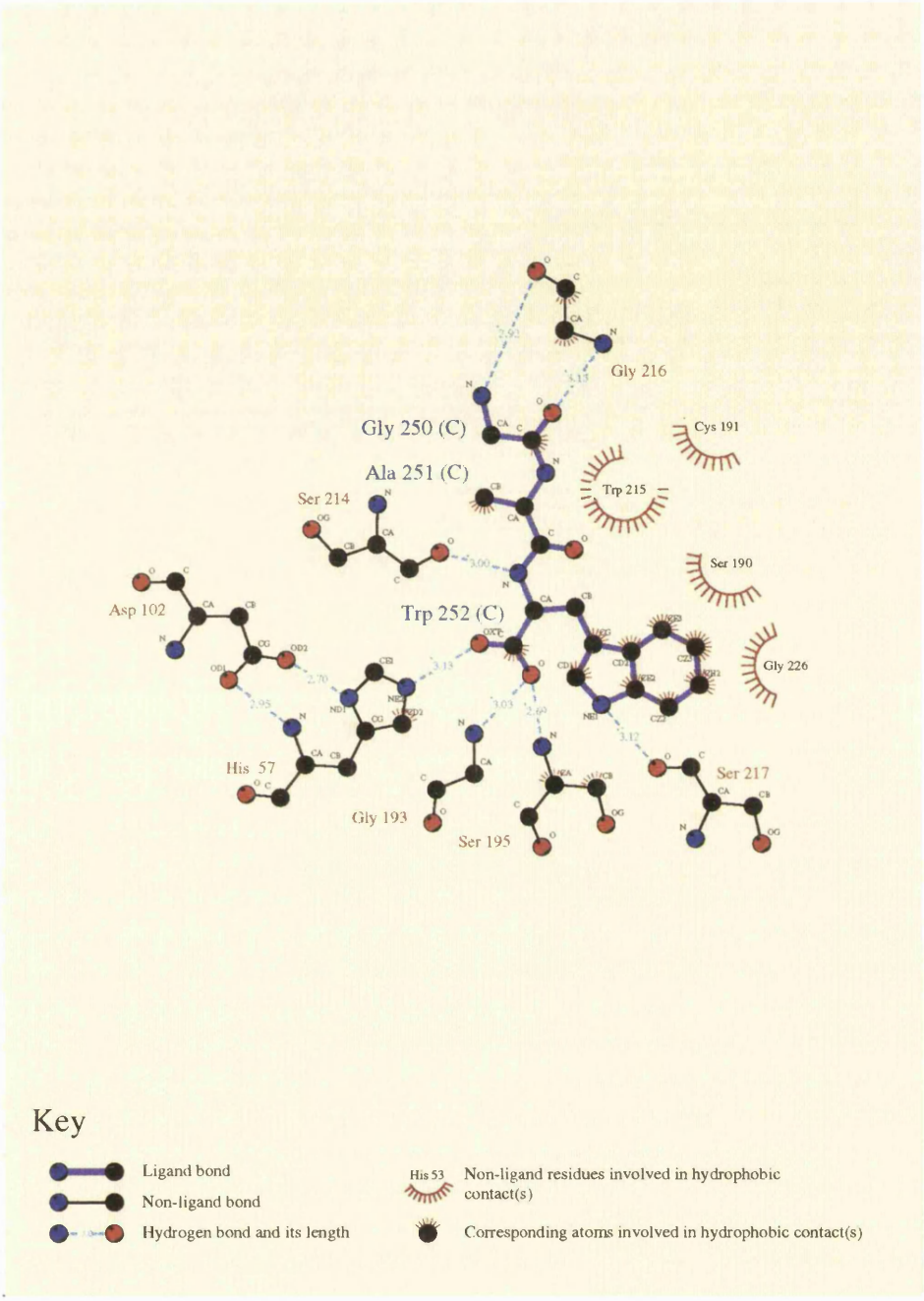


Figure 2.3: A LIGPLOT diagram of the active site of chymotrypsin (PDB code 8gch) complexed with the tripeptide Gly–Ala–Trp (residues 250–252, chain C). The bold bonds belong to the ligand, the thin bonds belong to the hydrogen-bonded residues from the protein, and the dashed lines represent the hydrogen bonds between ligand and protein. Hydrophobic contacts made with the protein are indicated by the spoked arcs pointing towards the ligand. Corresponding spokes on the ligand atoms indicate which atoms are involved in these contacts. Similarly, the atoms in the hydrogen-bonded groups involved in hydrophobic contacts are marked by spokes pointing in the direction of the contact atoms. For example, the C and CA atoms of the hydrogen-bonded group Gly 216 are involved in hydrophobic interactions with Trp 252 on the ligand. The letters in parentheses in the residue names are the corresponding chain identifiers. The diagram illustrates the catalytic triad of His 57, Asp 102 and Ser 195, as well as showing the ligand's Trp 252 residue nestling in the highly hydrophobic specificity pocket in the active site of the enzyme.

2.3.3 Flattening of ring groups

The third stage involves forcing flat all ring groups in the structure. This ensures that they are perfectly planar before the unrolling procedure of the next stage commences. Even small distortions in the rings can prevent the unrolling procedure from attaining a perfectly flat structure. The reason for this is that if a distorted ring separates one part of the structure from another, then the two halves can be independently flattened but, so long as the ring remains distorted, the entire structure will not be flat. A perfectly planar ring, on the other hand, circumvents this problem.

Many ring groups are expected to be planar, and any deviation from planarity is usually minor. Some groups, however, have standard non-planar conformations; for example, the pyranose ring-structure of glucose adopts the 'chair' conformation.

In LIGPLOT all the ring groups are forced flat in a fairly crude manner. A best-fit plane is first calculated through the ring atoms, and they are transformed so that this plane lies in the x-y plane. Also transformed are any atoms attached to the ring by end-bonds. The transformed atoms now have their z-coordinates set to zero, thus crudely flattening the ring and any end-bonds attached to it. Finally the transformed atoms, now perfectly planar, are transformed back into the structure by applying the reverse transformation. Depending on how unplanar the original ring-group is, this procedure can distort the ring's bond lengths and bond angles. More importantly, it can distort the bonds attached to the ring-group, such as the end-bonds, sometimes quite severely. This side-effect is remedied by checking all non-ring bonds and stretching or contracting them (moving the two halves of the structure either side of them accordingly) back to their original length.

2.3.4 Unrolling the structure

The structure is now progressively unrolled as follows. Each rotatable bond is taken in turn and the whole structure is transformed to place this bond along the negative x-axis with one or other of its atoms at the origin. The number of atoms bonded to the one at the origin (either covalently or through hydrogen bonds) is counted. The count does not include the bond's other atom. Figure 2.4 shows what happens in the cases where there are 1, 2 or 3 atoms attached.

The simplest case is where there is only one other atom connected (Figure 2.4*a*). The entire structure on that side of the bond is rotated about the x-axis, either clockwise or anti-clockwise, to place that atom in the x-y plane. The angle α shown in the figure remains unchanged.

If there are two atoms attached (Figure 2.4*b*), the normal to the plane defined by these two atoms and the one at the origin is computed. The entire structure on that side of the bond is then rotated, first about the z-axis and then about the x-axis, to bring this normal in line with the positive z-axis. This places the two atoms of interest in the x-y plane. The structure is then further rotated about the z-axis until the x-axis bisects the angle defined by the three atoms (angle β in Figure 2.4*b*).

In the case where there are three or more atoms attached, the procedure is a little more complicated. Each atom is taken in turn and the part of the structure connected to it is rotated about the origin to place the atom in the x-y plane at some splay angle relative to the x-axis. For three attached bonds (Figure 2.4*c*), the splay angle is 90 degrees, so the three atoms are splayed out at 90, 180 and 270 degrees relative to the x-axis. For four attached atoms, the splay angle is 72 degrees, and so on. Of course, in these situations, none of the original 3D geometry is retained other than the original bond lengths.

Once the required rotations have been applied to this end of the rotatable

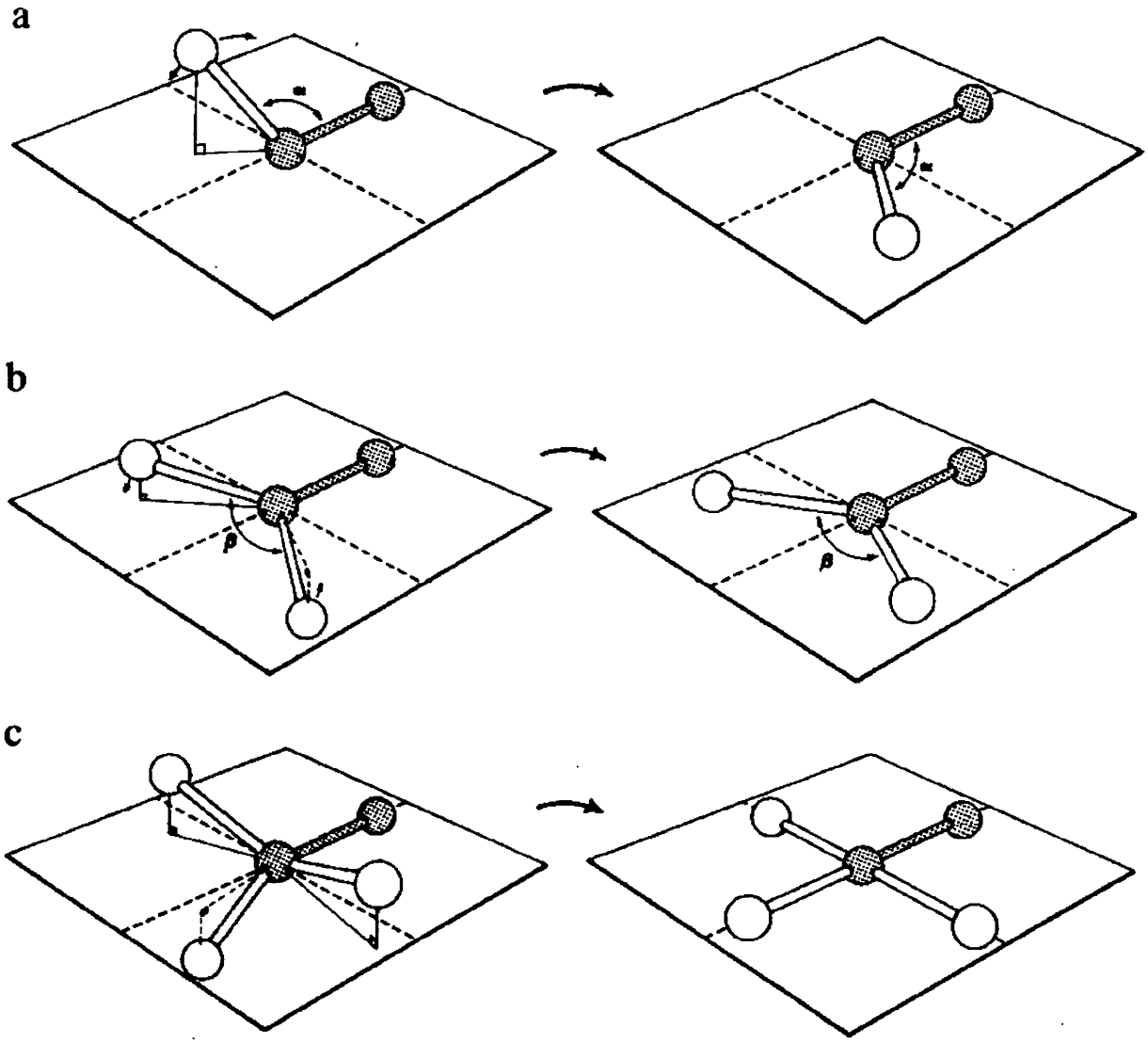


Figure 2.4: The different rotations applied when different numbers of atoms are bonded to a given rotatable bond. The left-hand pictures show the 'before' states for the three different situations depicted, while the right-hand pictures show how the atoms are transformed into the x-y plane in each case. The shaded atoms belong to the rotatable bond, lying along the x-axis, with the atom of interest at the origin. The three cases illustrated are: *a.* when one atom is attached to the rotatable bond the rotation applied brings the atom into the x-y plane, either clockwise or anti-clockwise, keeping the angle α fixed; *b.* with two atoms attached, the rotation maintains the angle β, placing the atoms in the x-y plane such that the angle β is bisected by the x-axis; *c.* with three atoms attached no attempt is made to retain any angles - the three atoms are placed at right-angles to one another in the x-y plane.

bond, the structure is flipped about the y-axis and transformed to place the bond's other atom at the origin. The procedure is then applied, if necessary, to this other end, getting the atoms attached to it into the x-y plane also. The unrolling continues in this manner until all rotatable bonds have been processed.

The end result is a completely flat structure in which all the atoms lie in the same plane, but in which parts of the structure may be folded back on one another with many atoms and bonds overlapping. Figure 2.2*b* gives an example.

2.3.5 Minimisation of atom and bond overlap

The final stage of LIGPLOT, prior to plotting, is the clean-up process in which the extent of the atom- and bond-overlaps is minimised to give as clear a diagram of the ligand's interactions as possible.

Once again, each rotatable bond is considered in turn. The structure on one side of the bond is flipped through 180 degrees about the bond to see if this flip reduces the severity of the atom-atom and bond-bond overlaps. If it does, the flipped conformation is retained, otherwise the flip is reversed to retrieve the original conformation.

The severity of the overlaps is evaluated using a simple energy function, E_{total}, consisting of two terms

$$E_{total} = E_a + wE_b.$$

where E_a is the energy due to close-contacts between non-bonded atoms in the structure, E_b is the energy due to bond overlaps, and w is a weighting factor that can be used to adjust the relative importance of the two term. Currently use $w = 0.05$; this value was derived empirically by testing LIGPLOT with various PDB files. The two energies are calculated as described below.

Atom-clash energy, E_a

The energy assigned to each atom-atom contacts is given by the inverse square of the distance between the atoms:

$$E_a = \begin{cases} 0, & \text{if } d_{ij} > d_{max}, \\ 1/d_{ij}^2, & \text{if } d_{min} < d_{ij} < d_{max}, \\ 1/d_{min}^2, & \text{if } d_{ij} < d_{min}, \end{cases}$$

where d_{ij} is the distance between atoms i and j, d_{max} is a cut-off beyond which the atoms are deemed not to be interacting, and d_{min} is a minimum distance cut-off used to guard against infinite energies arising when two atoms are practically on top of one another.

In LIGPLOT d_{min} is set to 0.1Å, and d_{max} is set to 2.5Å.

Bond overlap energy, E_b

As well as reducing atom clashes, it is also necessary to reduce the number of overlapping bonds in the picture as these can be very confusing. Only overlaps of hydrogen bonds with all other bonds (and atoms) needs to be considered; any overlaps between pairs of covalent bonds are already taken care of in the E_a energy above as any crossing covalent bonds will inevitably have their respective atoms clashing too.

Thus each hydrogen bond is considered in turn and is assigned an overlap energy according to the extent of its overlaps with the other bonds and atoms in the structure. The energy is calculated as follows. If the hydrogen bond crosses any other bond in the structure a fixed penalty C is added to the energy.

Any atoms within 0.7Å of the bond are taken to be clashing with it, provided that the perpendicular from the atom to the bond meets the bond between the two atoms defining it. The energy contribution E_{ih} from such a clash is given

by:-

$$E_{ih} = \begin{cases} 0, & \text{if } l_{ih} > l_{max}, \\ 1/l_{ij}^2, & \text{if } l_{min} < l_{ij} < l_{max}, \\ 1/l_{min}^2, & \text{if } l_{ij} < l_{min}, \end{cases}$$

where l_{ih} is the distance between the atom i and hydrogen bond h, l_{max} is the cut-off beyond which the atoms are deemed not to be interacting (here set to 0.7Å), and l_{min} is a minimum distance cut-off (set to 0.1Å) to guard against infinite energies arising when two atoms are practically on top of one another.

The net bond overlap energy E_b can be expressed as:-

$$E_b = \sum_{h=1}^{N_{h-bonds}} \left(o_h C + \sum_{i=1}^{N_{atoms}-2} E_{ih} \right),$$

where the outer summation is over the $N_{h-bonds}$ hydrogen bonds in the structure, the inner summation is over the $N_{atoms} - 2$ other atoms in the structure (*ie* excluding the two atoms in hydrogen bond h), o_h is the number of other bonds which overlap hydrogen bond h, and E_{ih} is the overlap energy between atom i and hydrogen bond h calculated from the perpendicular distance between them, as described above. In LIGPLOT the penalty C is taken to be 10; this value was derived empirically by testing LIGPLOT with various PDB files.

The total energy E_{total}, obtained from the atom-clash energy E_a and the bond overlap energy E_b, thus gives a means of measuring how much overlap there is in the picture, and the whole process becomes a minimisation problem. The flipping procedure is repeated through several cycles, allowing the structure to writhe this way and that, until the value of E_{total} ceases to change from one cycle to the next, and a minimum has been reached.

For large ligands, which may have many hydrogen-bonded sidechains attached,

the writhing can be severely cramped by all these extra groups (*eg* as is the situation in Figure 2.2*b*). To ease the process, the hydrogen-bonded groups are 'exploded' away from the ligand radially along the direction of the bond as shown in Figure 2.2*c*. That is, the sidechains of the protein are translated some distance out from the ligand, each progressively further than the last, stretching the hydrogen bonds as necessary. This gives the molecule the necessary space to unravel and, once done, the hydrogen-bonded groups can be drawn back in towards the ligand, swinging into more favourable positions and settling as near to their actual lengths as atoms clashes allow until the final picture is obtained.

2.3.6 Plot parameters

Once the clean up-process is complete a PostScript picture is generated of the final structure. The appearance of the final diagram can be modified to some extent by editing the parameters supplied in the parameter file. Some of the plot options available are:-

1. Produce a black-and-white or colour Postscript file. The colours of atoms, bonds and background colour can be defined by the user.
2. Show molecules in ball-and-stick representation, or as bonds only.
3. Include/exclude hydrogen-bonded groups and/or hydrophobic contacts.
4. Include additional residues not directly hydrogen-bonded to the ligand (*eg* the His-Asp pair in Figure 2.3). Up to 10 additional residue-pairs can be defined for inclusion.
5. Include/exclude water molecules.
6. Include/exclude internal ligand hydrogen bonds.
7. Show accessibility shading for all ligand atoms.

8. Label atoms and/or residues.

9. Produce a schematic peptide diagram of the hydrogen-bonded interactions only (see below). Here each ligand residue is represented by a single circle and non-ligand residues are represented by their name only.

10. Include/exclude a key explaining the symbols used in the diagram.

2.3.7 Placement of atom and residue names

If residue and atom names are required, these have to be placed with a minimum of overlap. Atom names are relatively straightforward to place, being located near the relevant atom at a point not interfering with any bond.

Residue names are more complicated to place. Each one is represented by a rectangle of the appropriate size, as defined by the length of the name and the height of the text characters used. This rectangle is then placed at successive trial locations on a grid of points encompassing the residue in question. At each grid-point the closest atom to the borders of the rectangle is found. If this atom does not belong to the correct residue, the trial location is discarded and the next is tried. If, on the other hand, the closest atom belongs to the correct residue, an 'energy', E_l, is computed using the closest distance d_t between the atom and the borders of the rectangle:

$$E_l = \frac{1}{d_{ideal}^2} - \frac{1}{d_t^2},$$

where d_{ideal} is an 'ideal' distance at which to place a residue label from an atom. In LIGPLOT, d_{ideal} is set to 0.6Å. If E_l is negative, the trial location is discarded as the label is too close to the atom (*ie* $d_t < d_{ideal}$). If E_l is zero, or close to zero, the label might be at a good distance from an atom, but it may also be in the vicinity of one or more atoms of *other* residues, and so cause confusion in the

final picture.

This is taken into account by calculating a new energy E_{label}:

$$E_{label} = E_l + \sum_{j=1}^{N_{other}} \frac{1}{d_t^2(j)},$$

where the sum is over the atoms belonging to residues *other* than the one of interest, and $d_t(j)$ is the nearest distance of each of these from the residue label. The more atoms of the wrong residue there are close to the label, the higher will the energy E_{label} be.

Thus the trial location with the lowest E_{label} is where the residue label is finally placed, being where it is as close to the ideal distance from one of the relevant atoms as possible while not being too close to other residue atoms.

2.3.8 Schematic peptide diagrams

Where the ligand is a large peptide, and hence has many interactions with the protein, a schematic peptide diagram can be produced to show very simply the hydrogen bonds involved. The schematic plot is based on those used to illustrate peptide–protein interactions in Zvelebil and Thornton (1993). Each residue in the ligand is represented by a single circle at the C_α position. The residue's sidechain is not shown unless it is involved in hydrogen-bond interactions with the protein.

2.3.9 Interactive modification of the diagrams

As mentioned above, the production of the final picture is essentially a minimisation procedure in which an attempt is made to minimise a somewhat arbitrary energy function. The minimisation procedure used in LIGPLOT is a rather crude one and has many of the common pitfalls associated with multiple-minima problems - namely the difficulty of finding a global, rather than just a local, minimum.

A more sophisticated approach might be to use simulated annealing, as is done by the TOPS program for optimizing its schematic topology diagrams of protein structures (Flores *et al.*, 1994).

Alternatively, the final diagram might be 'touched up' by hand on a graphics terminal. This can by done interactively using standard molecular modelling packages such as Quanta^{TM} (Molecular Simulations Inc., copyright 1986–1994). The coordinates of the final LIGPLOT picture are written out to a PDB-format file which can be read into the graphics package to be used. The picture will, of course, be perfectly flat, with all the z-coordinates of the atoms set to zero. Because some of the hydrogen bonds may be longer than their true lengths, they may need to be displayed using distance monitors defined between the atoms in question. It is then a simple matter to move the hydrogen-bonded groups around on the screen (using only translations in the x- and y-directions, and rotations about the z-axis) until the required arrangement is achieved. Residues involved in hydrophobic interactions are represented by a single carbon atom each and can also be moved around at will. Distance monitors corresponding to the hydrogen bonds (and possibly for hydrophobic interactions also) allow one to position these groups as close to their actual distances as is possible.

Once the final arrangement of the elements in the picture has been attained, the coordinates can be written out to a PDB file which can then be passed through LIGPLOT a second time to get the final PostScript picture.

2.4 Examples

Four examples of LIGPLOT outputs are given in Figure 2.3 and Figures 2.5 to 2.7.

Figure 2.3 shows a peptide substrate analogue (residues 250–252, chain C) bound to the active site of chymotrypsin, PDB code 8gch (Harel *et al.*, 1991). The

diagram illustrates the 'catalytic triad' of the enzyme, at the bottom right of the picture, comprising residues Ser 195 (which forms the acyl-enzyme intermediate), His 57 (which acts as a general acid) and Asp 102 (which orientates the His). Asp 102 is not directly hydrogen bonded to the ligand and so it was necessary to explicitly specify its inclusion in the LIGPLOT parameter file (option 4 above). Gly 193, on the right of Figure 2.3, is also catalytically important because it is thought to stabilise the transition state intermediate through hydrogen bonding. Of the other enzyme residues shown, Gly 216 and Ser 214 at the top right are responsible for hydrogen bonding to and binding the substrate. On the ligand, the Trp 252 binds into the specificity pocket as can be seen from the large number of hydrophobic contacts it makes (with residues Ser 190, Cys 191, Trp 215 and Gly 226, as well as with the Ser 195 and Gly 216 residues already mentioned because of the hydrogen bonds they make with the ligand).

Figure 2.5 is a LIGPLOT diagram of a transition state inhibitor (residue 935) bound to phospholipase A2, PDB code 1poe (Scott *et al.*, 1990). There is, as in chymotrypsin, a catalytically important His–Asp pair. His 47 in phospholipase has the same role as His 57 in chymotrypsin (*ie* as a general acid/base) and again the Asp serves to orientate the His. On this plot the atom accessibilities are indicated by the shading, showing which of the ligand atoms are near the surface of the protein (lighter shading) and which are buried (darker shading).

Figure 2.6 shows a tyrosine-phosphorylated peptide bound to the phosphotyrosine recognition domain SH2 of v-*src* (Waksman *et al.*, 1992), PDB code 1sha. Here the important residues are those responsible for the specific recognition of the phosphotyrosine; namely those involved in hydrogen-bonding to the phosphate oxygens (Arg 12, Ser 34, Glu 35, Thr 36 and in particular the Arg 32 which forms an ion pair with the phosphate group) and the hydrophobic interaction of the Lys 60 which forms a 'hydrophobic platform' for the ring of the

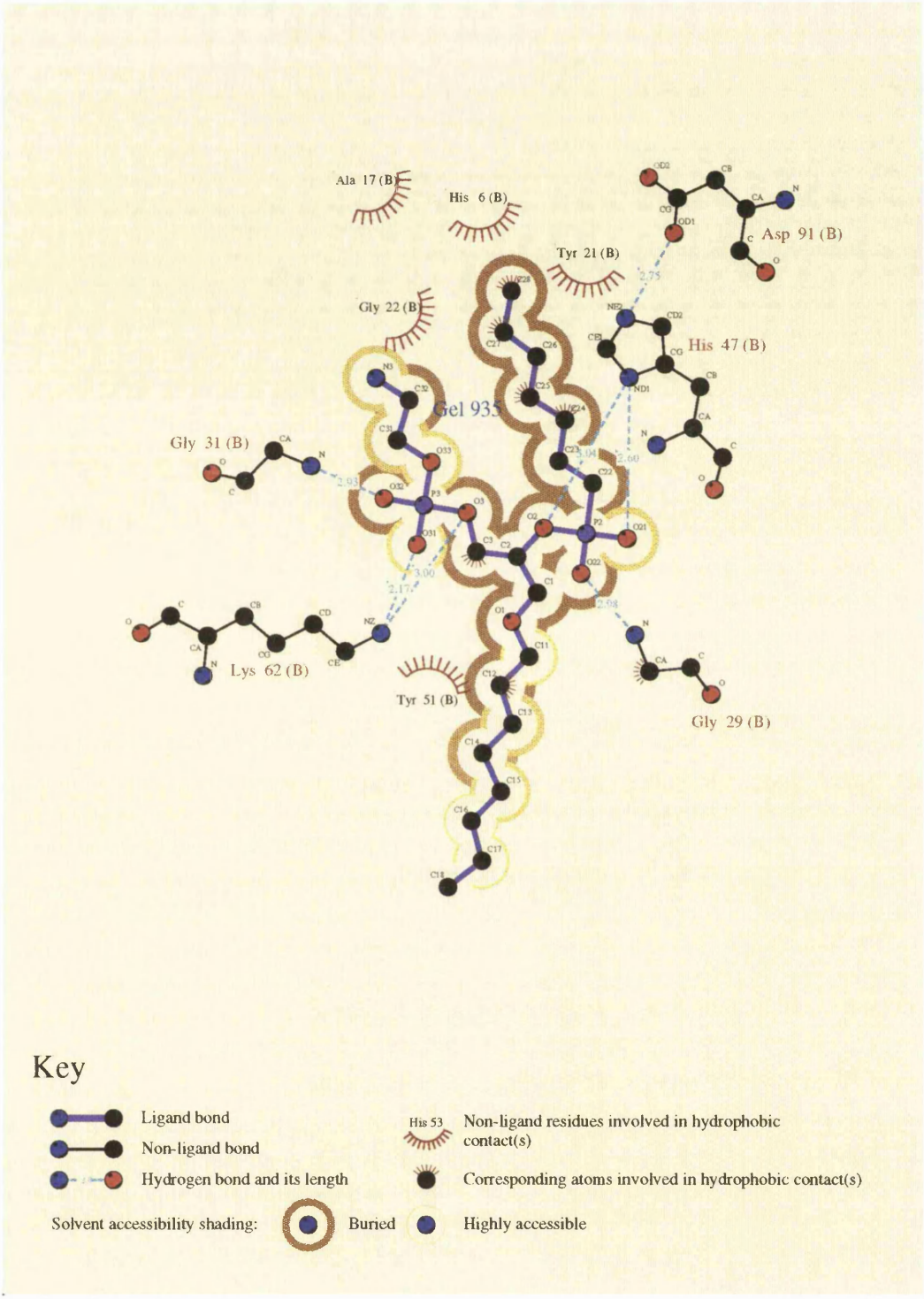


Figure 2.5: A LIGPLOT diagram of phospholipase A2 (PDB code 1poe) bound to the transition state inhibitor (residue Gel 935). The shading behind each of the ligand atoms gives a measure of their accessibility, with the darker the shade the more buried and inaccessible the atom. The key illustrates the meaning of the various symbols in the diagram; further description is given in the legend to Figure 2.3.

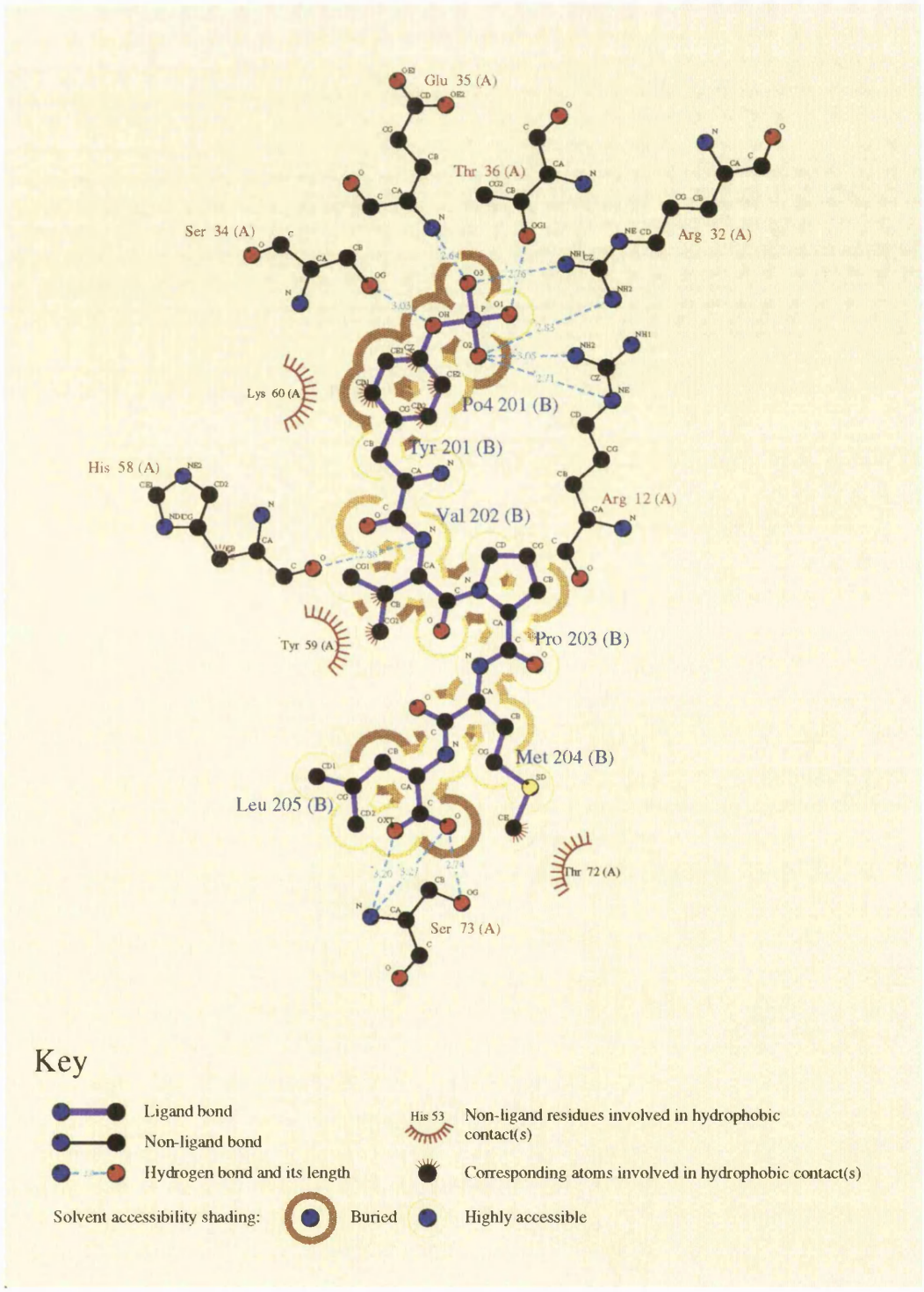


Figure 2.6: A LIGPLOT diagram of a SH2 domain-peptide complex (PDB code 1sha, ligand residues 201–205 of chain B). The peptide is a phosphotyrosine, with the phosphorylated tyrosine shown at the top of the picture with its network of hydrogen bonds to the residues of the SH2 domain of the v-*src* oncogene product (Waksman *et al.*, 1992). The accessibility shading shows the phosphate and three of its oxygens as being buried while the remainder of the peptide is largely exposed, making contact with the SH2 domain only at certain positions along its length.

tyrosine (Waksman *et al.*, 1992). The amino-aromatic interaction between Arg 12 and the ring of the tyrosine is not shown in the LIGPLOT diagram, but such interactions are of relatively minor importance (Mitchell *et al.*, 1994).

The shading in Figure 2.6 indicates the solvent accessibility of each of the peptide atoms. The darker shades correspond to atoms that are less exposed to solvent. Thus the phosphate group is largely buried while the remainder of the peptide is largely exposed, making contacts with the SH2 domain only at certain positions along its length.

Finally, Figure 2.7 shows an example of LIGPLOT's 'schematic peptide' representations based on the diagrams in Zvelebil and Thornton (1993). The diagram shows an antibody-peptide complex: the complex of Fab'B1312 with a fragment of myohaemerythrin, PDB code 2igf, ligand residues 69–75, chain P, (Stanfield *et al.*, 1990). For the peptide, only the sidechains involved in hydrogen bonding are shown while for the protein, only the atoms that make hydrogen bonds are shown, being represented only by their names and residue details. The figure corresponds closely to Figure 1*b* of Zvelebil and Thornton (1993).

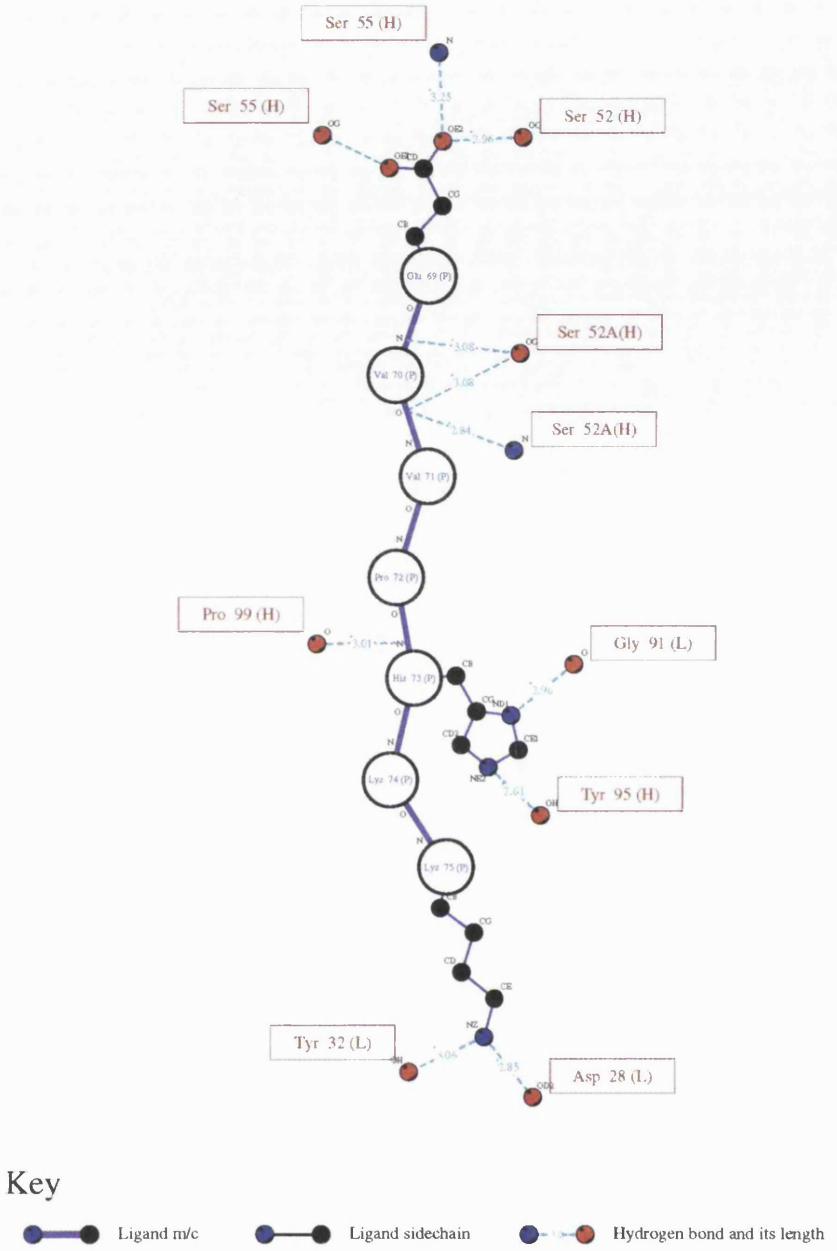


Figure 2.7: An example of a 'schematic peptide' LIGPLOT diagram. The molecule shown is the Fab'B1312-myohaemerythrin complex (PDB code 2igf) which is an antibody-peptide complex (Stanfield *et al.*, 1990) - ligand residues 69–75, chain P. Each peptide residue is shown by a circle at the C_α position, and only those sidechains which are involved in hydrogen bonds are depicted. The diagram corresponds closely to Fig.1*b* of Zvelebil and Thornton (1993) which was drawn by hand, whereas here it has been produced automatically by LIGPLOT, directly from the PDB coordinates.

2.5 References

Adobe Systems Inc. (1985) *PostScript Language Reference Manual.* Addison-Wesley, Reading,MA.

Bernstein F.C., Koetzle T.F., Williams G.J.B., Meyer E.F.Jr., Brice M.D., Rogers J.R., Kennard O., Shimanouchi T. & Tasumi,M. (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures *J. Mol. Biol.* **112** 535–542

Flores T.P., Moss D.S. & Thornton J.M. (1994) An algorithm for automatically generating protein topology cartoons *Protein Eng.* **7** 31–37

Harel M., Su C.-T., Frolow F., Silman I. & Sussman J.L. (1991) γ–chymotrypsin with its own autolysis products *Biochemistry* **30** 5217–5225

Hubbard S.J. (1991) ACCESS, computer program. Department of Biochemistry & Molecular Biology, University College, London.

Hutchinson E.G. & Thornton J.M. (1990) *hera* – a program to draw schematic diagrams of protein secondary structures *Proteins* **8** 203–212

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

McDonald I.K. & Thornton J.M. (1994) Satisfying hydrogen bonding potentials in proteins *J. Mol. Biol.* **238** 777–793

Mitchell J.B.O., Nandi C.L., McDonald I.K., Thornton J.M. & Price S.L. (1994) Amino aromatic interactions *J. Mol. Biol.* **239** 315–331

Scott D.L., Otwinowski Z., Gelb M.H. & Sigler P.B. Crystal structure of bee–venom phospholipase A_2 in a complex with a transition state analog (1990) *Science* **250** 1563–1566

Stanfield R.L., Fieser T.M., Lerner R.A. & Wilson I.A. (1990) Crystal structure of an antibody to a peptide and its complex with peptide antigen at 2.8Å.*Science* **248** 712–719

Waksman G., Kominos D., Robertson S.C., Pant N., Baltimore D., Birge R.B., Cowburn D., Hanafusa H., Mayer B.J., Overduin M., Resh M.D., Rios C.B., Silverman L. & Kuriyan J. (1992) Crystal structure of the phosphotyrosine recognition domain SH2 of V-SRC complexed with tyrosine–phosphorylated peptides *Nature* **358** 646–653

Zvelebil M.J.J.M. & Thornton J.M. (1993) Peptide–protein interactions – An overview *Quarterly Reviews of Biophysics* **26** 333–363

Chapter 3

A structural comparison of the Ser-His-Asp catalytic triads in the serine proteinases and lipases

3.1 Introduction

One of the best known functional units in enzymes is the Ser–His–Asp catalytic triad; it was first identified following the elucidation of the X–ray structure of the serine–proteinase chymotrypsin (Matthews *et al.*, 1967; Wright *et al.*, 1969; Blow *et al.*, 1969). These three residues, which occur far apart in the amino acid sequence of the enzyme, come together in a specific conformation in the active site to perform the hydrolytic cleavage of the appropriate bond in the substrate.

Serine–proteinases are a ubiquitous group of proteolytic enzymes responsible for a range of physiological responses such as the onset of blood clotting (Mann *et al*, 1987) and digestion (Blow, 1976). They also play a major role in the tissue destruction associated with arthritis, pancreatitis and pulmonary emphysema. Each enzyme is highly specific for its own peptide substrate and this specificity

is governed by the substrate residue that fits into the P_1 subsite, or specificity pocket, immediately adjacent to the scissile bond. The shape of this pocket depends upon only a few amino acids. For example, in chymotrypsin the substrate binding pocket is specific for aromatic residues, whereas in trypsin an Asp residue in its binding pocket makes it specific for the sidechains of lysine and arginine; in elastase, valine and threonine residues in the binding pocket make the enzyme specific for non-bulky uncharged residues at this position in the peptide substrate. Perona and Craik (1995) have recently published a comprehensive review of the structural basis of substrate specificity in serine proteinases.

Prior to the elucidation of the structure of chymotrypsin, several key experiments gave insight into the mechanism of action of the serine–proteinases. As described in Chapter 1, in 1946 Mazur & Bodansky found that diisopropyl fluorophosphate (DFP) irreversibly inhibits acetylcholinesterase and in 1949 Jansen and coworkers demonstrated a 1:1 stoichiometric reaction of DFP with chymotrypsin Ser 195. His 57 was implicated in the mechanism when it specifically bound tosyl-L–phenylalanine chloromethyl ketone (Schoellman & Shaw 1953). Work in Brian Hartley's laboratory (Hartley B.S. & Kilby B.A., 1954) on the hydrolysis of *p*–nitrophenylacetate suggested that there were two phases to the catalytic reaction. Gurfreund & Sturtevant (1955) performed further stopped–flow experiments with the same enzyme and substrate. They showed that the reaction could be described by a mechanism involving three distinct steps: rapid absorption of the substrate on the enzyme followed by acylation of the enzyme and finally liberation of product. In addition they implicated the Ser O^{γ} in the acylation step. The chromophoric inhibitor displacement experiments by Bernhard & Gutfreund (1965) used proflavin as a competitive inhibitor of chymotrypsin; this undergoes a large change in absorbance upon binding to the enzyme. If an ester is also mixed in the experimental solution, the proflavin will be displaced. As the

acyl–enzyme intermediate is formed all the proflavin will be displaced and the absorbance remains constant until the ester is depleted. The dissocation constant of the enzyme–substrate can be calculated from the magnitude of the initial rapid displacement.

The stopped-flow kinetic experiments decribed above detect intermediates that accumalate; steady state kinetics can detect intermediates that do not accumalate. If several substrates generate the same intermediates and its breakdown is rate determining, then they should all hydrolyse with the same value of k_{cat}. Gutfreund & Hammond (1959) compared the reaction parameters of the chymotrypsin–catalysed hydrolysis of the amide, ethyl ester and *p*–nitophenyl of tyrosine. The results, along with subsequent experiments in other laboratories, gave futher support to the three–step reaction scheme.

Epand & Wilson (1962) measured the fraction of ester converted to hydroxamic acid by incubating chymotrypsin in 10 esters of hippuric acid. All the esters produced the same ratio of hippuric acid and hippurylhydroxamic acid. In constrast, when the esters were hydrolysed by water, it results in variable product ratios. This provided good evidence for a common intermediate.

Dixon (1953) showed that deductions about the enzyme–substrate complex and ionization constants of the groups involved can be made from the effects of pH on substrate affinity. Experimentally, the most studied enzyme in this context is chymotrypsin. Bender *et al.* (1964) found that the pH dependence of k_{cat}/K_m for the hydrolysis of substrates follows a bell–shaped curve with a maximum at pH 7.8 and the reaction is dependent on two ionizable groups of pK_a 6.8 and 8.8. Subsequently, Renard & Fersht (1973) showed that the hydrolysis of acetyl-L–tryptophan *p*–nitrophenyl ester was an exception to this scheme: it followed a titration curve of pK_a 6.5 and a maximum rate constant of 3.1 x 10^7 sec^{-1} M^{-1}. It was found to be consistent with the association of enzyme and substrate being

rate determining at high pH and gives an example where steady state kinetics may be analysed to give constants for several steps in the reaction pathway.

At low pH, the k_{cat}/K_m lowers because the catalytically important base (*i.e.* His 57 in chymotrypsin, pK_a 7) becomes protonated. The combination of X–ray diffraction and solution studies on α–chymotrypsin showed that chymotrypsin existed in two conformations between pH 2 and 12. Fersht (1972) showed that the equilibrium between these two conformations is controlled by a salt bridge between Ile 16 and Asp 194 with an apparent pK_a of 9.1. At pH 9, the salt bridge is deprotonated and the conformation change in the protein occurs, rendering it catalytically inactive. These two factors explain the pH dependency of chymotrypsin's catalytic activity.

The catalytic mechanism of the serine proteinases is illustrated in Figure 3.2. The catalytic reaction proceeds by the Ser O^{γ} both donating its proton to the His imidazole ring and attacking the electrophilic carbonyl carbon of the substrate scissile bond (Figure 3.2a). This forms the first tetrahedral intermediate (Figure 3.2b) which rapidly breaks down releasing the first product, an amine in the serine proteinases or an alcohol in the lipases, as well as forming the acyl–enzyme intermediate (Figure 3.2c). A water molecule then attacks this intermediate (Figure 3.2d) forming the second tetrahedral intermediate (Figure 3.2e) and this rapidly breaks down to form product. Figure 3.1 shows a schematic diagram of the catalytic triad in chymotrypsin, PDB code *8gch* (Harel *et al*, 1991), with the tri–peptide Gly–Ala–Trp bound in the protein's active site. The general acid/base His 57 is hydrogen bonded to the substrate, and to the O^{δ_1} and O^{δ_2} of Asp 102. The nucleophilic Ser 195 O^{γ} would usually attack the electrophilic C^{α} carbonyl group on the peptide substrate, and the diagram shows Ser 195 hydrogen bonding in the vicinity of this carbonyl group. The His 57 abstracts the proton from Ser 195 O^{γ} and the Asp acts to stabilise the positive charge

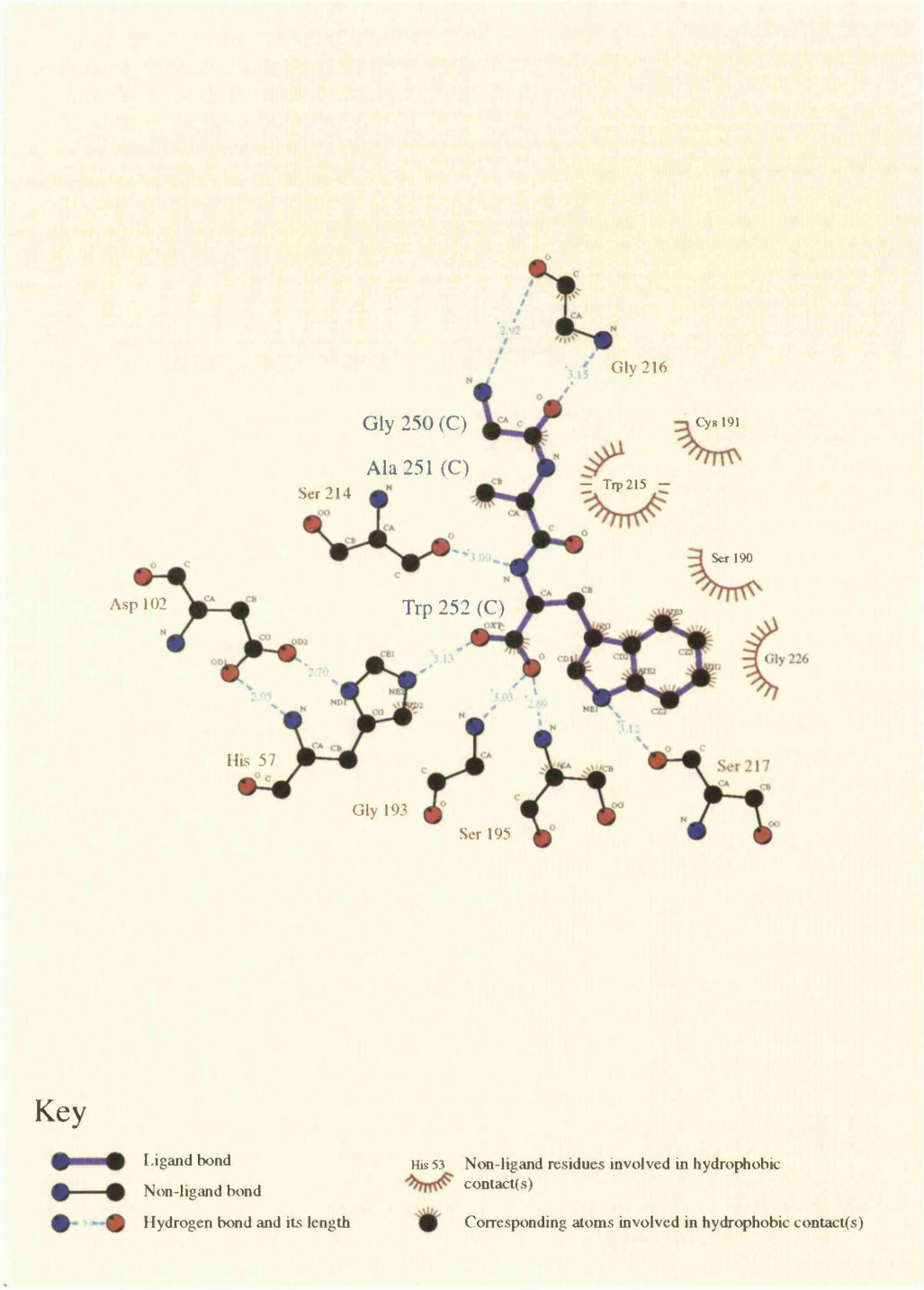


Figure 3.1: Schematic diagram of the tri–peptide Gly–Ala–Trp (residues 250–252 C) bound to the active site of chymotrypsin, 8*gch* (Harel *et al*, 1991), showing the hydrogen bonds and hydrophobic interactions the tripeptide makes with the residues of the active site. The diagram illustrates the catalytic triad of His 57, Asp 102 and Ser 195, as well as the ligand's Trp 252 residue nestling in the enzyme's hydrophobic specificity pocket.

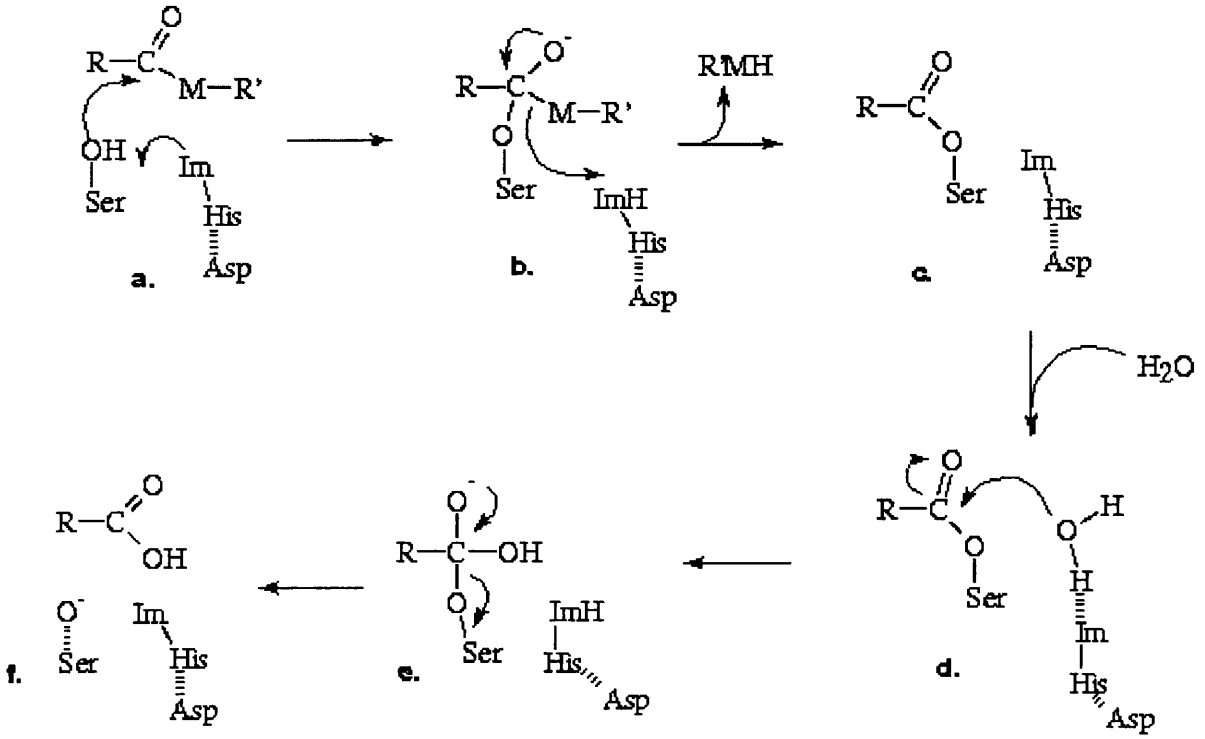


Figure 3.2: Schematic diagram representing the generalised catalytic mechanism of serine proteases and lipases. *M* is a nitrogen atom for proteases (amide bond) or an oxygen atom for the lipases (ester bond). *Im* is the imidazole sidechain of the His residue. *a.* The reaction proceeds by the His deprotonating the catalytic Ser O^{γ}. This Ser O^{γ} acts as a nucleophile, attacking the carbonyl group of the scissile bond. *b.* The first tetrahedral intermediate is formed. *c.* This rapidly breaks down to form the activated energy acyl–enzyme intermediate, releasing an amine in the serine proteases or an alcohol in the lipases. *d.* The intermediate is then hydrolysed by a water molecule. *e.* The second tetrahedral intermediate is formed. *f.* This then breaks down to form product.

on the protonated His (Polgar, 1989). The deprotonated Ser 195 nucleophile attacks the substrate and forming an acyl–enzyme intermediate. The Ser O^{γ} lies in the plane of the His and hydrogen bonds to the His $N^{\epsilon 2}$, although the angle it makes with the $N^{\epsilon 2}$ is not suitable for the formation of a strong hydrogen bond (Sawyer *et al.*, 1978). The substrate's scissile peptide carbonyl group is stabilised by hydrogen bond interactions. In elastase, for example, the carbonyl group hydrogen bonds to the two backbone nitrogens of Gly 193 and Ser 195, forming the so called 'oxyanion hole' (Kraut, 1977). Similarly, in subtilisin the equivalent of these two amido groups are Asn 155 $N^{\delta 2}$ and Ser 221 N and these are equally important in stabilising the substrate. Mutational experiments on the oxyanion hole in subtilisin have shown a 4 kcal/mol reduction in the binding energy for a mutation of Asn 155 to Ala and 2 kcal/mol for Thr 220 to Ala (Braxton & Wells, 1991). Mutation of Asn 155 in combination with the catalytic His 62 destablises the transition state to the same extent as the single mutation (Carter *et al.*, 1991). Similarly, the importance of Ser 195, His 57 and Asp 102 in chymotrypsin has been assessed by mutational studies of each residue (Corey & Craik, 1992). The binding energy decreased by 4 kcal/mol upon mutation of Asp 102 and by around 7 kcal/mol for Ser 195 or His 57. Mutation of the entire catalytic triad does not destabilise the transition state any further. This suggests that the oxyanion hole and the catalytic triad of the serine proteinases function cooperatively to enhance the catalytic rate. In the absence of the Ser–His–Asp triad, trypsin and subtilisin still achieve catalytic rates 10^4 or 10^5 greater than the uncatalysed rate. Thus, as well as the residues involved in direct chemical catalysis, interactions that contribute to binding and conformational positioning of the extended substrate in the transition state contribute to rate enhancement.

The digestive serine–proteinases, such as trypsin, are stored in the pancreas as inactive precursors and are activated by proteolysis. Trypsinogen, for example, is

converted to trypsin by removal of the N–terminal hexapeptide between residues Lys 6 and Ile 7 by enterokinase. In chymotrypsinogen, the cleavage occurs between Arg 15 and Ile 16 (Bode & Huber, 1986). This allows Ile 16 to interact with Asp 194. This electrostatic interaction causes a conformational change in the protein: Met 192 moved from a deeply buried position to the surface of the protein; residues 187 to 193 become more extended. This results in the formation of the substrate specificity pocket. In addition, the mainchain nitrogens of Gly 193 and Ser 195 move into a position to form the oxyanion hole.

The other main group of enzymes containing the Ser–His–Asp catalytic triads are the triacylglycerol lipases which are responsible for hydrolysing triglycerides into diglycerides and subsequently monoglycerides and free fatty acids. For example, pancreatic lipase hydrolyses water–insoluble triacylglycerols in the intestinal lumen and thereby plays an important role in dietary fat absorption. Lipases are stable in both aqueous and organic media and this makes them suitable as catalysts for a number of synthetic processes which would otherwise require harsh conditions to proceed. Like the serine proteinases, the catalytic mechanism is effected by way of a catalytic serine (Blow, 1990; Brady *et al.*, 1990). The catalytic site is buried beneath a short stretch of helix, known as the 'lid'. A number of crystallographic studies have confirmed the hypothesis that the lid is displaced during activation (Brzozowski *et al.*, 1991; Derewenda *et al.*, 1992), being rolled back as a rigid body into a hydrophilic trench previously filled by water molecules, exposing the active site.

In this chapter we investigate the 3D conformations of the Ser–His–Asp catalytic triads in both the serine proteinases and the lipases, using the structures deposited in the Protein Data Bank (PDB) (Bernstein *et al.*, 1977). There are around 1500 enzyme structures in the January 1995 release of the PDB, and we were able to extract 192 serine proteinases, 4 serine-type carboxypeptidases and 9

triacylglycerol lipases. We have grouped these 205 proteins into classes firstly by making structure and sequence comparisons, and secondly by the functional classification given by the protein's Enzyme Classification, or E.C. number (Bielka *et al.*, 1992).

There have already been studies that identify recurring non–sequential motifs in protein structures, such as the algorithm by Fischer *et al.*, (1994). Barth *et al* (1993, 1994) have classified the Ser–His–Asp triad serine proteinases according to the chymotrypsin and subtilisin families and suggest a new catalytic mechanism based on the differences between tonin and kallikrein. In addition, they discuss the catalytic implications of the conserved non–catalytic Ser 214 residue that lies in the vicinity of the Ser–His–Asp catalytic triad.

Our aim is to extract and compare all the available Ser–His–Asp conformations and see how they are conserved or differ across the different fold types and functional classes. This should identify the most important aspects of the triad's conformation and how the different enzymes go about achieving it. We find that the orientation of the Asp and Ser sidechain atoms differ quite considerably between the various fold groups and it appears that only the positions of the Asp carboxyl oxygen hydrogen–bonded to the His $N^{\delta 1}$, and of the Ser $O\gamma$ hydrogen–bonded to the His $N^{\epsilon 2}$ are critical.

In addition, we obtain a template defining the Ser–His–Asp catalytic conformation and use this template to search for similar triads in other proteins, including non-enzymes, to see how often they occur outside the serine proteinases and lipases. We found two examples of non–enzyme 'catalytic' triads but, upon inspection, these triads appear to be sterically hindered by surrounding hydrophobic residues or are in an unsuitable position in the protein molecule to perform catalysis. Therefore, to date, the catalytically active form of the Ser–His–Asp triad, as defined by the positions of the His sidechain, the Asp carboxyl oxygen

and the Ser Oγ atoms, only occurs in the serine proteinases and lipases.

3.2 Methods

3.2.1 The datasets

Two datasets were used, both extracted from the January 1995 release of the PDB. The first comprised the serine proteinases and lipases and was used to study the structural similarites of the catalytically active forms of the Ser-His-Asp triad. These enzymes were extracted from the PDB by first cross–referencing every structure's sequence against SWISS–PROT (Bairoch and Boeckmann, 1994; March 1995 release). This is a sequence database which enables the accurate identification of the E.C. number (Bielka *et al.*, 1992) of every enzyme structure in the PDB.

The dataset comprised 192 serine proteinases, 4 serine–type carboxypeptidases and 9 triacylglycerol lipases. Since each enzyme can have more than 1 chain and therefore more than 1 catalytic triad, we in fact had a dataset of 205 serine proteinase chains, 7 serine–type carboxypeptidases and 13 lipases. The enzymes were first grouped into families according to sequence similarity, the enzymes in each group having a sequence identity of more than 30% with at least one other member of the group. To identify more remote homologues we further classified the enzymes according to their structural similarity because the structure of an enzyme will reflect its evolutionary origin and this may influence the conformation of the catalytic triad. The structural classification was achieved using the program SSAP (Orengo *et al.*, 1993) which computes a similarity score between two proteins (SSAP score) between 0 and 100; the higher the score the more similar the overall structures. We used a SSAP score of > 80 to group together the enzymes having similar overall folds. This is the minimum score

generally used to identify homologues.

Table 3.1 shows the dataset used, classified according to the four different fold groups and E.C. subgroupings. The serine proteinases come in two distinct folds: a β-sandwich fold, characterised by trypsin, and an alternating α/β fold, characterised by subtilisin. These make up Groups 1 and 2 in Table 3.1, respectively. Groups 1.*a*, *b* and *c* have lower than 30% sequence identity but SSAP score > 80, which indicates that they have very similar overall structures and are almost certainly derived from a common ancestor. Figure 3.3 shows a 3D representation of the Group 1 β-sandwich structure of chymotrypsin, 1*cho* (Fujinaga *et al.*, 1987), with the Ser–His–Asp catalytic triad lying in the binding groove of the enzyme. Group 2 consists of serine–proteinases having an alternating α/β structure with doubly–wound topology as shown in Figure 3.4 for subtilisin, 2*sic* (Takeuchi *et al.*, 1991). The enzyme has a central core of a seven–stranded parallel β–sheet and nine α–helices which are packed, mainly antiparallel, against the sheet. Group 3 contains the serine-type carboxypeptidases; its overall fold is also an alternating α/β structure (Figure 3.5), but differs from that of Group 2 in that it consists of an 11-stranded β–sheet surrounded by 15 helices with different connectivity. The Ser–His–Asp catalytic triad is buried in a deep bowl-like depression in the enzyme surface. Finally, Group 4 contains all the triacylglycerol lipases. Again, the overall fold is α/β with 6 helices surrounding a 5 to 11 stranded β–sheet (Figure 3.6). The SSAP scores between members of this Group and members of Groups 2 and 3 are < 70, reflecting the structural differences.

The second of our two datasets was used as a representative set of protein structures from the PDB, used for searching for possible occurrences of the Ser-His-Asp triads in the catalytic conformation, in protein structures in general. In this dataset we wished to include all unique protein chains, including homologues, but excluding identical or trivially different chains such as single-residue mutants.

GROUP 1: serine proteases
β–sandwich - trypsin-like fold

1.*a*

chymotrypsin E.C.3.4.21.1

1acb E	1cgi E	1cgj E	2cga A	2cga B	1chg	1cho E	2cha	4cha A	4cha B	5cha A	5cha B
6cha A	6cha B	1gcd	1gct A	2gch	2gct A	3gch	3gct A	4gch	5gch	6gch	7gch
8gch	1gha E	1ghb E	1gmc A	1gmd A	1gmd B	1gmh	2gmt				

trypsin E.C.3.4.21.4

1bit	1bra	1brb E	1brc E	1gbt	1mct A	1ntp	1ppc E	1ppe E	1pph E	2ptc E	2ptn
3ptb	3ptn	4ptp	1sgt	1smf E	1tab E	2tbs	1tgb	1tgc	1tgn	1tgs Z	1tgt
2tga	2tgp Z	2tgt	1tld	1tng	1tnh	1tni	1tnj	1tnk	1tnl	1tpa E	1tpo
1tpp	3tpi Z	4tpi Z									

thrombin E.C.3.4.21.5

1abi H	1abj H	1bbr H	1bbr K	1bbr N	1dwb H	1dwc H	1dwd H	1dwe H	1etr H	1ets H	1ett H
1fph H	1hag E	1hah H	1hai H	2hat H	1hgt H	2hgt H	1hlt H	1hlt K	2hnt E	2hpp H	2hpq H
1hrt H	4htc H	1hut H	1ihs H	1iht H	1nrn H	1nrn R	1nro H	1nro R	1nrp H	1nrp R	1nrq H
1nrr H	1nrs H	1ppb H	1thr H	1ths H	1tmb H	1tmt H	1tmu H				

tissue kallikrein E.C.3.4.21.35

2kai A	2kai B	2pka B	2pka Y	1ton

pancreatic elastase E.C.3.4.21.36

1ela A	1elb A	1elc A	1esa	1esb	1est	2est E	3est	4est E	5est E	6est	7est E
8est E	9est	1inc	1jim								

leukocyte elastase E.C.3.4.21.37

1hne E	1ppf E	1ppg E

1.*b*

α–lytic protease E.C.3.4.21.12

2alp	1lpr A	2lpr A	3lpr A	4lpr A	5lpr A	6lpr A	7lpr A	8lpr A	9lpr A	1p01 A	1p02 A
1p03 A	1p04 A	1p05 A	1p06 A	1p08 A	1p09 A	2p07	1p10 A	1p11 E	1p12 E		

streptogrisin A E.C.3.4.21.80

1sgc	2sga	3sga E	4sga E	5sga E

streptogrisin B E.C.3.4.21.81

3sgb E	4sgb E

1.*c*

1.3 lysyl endopeptidase E.C.3.4.21.50

1arb	1arc

GROUP 2: serine proteases
alternating α/β - subtilisin-like fold

subtilisin E.C.3.4.21.62

1cse E	1mee A	1s01	1s02	1sbc	1sbn E	1sbt	2sbt	1sca	1scb	1scd	1scn E
1sel A	1sel B	2sec E	1sib E	2sic E	3sic E	5sic E	2sni E	1st2	1st3	2st1	1sub
1suc	1sud										

endopeptidase K E.C.3.4.21.64

1pek E	2pkc	2prk	3prk E	1ptk

thermitase E.C.3.4.21.66

1tec E	2tec E	3tec E	1thm

GROUP 3: serine-type carboxypeptidase
alternating α/β

serine type carboxypeptidase E.C.3.4.16.5

3sc2 A	3sc2 B	1whs A	1whs B	1ysc

GROUP 4: triacylglycerol lipase
α/β

triacylglycerol lipase E.C.3.1.1.3

1crl	1hpl A	1hpl B	1tah B	1tah A	1tah C	1tah D	1tgl	3tgl	4tgl	5tgl	1thg
1trh											

Table 3.1: Dataset of enzymes containing the Ser–His–Asp catalytic triad.

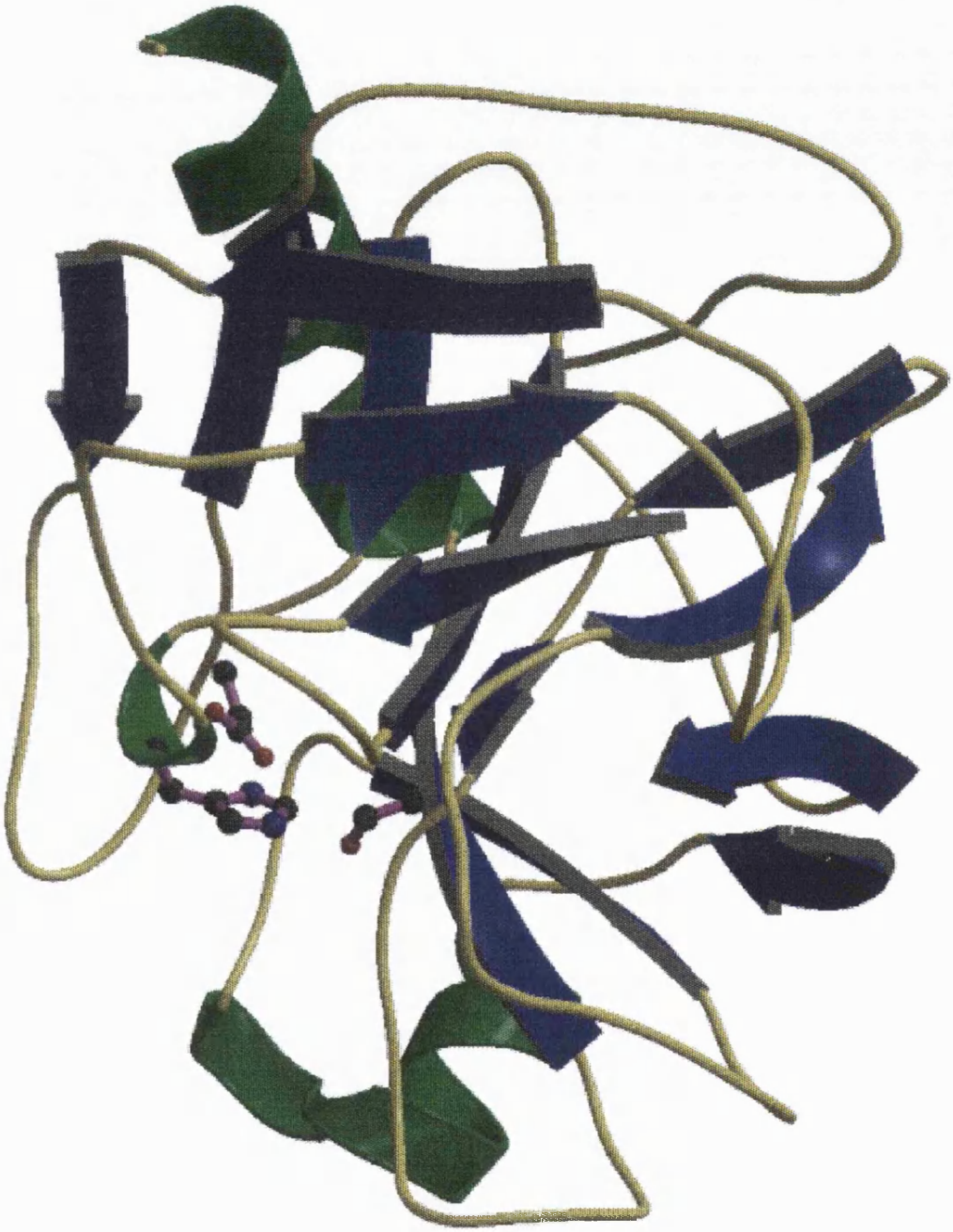


Figure 3.3: MOLSCRIPT (Kraulis, 1991) diagram of Group 1 proteins: The β-sandwich structure of chymotrypsin, 1*cho* (Fujinaga *et al.*, 1987).

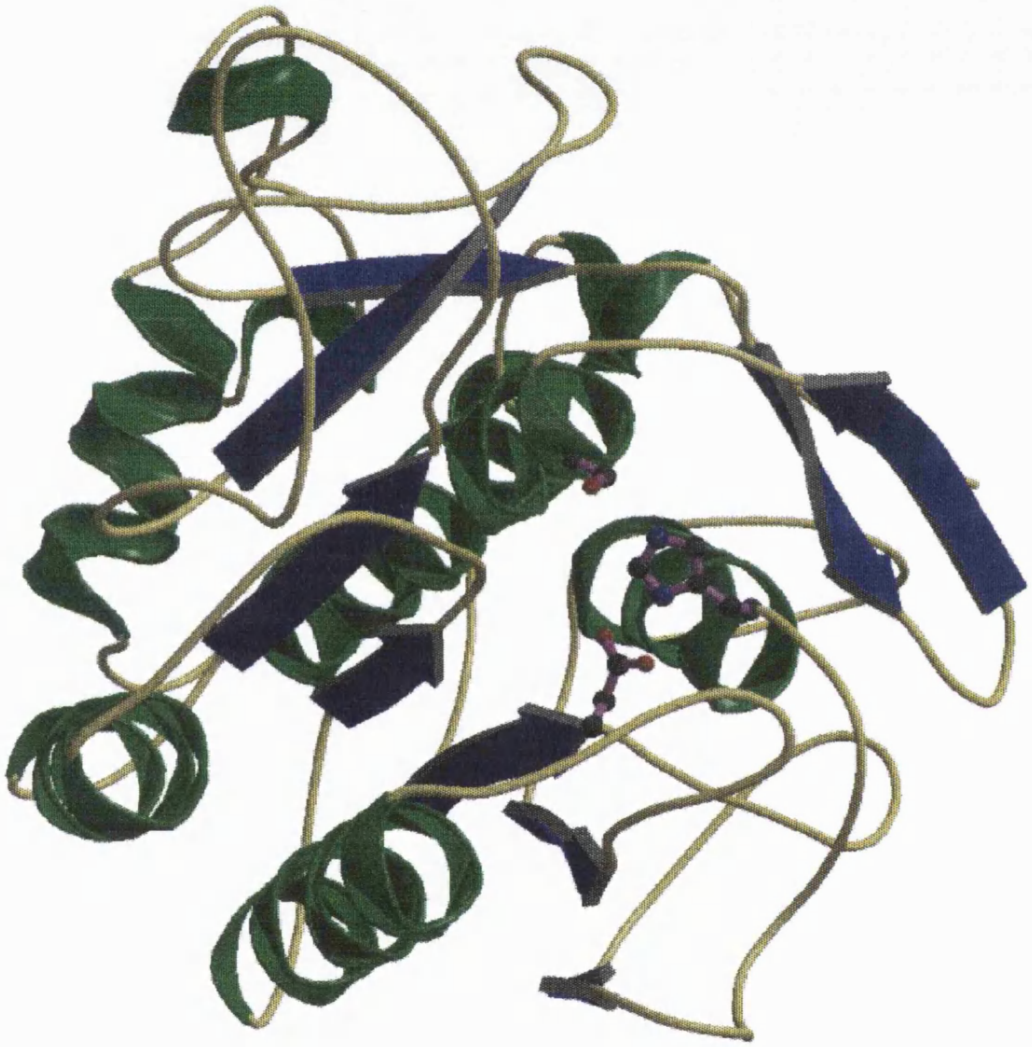


Figure 3.4: MOLSCRIPT (Kraulis, 1991) diagram Group 2 proteins: the doubly-wound α/β structure of subtilisin 1*s*01 (Pantoliano *et al*, 1989).

Figure 3.5: MOLSCRIPT (Kraulis, 1991) diagram of Group 3: The α/β structure of serine-type carboxypeptidase 3*sc*2 (Liao *et al*, 1992).

Figure 3.6: MOLSCRIPT (Kraulis, 1991) diagram of Group 4: The α/β structure of lipase *4tgl* (Derewenda *et al*, 1992).

The protein chains were again extracted from the PDB but this time so that no two had a sequence identity greater than 95%. The resultant 639 protein chains are listed in Table 3.2.

3.2.2 Extraction of catalytic triads

The Ser–His–Asp catalytic triads were automatically extracted from the enzyme dataset by first locating all interacting triplets of Ser, His and Asp residues using a program called DISTRIB (R.A.L). Residues were considered to be interacting if at least one interatomic contact is less than the sum of the van der Waals radii of the two atoms plus 1Å. The triplets extracted were transformed onto a common reference frame defined by the planar ring of the His. Each His was placed in the x–y plane with its C^{γ} at the origin, its C^{β} on the negative y–axis and its N^{δ_1} atom with positive x and y values.

After extracting all the Ser, His, Asp interacting triplets, we wished to filter out the catalytic triads from the ordinary non–catalytic associations. The goal was to derive a 3D consensus template which would allow the automated identification of all catalytic Ser–His–Asp triplets with the exclusion of all other triads in the PDB. This filtering process involved identifying those triplets where the Asp and Ser residues were in approximately the correct positions relative to the His.

However, rather than merely use a simple distance cut-off to achieve this filtering, we aimed to be more specifically selective towards the triads. We did this by iteratively calculating a mean position for the atoms in the Asp and Ser. We evolved a rather complex procedure for defining the consensus templates first within each homologous family of enzymes in the dataset, and then over all catalytic triads. Initially we used all atoms in the Asp and Ser sidechains but it soon became apparent that only the position of the catalytic Ser $O\gamma$, and

95% by sequence non-homologous dataset									
119l	135l	155c	1aaf	1aaj	1aak	1aap A	1aat	1ab2	1aba
1abk	1abm A	1abt A	1aca	1ace	1aco	1acp	1acx	1adn	1adr
1ads	1aec	1aep	1afc A	1ahd P	1ain	1aiz A	1ak3 A	1ake A	1ala
1alb	1alc	1ald	1alk A	1aoz A	1apa	1apm E	1apo	1aps	1arb
1arp	1arq A	1atn A	1atx	1ave A	1avh A	1ayh	1baf H	1bal	1bbh A
1bbi	1bbo	1bbp A	1bbt 1	1bds	1bgc	1bge A	1bgh	1bha	1bia
1bll E	1bmv 1	1bod	1bov A	1brn L	1bsr A	1btc	1bus	1bw3	1c2r A
1c5a	1caa	1cau A	1cb1	1cbn	1cc5	1ccd	1ccr	1cd8	1cdb
1cde	1cdg	1cdt A	1cew I	1cgi I	1cgt	1chb D	1chr A	1cid	1cll
1cmb A	1cob A	1col A	1cor	1coy	1cpb	1cpc A	1cpt	1crl	1csc
1cse E	1ctf	1cth A	1cvo	1cy3	1d66 A	1dbb H	1dfb H	1dhr	1drf
1dri	1dtk	1dtx	1dxi A	1eaf	1eca	1ede	1egf	1ego	1end
1etr H	1ezm	1f3g	1fas	1fba A	1fc2 C	1fcb A	1fdh G	1fdl H	1fdx
1fgv H	1fha	1fia A	1fkb	1flv	1fnr	1frr A	1fus	1fvc A	1fvd A
1fxa A	1fxd	1fxi A	1gal	1gat A	1gb1	1gca	1gct A	1gd1 O	1gdh A
1gf1	1gf2	1ggb H	1ghl A	1gky	1gla G	1glt	1glu A	1gly	1gmf A
1gmp A	1gof	1gox	1gp1 A	1gpb	1gpr	1gps	1gpt	1gsr A	1gss A
1guh A	1hbg	1hbq	1hcc	1hdd C	1hds A	1hdz A	1hem	1hev	1hfh
1hge A	1hhl	1hil A	1hip	1hiv A	1hle A	1hmy	1hna	1hne E	1hoe
1hra	1hrh A	1hsa A	1hsb A	1hsp	1hst A	1huw	1hyp	1il1b	1ifc
1igf H	1ind H	1ipd	1isu A	1ith A	1jhl H	1kdu	1kst	1lab	1lcc A
1lct	1ldn A	1lec	1len A	1lfb	1lfi	1lga A	1lh1	1lis	1lla
1llc	1lld A	1lmb 3	1lpe	1lpf A	1ltb C	1lte	1lts A	1lvl	1lya A
1lz1	1maj	1mam H	1mat	1mba	1mbd	1mbs	1mcp H	1mct A	1mda H
1mdc	1mee A	1mfa H	1mfa L	1min A	1mio A	1mpp	1mup	1myg A	1myp A
1myt	1nar	1nbt A	1nbv H	1nca H	1ndk	1nea	1nip A	1noa	1nor
1npc	1npx	1nrc A	1nrd	1nsc A	1ntx	1nxb	1ofv	1oma	1omf
1onc	1opa A	1osa	1ova A	1ovb	1paf A	1pal	1paz	1pba	1pbx A
1pca	1pda	1pdc	1pdg A	1pfk A	1pgd	1pgx	1pha	1phh	1pho
1pi2	1pii	1pk4	1pkp	1pkr	1plc	1pnj	1poa	1poc	1pod
1poh	1pox A	1pp2 L	1ppa	1ppb H	1ppl E	1ppn	1ppo	1prc C	1ptf
1pya A	1pyp	1r09 1	1r69	1rai A	1rbp	1rcb	1rdg	1rds	1rec
1rei A	1rfb A	1rhd	1rhg A	1rib A	1ril	1rip	1rne	1rop A	1rro
1rtc	1rtp 1	1rve A	1s01	1sbp	1sdy A	1sgt	1sh1	1sha A	1shf A
1shg	1shp	1sim	1siv A	1slt A	1smr A	1sos A	1spa	1srd A	1sry A
1st3	1stf I	1stp	1sub	1tab I	1tbs	1ten	1tet H	1tfd	1tfi
1tgl	1tgs I	1thb A	1thg	1thm	1tie	1tim A	1tlk	1tme 1	1tml
1tnc	1tnf A	1ton	1top	1tpk A	1tpl A	1tpm	1trb	1tre A	1trm A
1tta A	1ttf	1ubq	1ula	1utg	1vaa A	1vab B	1vil	1vna	1vsg A
1wsy A	1xim A	1xis	1xla A	1yat	1ycc	1yea	1yeb	1ymb	1ypc I
1ypi A	1ysa C	1zaa C	256b A	2aaa	2aai B	2abx A	2ach A	2act	2alp
2apr	2atc A	2bat	2bb2	2bbk H	2bjl 1	2bop A	2bpa 1	2cab	2cas
2cba	2ccx	2ccy A	2cdv	2cmd	2cna	2cpl	2cro	2ctc	2cts
2ctv A	2ctx	2cyp	2dnj A	2ech	2er7 E	2fb4 H	2fbj H	2fcr	2fx2
2fxb	2gbp	2gcr	2gst A	2hhm A	2hhr A	2hip A	2hmb	2hmq A	2hpd A
2hpr	2ig2 H	2igg	2ihl	2imn	2ldx	2lhb	2ltn A	2mad H	2mcg 1
2mcm	2mev 1	2mhb A	2mhr	2mip A	2mm1	2mnr	2msb A	2mta C	2nck L
2nn9	2ohx A	2ovo	2pcb B	2pf1	2pia	2pka A	2pkc	2plt	2plv 1
2pmg A	2pna	2pol A	2reb	2rhe	2rn2	2rsp A	2sas	2sga	2sic I
2sn3	2sns	2snv	2stv	2tbv A	2tgf	2tgi	2tmd A	2tmn E	2tmv P
2tpr A	2trx A	2ts1	2tsc A	2uce	2wrp R	2yhx	2yhx	351c	3adk
3b5c	3bcl	3blm	3c2c	3cd4	3chy	3cla	3cms	3dfr	3eca A
3est	3fxc	3gap A	3grs	3hfm H	3il8 A	3ink C	3lad A	3ldh	3mds A
3mon A	3ovo	3p2p A	3pfk	3pgk	3pgm	3psg	3rp2 A	3rub L	3sc2 A
3sdh A	3sdp A	3sgb E	3trx	3xia	4azu A	4bp2	4cpv	4dfr A	4enl
4fab H	4fgf	4fxn	4gcr	4gpd 1	4hvp A	4icb	4mdh A	4mt2	4ptp
4rcr H	4sbv A	4sgb I	4tms	5cyt R	5fbp A	5fd1	5ldh	5p21	5pal
5pti	5rub A	5tim A	6ins E	6ldh	6rxn	6taa	7aat A	7api A	7cat A
7fab H	7icd	7pcy	7rsa	8abp	8dfr	8fab A	8i1b	8rub L	8rxn A
9ldt A	9pcy	9rnt	9wga A						

Table 3.2: Non–identical dataset of enzyme and non–enzyme proteins, where no two proteins have a sequence identity greater than 95%.

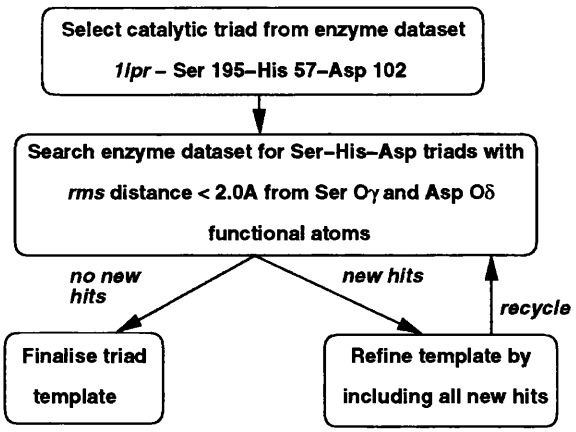


Figure 3.7: A flow diagram showing the main steps involved in the calculation of the 3D template triad.

whichever of the Asp carboxyl oxygens is hydrogen bonded to the His ring were conserved.

The first step in this procedure, which is summarised in Figure 3.7, was to select one of the known catalytic triads as a starting-point. The one chosen was Ser 195–His 57–Asp 102 from α–lytic proteinase, *1lpr* (Bone *et al.*, 1991*a*). The relative positions of its two functional oxygens, Asp $O^{\delta 2}$ and Ser O^{γ}, were taken as reference points. When a given Ser–His–Asp triad was transformed so that its histidine was superimposed on the reference His, the root mean square (*rms*) distance of its closest Asp carboxyl oxygen and its Ser O^{γ} to the reference atoms was calculated, the smaller the *rms* distance value the closer the triplet to the reference triplet from *1lpr*. This provides a means of filtering out the catalytic triads from the general Ser–His–Asp associations. However, to avoid bias caused by the initial choice of the reference enzyme, a procedure was used whereby, having filtered out the catalytic triads, they were used to calculate a mean position for both the two functional oxygens. Initially, a separate mean

was computed for each of the four structural groups shown in Table 3.1. These means were averaged to give an overall mean 3D consensus template. Using this template as the new starting point, we were able to calculate mean templates for the catalytic triads in each enzyme and fold group; first for just the two functional oxygens, as before, and secondly for all sidechain atoms of the Asp and Ser (namely Ser C^{α}, Ser C^{β}, Ser O^{γ}, Asp C^{α}, Asp C^{β}, Asp O^{δ_1} and Asp O^{δ_2}). These templates shall be referred to as 'functional' and 'sidechain' templates respectively.

The calculation of the mean coordinates of the sidechain atoms was slightly complicated by the Asp having two carboxyl oxygens, either one of which, O^{δ_1} or O^{δ_2}, might be the functional one; their names being defined solely by the appropriate torsion angle. Thus, in calculating the 3D mean consensus template, we identified which Asp carboxyl oxygen is hydrogen bonded to the His imidazole ring and took this to be the functional oxygen; the other oxygen was considered the non–functional one. In computing the overall *rms* distance, the distances between corresponding functional and non–functional oxygens were used. In some cases this procedure can artificially increase the *rms* distance value. For example, consider the case where the two functional oxygens of equivalent Asps coincide but their non–functional oxygens are on opposite sides of this oxygen. Our *rms* distance will be larger than the value a standard *rms* distance comparison would give. However, this has the advantage of allowing those triads with unusual sidechain conformations to be more easily identified.

The overall 3D consensus template was used on the second dataset to locate conformations resembling the Ser–His–Asp catalytic triads in proteins other than the lipases and serine proteinases. The results are discussed below.

3.3 Results

3.3.1 Conformations of the catalytic Asp and Ser sidechains

Figure 3.8 shows a representative Ser–His–Asp triad from each of the four structural groups in Table 3.1, and one can see that there are quite marked differences between them. These differences can be quantified by comparing the template triads. Table 3.3 shows the *rms* distances between both the 'functional' and 'sidechain' templates of the four groups. The *rms* deviation of each of the 'functional' fold group templates from the combined template is between 0.39Å and 0.65Å, indicating a high degree of structural conservation. However, the 'sidechain' forms of the same templates have *rms* values varying from 1.49Å to 3.27Å, indicating that the sidechain atoms originate from different orientations across the four fold groups. In contrast, as expected, the catalytic triad conformation is more conserved within each structural group; the 'functional' template mean *rms* deviations vary from 0.45Å to 0.65Å whereas the 'sidechain' mean *rms* deviations are from 0.67Å to 1.06Å.

Comparison of the four structural group triads indicates that the 'sidechain' templates of fold groups 3 and 4 are very similar (*rms* 0.87Å), whereas the sequence identity between these two groups is low at 11%. Groups 1 and 2 are also reasonably similar (*rms* 1.33Å), but the sequence identity between these two groups is also low at 16%. The *rms* distance values increase to around 2Å when either of the first two groups are compared to the last two groups. The Ser sidechain atoms of Groups 1 and 2 originate from below the plane of the His whereas those of Groups 3 and 4 come from above. There are also differences in the conformations of the Asp sidechains for each of the 4 Groups, most noticeably in the subtilisins (Group 2), which have a different oxygen as the functional one.

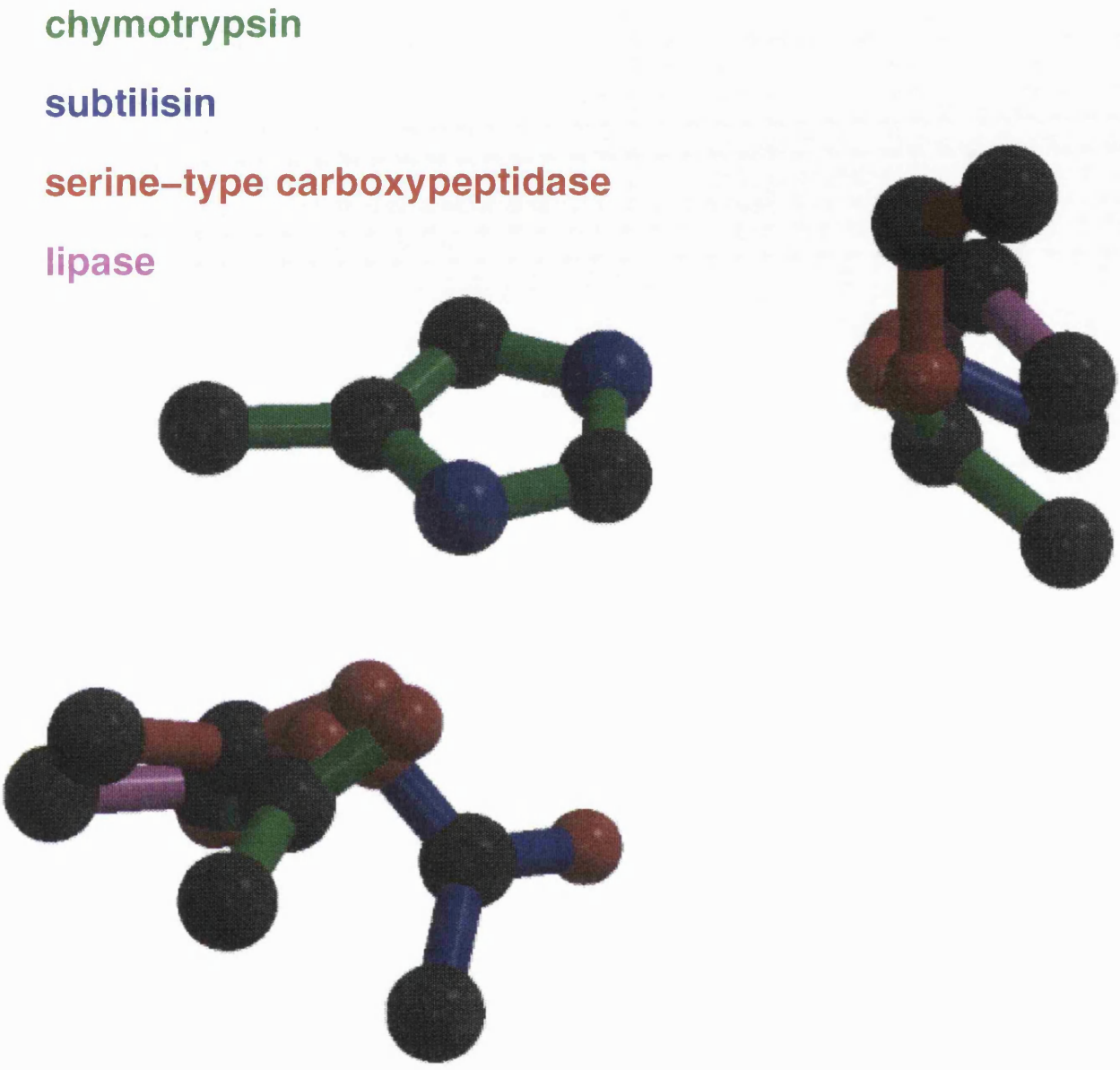


Figure 3.8: Conformations of representative catalytic triads from each of the 4 fold groups: chymotrypsin 1*cho* (Fujinaga *et al.*, 1987), subtilsin 2*sic* (Takeuchi *et al.*, 1991), serine–type carboxypeptidase 3*sc*2 (Liao *et al*, 1992) and lipase 1*tah* (Noble *et al.*, 1993), showing the different conformations adopted by the Ser and Asp sidechains. The triads have all been superimposed on their histidine residue. Diagrams produced using Raster3d (Bacon and Anderson, 1988; Merritt and Murphy, 1994)

	Number chains	Number catalytic triads	Mean *rms* distance of group	**Combined template**	**Fold group 1**	**Fold group 2**	**Fold group 3**	**Fold group 4**
Combined template:								
sidechain	225	195	1.32	0.00	1.49	3.27	2.39	1.59
functional			0.77	0.00	0.47	0.17	0.65	0.39
Fold gp. 1								
sidechain	170	152	0.67	1.49	0.00	1.33	2.23	1.74
functional			0.62	0.47	0.00	0.34	1.04	0.71
Fold gp. 2								
sidechain	35	29	0.70	3.27	1.33	0.00	2.76	2.30
functional			0.58	0.17	0.34	0.00	0.80	0.42
Fold gp. 3								
sidechain	7	4	0.83	2.39	2.23	2.76	0.00	0.87
functional			0.65	0.85	1.04	0.80	0.00	0.87
Fold gp. 4								
sidechain	13	10	1.06	1.59	1.74	2.30	0.87	0.00
functional			0.45	0.39	0.71	0.42	0.87	0.00

Table 3.3: Comparison of the consensus triad template derived for each fold group individually and also combined to give the mean triad. *Rms* distances are given for each fold group triad against all others for all sidechain atoms of the catalytic Asp and Ser and 'functional' atoms Asp O^{δ_1} and Ser O^{γ}. 'Number chains' are the total number of chains in the enzyme dataset. 'Number catalytic triads' is the number of catalytic triads identified in the enzyme dataset The discrepancy between number of chains and number of triads is explained in the text. 'Combined template' are the mean coordinates of the four structural group triads. 'Mean *rms* deviation of group' is the mean *rms* deviation of each of the sub-group members from their respective mean catalytic triads.

Since no two fold groups share more than 16% sequence identity between them, there is no clear link between sequence identity and catalytic triad conformation. The structural aspect of each group's catalytic triad will be discussed below.

Group 1 - β-sandwich trypsin-like fold

152 catalytic triads identified from 170 chains. Mean *rms* deviation from Group 1 template: 'functional' 0.62Å, 'sidechain' 0.67Å.

A 3D representation of the catalytic triad of a Group 1 enzyme, elastase *4est* (Takahashi *et al.*, 1989) is shown in Figure 3.9. The Asp 102 $O^{\delta 2}$ is hydrogen bonded to His $N^{\delta 1}$ and in the same plane as the His 57 imidazole ring, while its $O^{\delta 1}$ atom is hydrogen bonded to the mainchain nitrogen of the His. The Ser 195 O^{γ} is hydrogen–bonded to the His $N^{\epsilon 2}$. It lies slightly below the the plane of the His imidazole ring.

The Group 1 enzymes have been divided into three sub-groups (*1.a*, *1.b*, *1.c*) according to sequence identity, each sub–group member having $< 30\%$ sequence identity with the other sub–groups (Table 3.1). To show the structural conservation of the catalytic triads in this group, mean catalytic triads were calculated for each of the enzyme groups and the *rms* deviation of each of these mean templates from each other enzyme triad was calculated. For example, Figures 3.10 and 3.11 show a 3D representation of one triad from each of the 3 subgroups, *1.a* chymotrypsin, *1cho* (Fujinaga *et al.*, 1987), *1.b* α–lytic proteinase (*1lpr*) and *1.c* lysyl endopeptidase, *1arb* (Tsunasawa S. *et al.*, 1989) showing the strong structural similarity of these 3 sub–group triads. Indeed, the sidechain template of chymotrypsin is found to be only 0.35Å and 0.53Å respectively from the templates of subgroup *1.b* α–lytic proteinase and *1.c* lysyl endopeptidase.

In some of the structures in our dataset, the catalytic triads are distorted by the presence of an inhibitor in the active site. In Group 1 there are 18 such

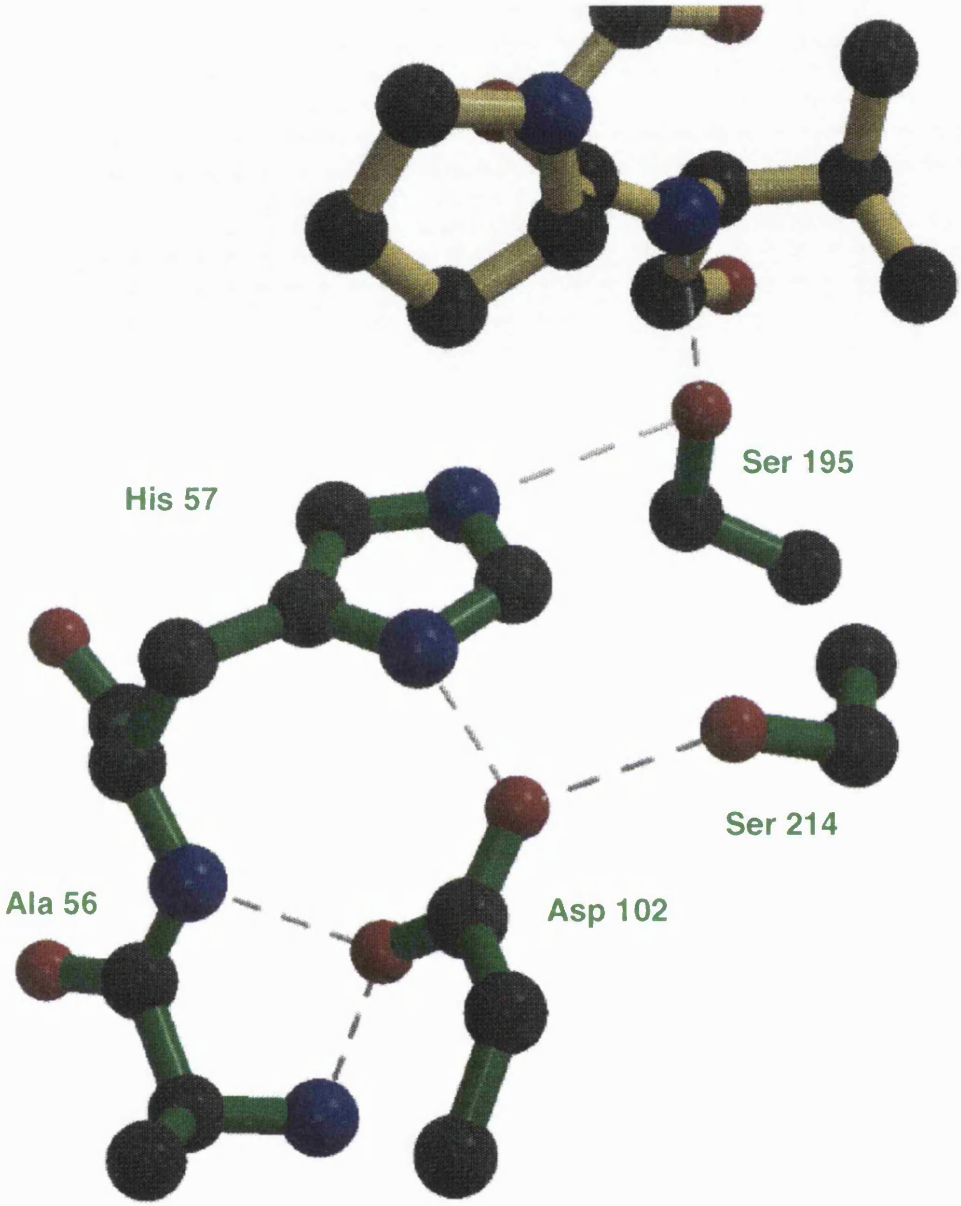


Figure 3.9: The catalytic triad of elastase 4*est* (Takahashi *et al.*, 1989) and its inhibitor, a modified tri-peptide. The diagram shows the hydrogen bond interaction of Asp O^{δ_1} with His N^{δ_1} and Ser $O\gamma$ with N^{ϵ_2}. In addition, the Ser $O\gamma$ is hydrogen bonded to the mainchain of the peptide inhibitor and this would be near the site of cleavage in the actual substrate. Ser 214 is found in a structurally conserved position in fold Group 1 enzymes. The figure also shows the non-catalytic Asp oxygen hydrogen bonding to the backbone of His 57.

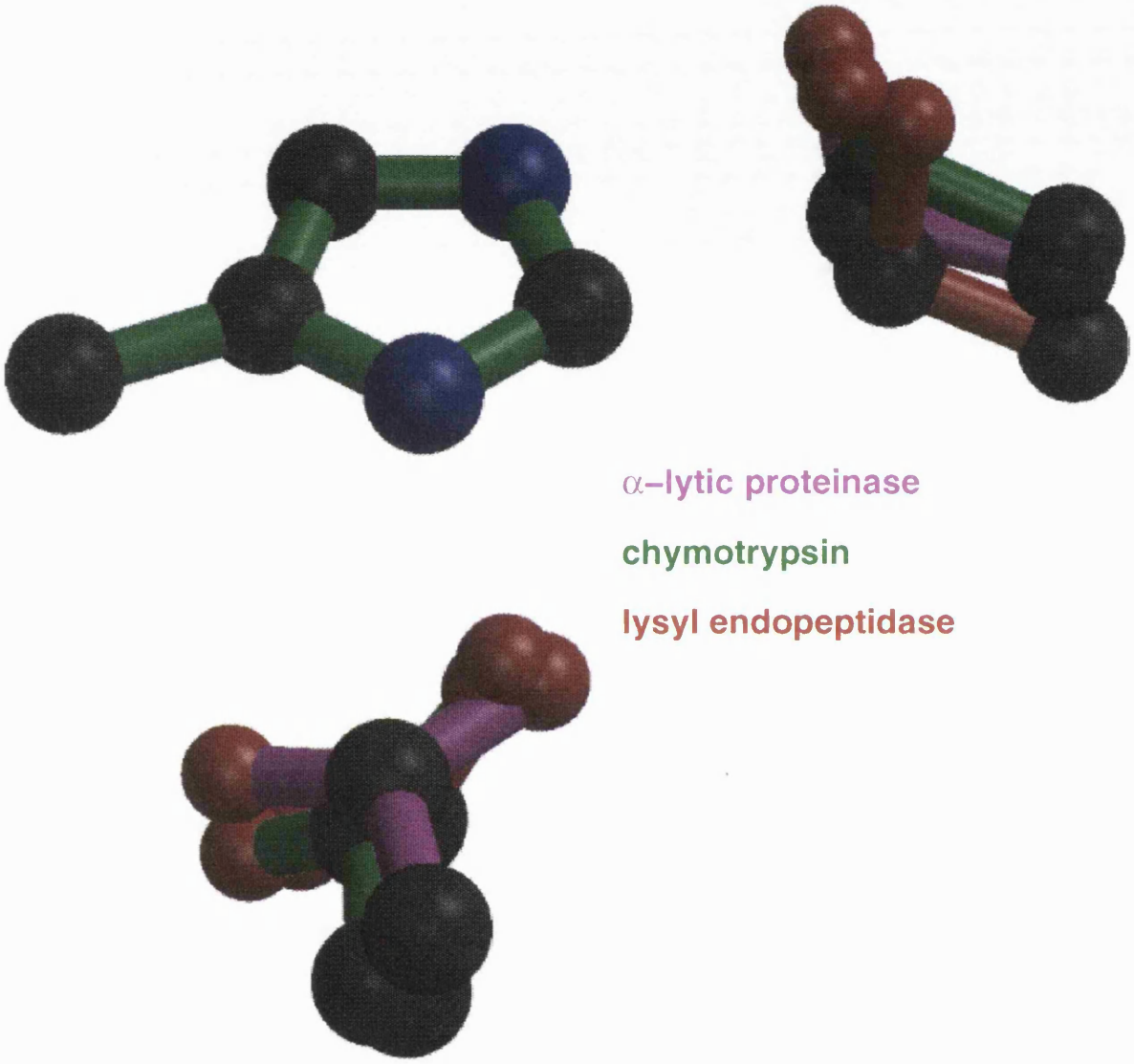


Figure 3.10: A superposition of three catalytic triads from enzymes in fold Group 1, each from a different subgroup: Group 1*a* chymotrypsin (1*cho*, Fujinaga *et al.*, 1987), Group 1*b* α–lytic protease (1*lpr* Bone *et al.*, 1991*a*) and Group 1*c* lysyl endopeptidase, 1*arb* (Tsunasawa S. *et al.*, 1989). These 3 enzymes have less than 30% sequence identity but their structures are highly similar and this is reflected in the similarity in the conformation of their Ser–His–Asp catalytic triads.

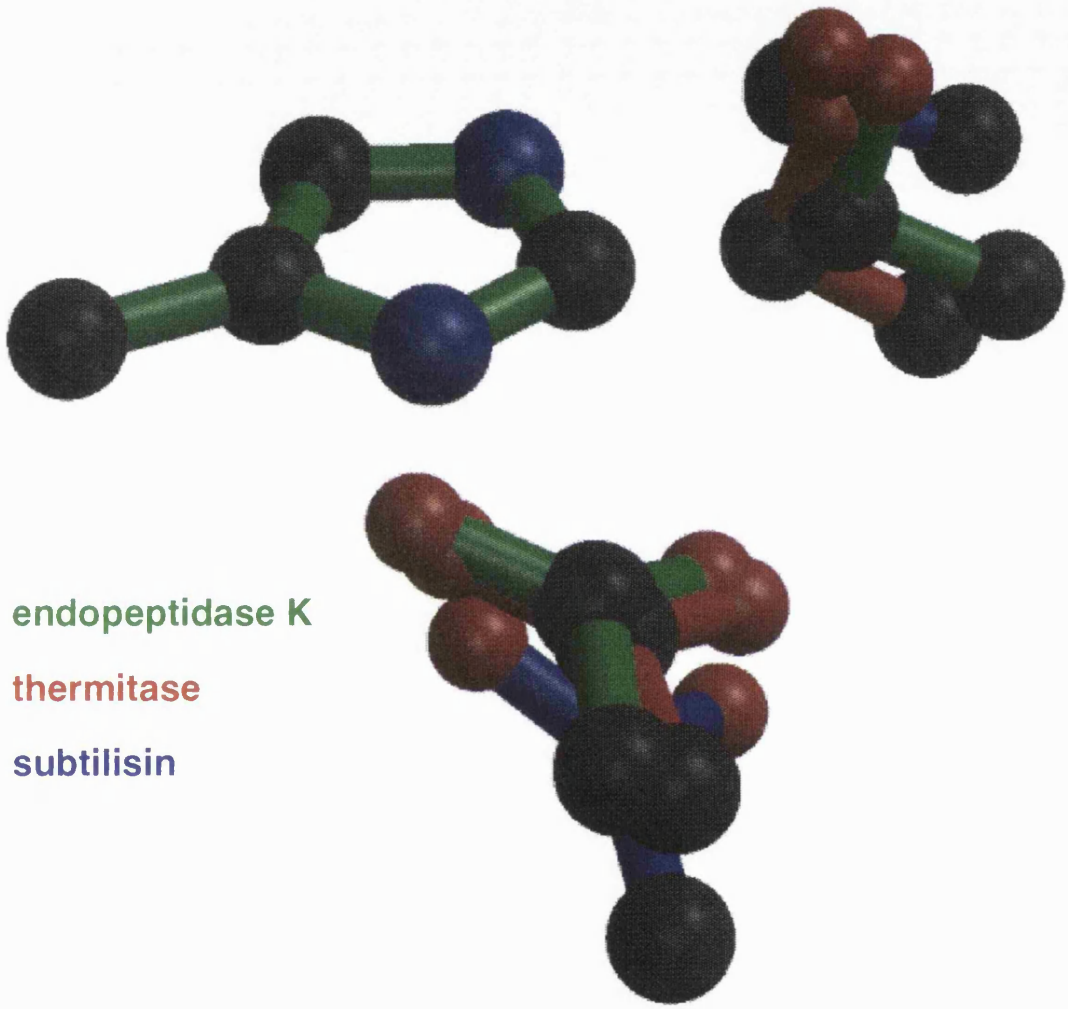


Figure 3.11: A superposition of three catalytic triads from enzymes in fold Group 2, each corresponding to a different E.C. number: subtilisin 2*sic* (Takeuchi *et al.*, 1991), E.C.3.4.21.62, endopeptidase K 2*pkc* (Bajorath *et al.*, 1989), E.C.3.4.21.64 and thermitase 1*thm* (Teplyakov *et al.*, 1990) E.C.3.4.21.66.

Residue	Number	Atom	x	y	z
Ser	195	O^{γ}	-1.15	4.87	-0.07
Asp	102	O^{δ_2}	3.68	0.06	0.06
His	57	C^{δ_2}	-1.09	0.80	0.00
His	57	C^{γ}	0.00	0.00	0.00
His	57	N^{δ_1}	1.11	0.82	0.00
His	57	C^{β}	0.07	-1.50	0.01
His	57	C^{ϵ_1}	0.70	2.09	-0.00
His	57	N^{ϵ_2}	-0.66	2.09	-0.00
Ser	214	O^{γ}	5.01	2.26	1.71
Ser	125	O^{γ}	2.28	5.71	-2.81

Table 3.4: Coordinates of the 'functional oxygens' and histidine sidechain of the consensus template triad. The mean position of the Ser 214 O^{γ} atom from structural group 1 and Ser 125 O^{γ} from structural group 2 is also given.

triads. The degree of the perturbation depends entirely on the type of inhibitor. In Figure 3.9 the inhibitor binds into the substrate binding site and apparently leaves the triad undisturbed. The *rms* distance of this catalytic triad is only 0.56Å from the overall template triad in Table 4.

Figure 3.12 shows a modified di–peptide inhibitor bound to another elastase structure, *7est* (Li De La *et al.*, 1990), that appears to bind adjacent and not parallel to the His sidechain imidazole ring. This forces the catalytic Ser away from the His ring, giving a *rms* distance of 2.23Å from the overall template triad. Two even more extreme examples of this are found in Figures 3.13 and

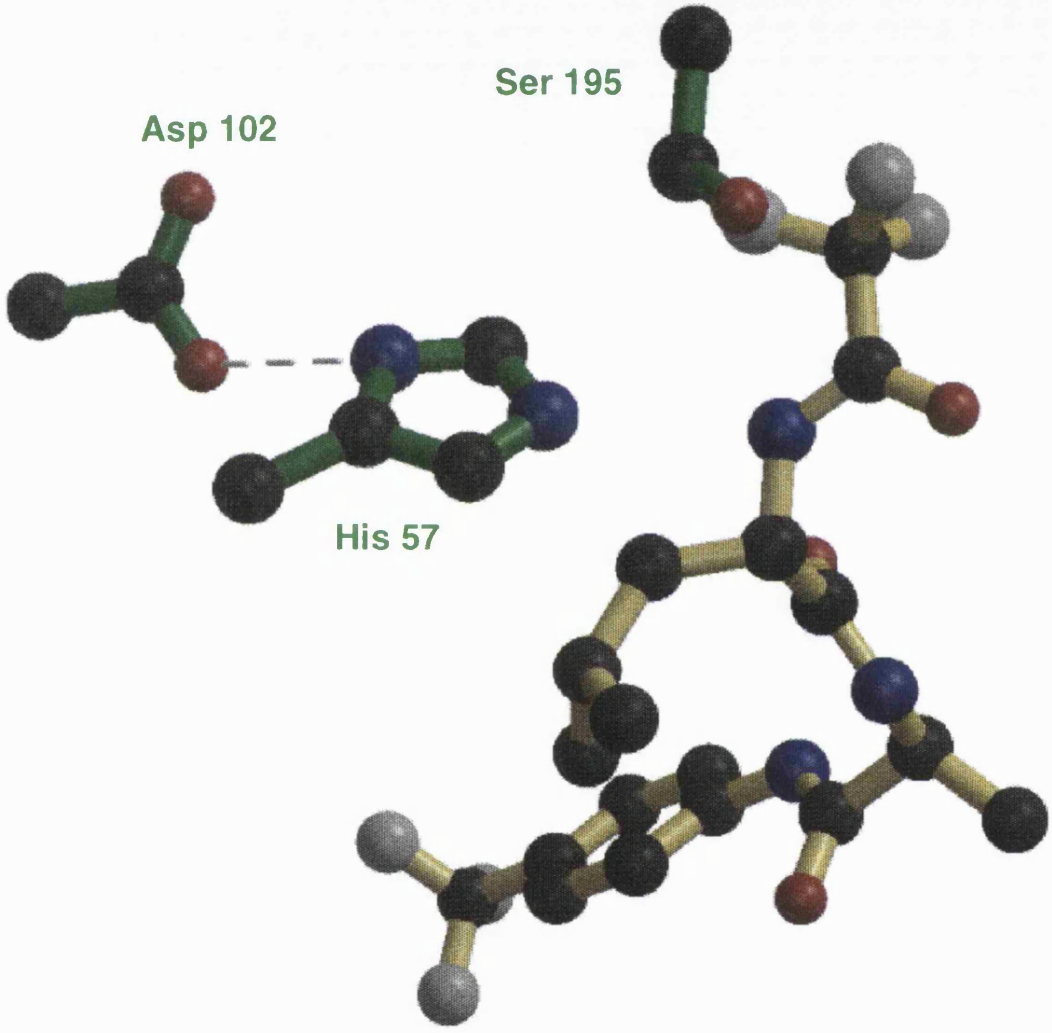


Figure 3.12: A modified di–peptide from *7est* (Li De La *et al.*, 1990) that binds adjacent, and not parallel, to the catalytic His ring (*cf.* Figure 3.9), perturbing the catalytic Ser out of its usual position

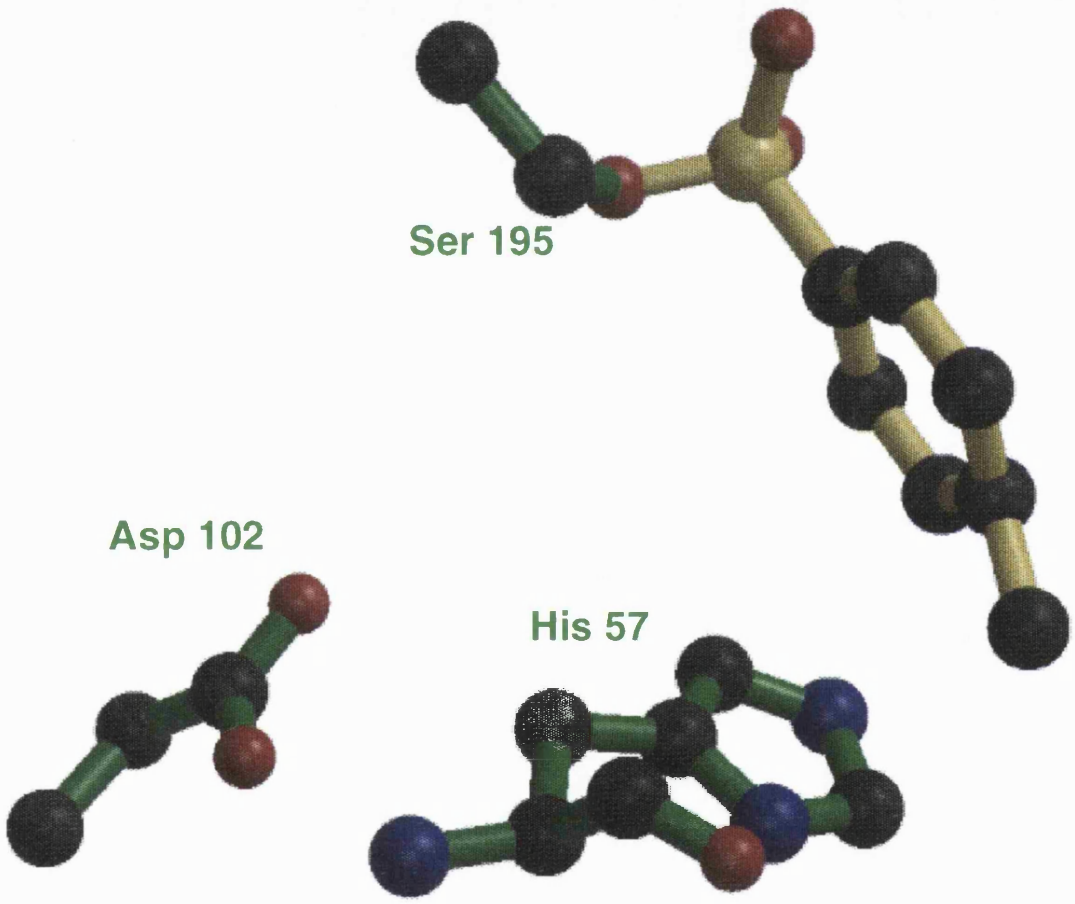


Figure 3.13: A tosyl group bound to the active site of 1*est* (Sawyer *et al.*, 1978).

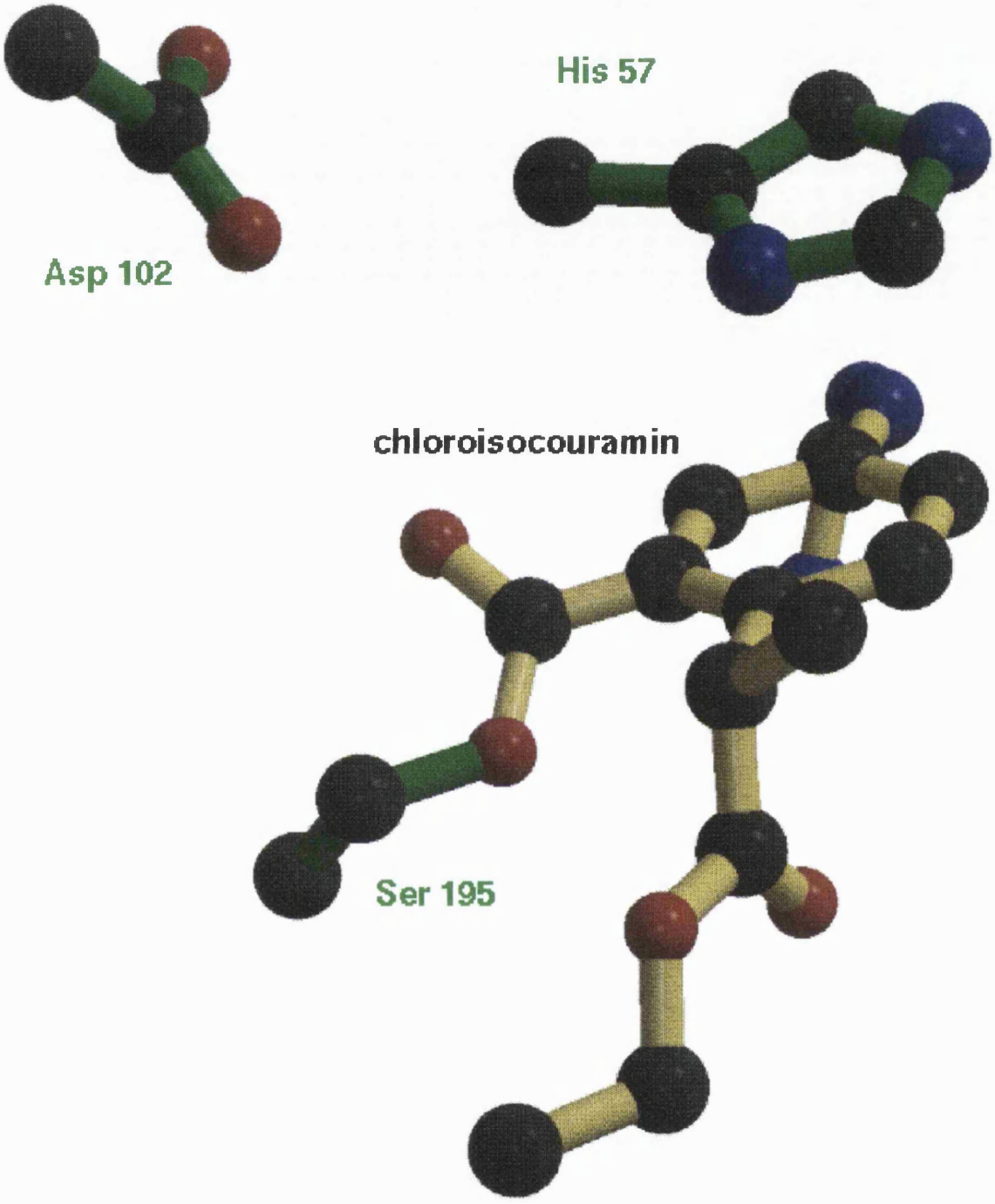


Figure 3.14: 7–substituted 3–alkoxy–4–chloroisocoumucarin inhibitor bound to the active site of *8est* (Powers *et al.*, 1990). The inhibitor is situated below the His sidechain, again perturbing the Ser and Asp catalytic residues out of their usual conformation. The catalytic residues are in an unrecognisable conformation when compared to Figure 3.9.

3.14, which have inhibitors that covalently bind to the Ser 195 in the active sites. The first is the elastase structure 1*est* (Sawyer *et al.*, 1978) which has a tosyl group inhibitor. The catalytic Ser and Asp are forced into unrecognisable conformations when compared to Figure 3.9. Figure 3.14 shows the active site of the elastase structure, 8*est* (Powers *et al.*, 1990) which has a 7–substituted 3–alkoxy–4–chloroisocoumarin molecule bound. This heterocyclic compound reacts initially by acylation of the active site Ser 195 forming the acyl–enzyme (as in Figure 3.14) and then undergoes further reactions with other active site residues giving an extremely stable inactivated enzyme inhibitor complex.

In fact, these last two enzymes have catalytic triads whose conformations are so perturbed that they are not even identified as an interacting Ser–His–Asp triplet by the program DISTRIB. The same applies to two other elastase structures 1*inc* (Radhakrishnan *et al.*, 1987), 1*jim* (Meyer *et al.*, 1985), five thrombin structures 3*htc*, 1*nrn*, 1*nro*, 1*nrp* and 1*nrq* (Mathews *et al.*, 1994), trypsin 1*tpa* (Marquart *et al.*, 1983), four trypsin structures 2*tgd*, 1*trm*, 2*trm* (Rydel *et al.*, 1990) and 2*tld* (Takeuchi *et al.*, 1992), α–lytic proteinase 1*p*11 (Bone *et al.*, 1991*b*), tonin 1*ton* (Fujinaga *et al.*, 1987) and streptogrisin A 3*sga* (James *et al.*, 1980). These unidentified triads account for the discrepancies in the number of chains in our dataset in Table 3.3 (*e.g* 170 for group 1) and the number of triads identified.

Group 2 - doubly-wound α/β subtilisin-like fold

29 catalytic triads identified from 35 chains. Mean *rms* deviations from Group 2 template: 'functional' 0.58Å, 'sidechain' 0.70Å

Figure 3.15 shows the catalytic triad of one of the Group 2 enzymes: Ser 221, His 64, Asp 32 in subtilisin, 2*sic* (Takeuchi *et al*, 1991). The Ser 125 mainchain carbonyl oxygen is hydrogen bonded to the functional Ser 221 O^{γ} and the Ser

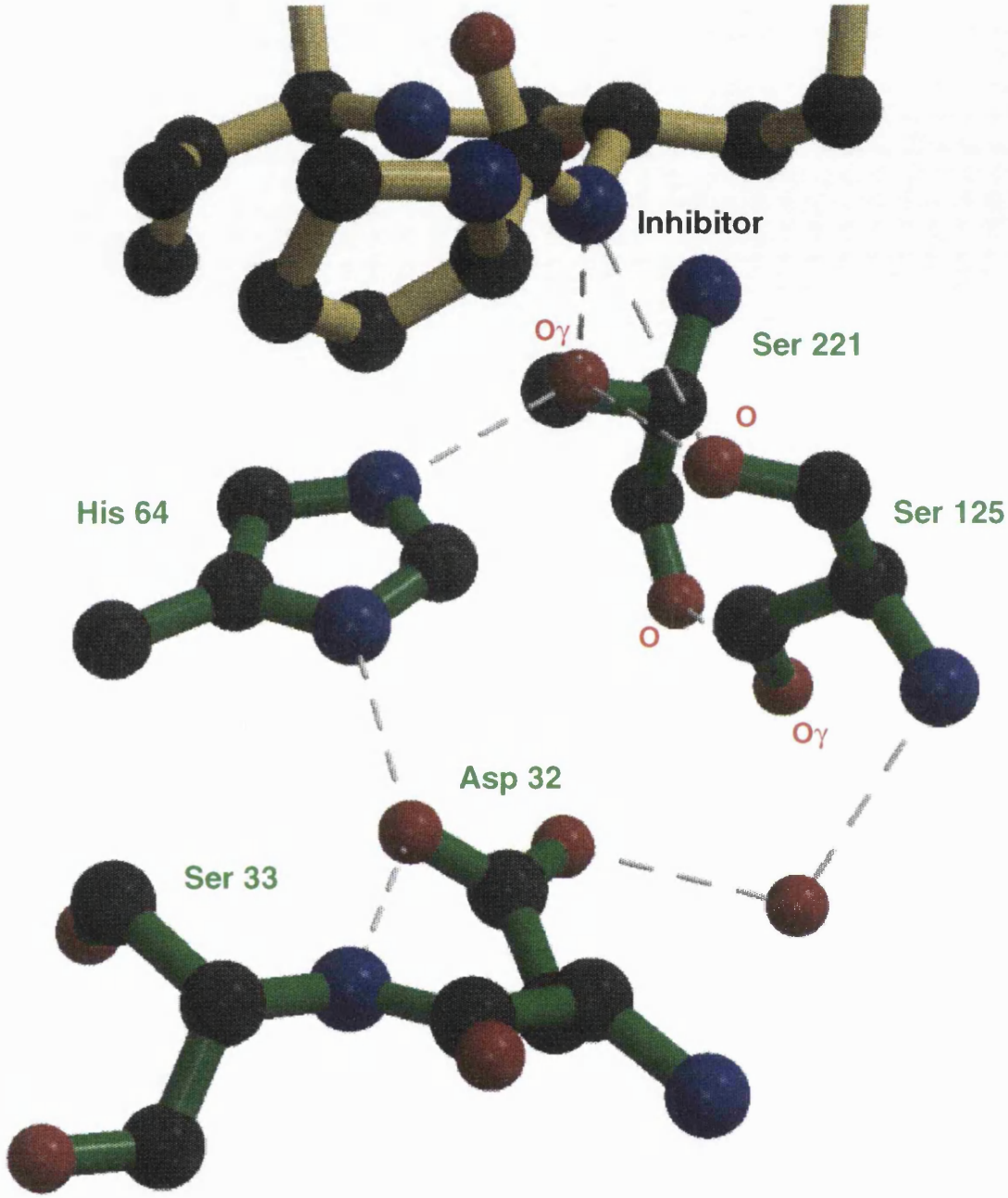


Figure 3.15: The Ser 221–His 64–Asp 32 catalytic triad, plus inhibitor and the Ser 125 residue that is found in a structurally conserved position in the active site of subtilisin. The latter residue is analogous to the Ser 214 residue of the Group 1 enzymes in that it plays a functional role in hydrogen bonding to and orientating the catalytic triad residues even though it hydrogen bonds to different atoms.

125 O^{γ} is hydrogen bonded to the Ser 221 mainchain carbonyl oxygen thereby orientating this residue to optimise its nucleophilic character. In addition, the backbone nitrogen of Ser 125 hydrogen bonds to a water, which in turn is hydrogen bonded to the 'non–functional' carboxyl oxygen of Asp 32. The functional Asp 32 carboxyl oxygen is held in position by the neighbouring backbone N of Ser 33, which serves a similar role to the backbone His 57 N in the Group 1 enzymes.

Figure 3.11 shows an example of each of the enzymes in this group - subtilisin *2sic* , endopeptidase K *2pkc* (Bajorath *et al.*, 1989) and thermitase *1thm* (Teplyakov *et al.*, 1990). The triads of this group are structurally conserved with the *rms* deviation of the mean subtilisin triad being only 0.54Å and 0.52Å respectively from the mean triads of endopeptidase K and thermitase.

There are four subtilisin structures which do not have their catalytic triads identified by the 3D consensus template. Three of these, *1sub*, *1suc* and *1sud* (Gallagher T. *et al.*, 1993) are missed by DISTRIB because their catalytic Ser 221 residue has been mutated to a Cys and also *1sel* which has its catalytic Ser 221 covalently bound to a selenium. This explains the discrepancy between the number of chains and number of triads identified for Group 2.

Group 3 - serine-type carboxypeptidases

4 catalytic triads identified from 7 chains. Mean *rms* deviations from Group 3 template: 'functional' 0.65Å, 'sidechain' 0.85Å.

Figure 3.8 compares a catalytic triad of the serine-type carboxypeptidase *1wht* (Liao *et al.*, 1992) with representatives from each of the other fold groups subtilisin, chymotrypsin and lipase. The most striking difference which distinguishes this group from Groups 1 and 2, is that the catalytic Ser of the serine–type carboxypeptidase sidechain delivers its Ser O^{γ} from above the plane of the His rather than from below as in chymotrypsin and subtilisin. In contrast, the Asp looks

very similar to lipase. However, the functional oxygens from both the Asp and Ser overlap the other functional groups very well (*rms* deviation from combined template is 0.83Å).

Group 4 - triacylglycerol lipases

10 catalytic triads identified from 13 chains. Mean *rms* deviations from Group 4 template: 'functional' 0.45Å, 'sidechain' 1.04Å.

There are two distinct triad conformations that occur in fold Group 4. The first is that shown in Figure 3.8 with representatives from the 3 other fold groups. It is the conformation of a typical bacterial lipase, *3tgl* (*e.g.* Brady *et al.*, 1990; Noble *et al.*, 1991). The Ser sidechain atoms are in a similar conformation to the serine–type carboxypeptidase; that is, above the plane of the His ring when compared to Groups 1 and 2. The other catalytic triad conformation in Group 4 is shown in Figure 3.16 for dimeric horse pancreatic lipase, *1hpl* (Bourne *et al.*, 1993). In this structure, the Asp sidechain is above and adjacent to the His ring. Once again this illustrates that the position of the Asp sidechain in a catalytic triad is restricted only in so far that one of its carboxyl oxygens is delivered to a favourable hydrogen bonding position relative to the His imidazole ring.

3.3.2 Position of the oxygens in the catalytic triad

It is clear that the position of the sidechain atoms of both the Ser and the Asp residues with respect to the His imidiazole ring in catalytically active Ser–His–Asp triads is highly variable. However, all these catalytic triads, except those that are severely distorted by the presence of an inhibitor, have one of their Asp carboxyl oxygens in a conserved position that enables it to hydrogen bond to the His $N^{\delta 1}$, and the Ser O^{γ} in a hydrogen bonding position with the His $N^{\epsilon 2}$ (see Figure 3.8).

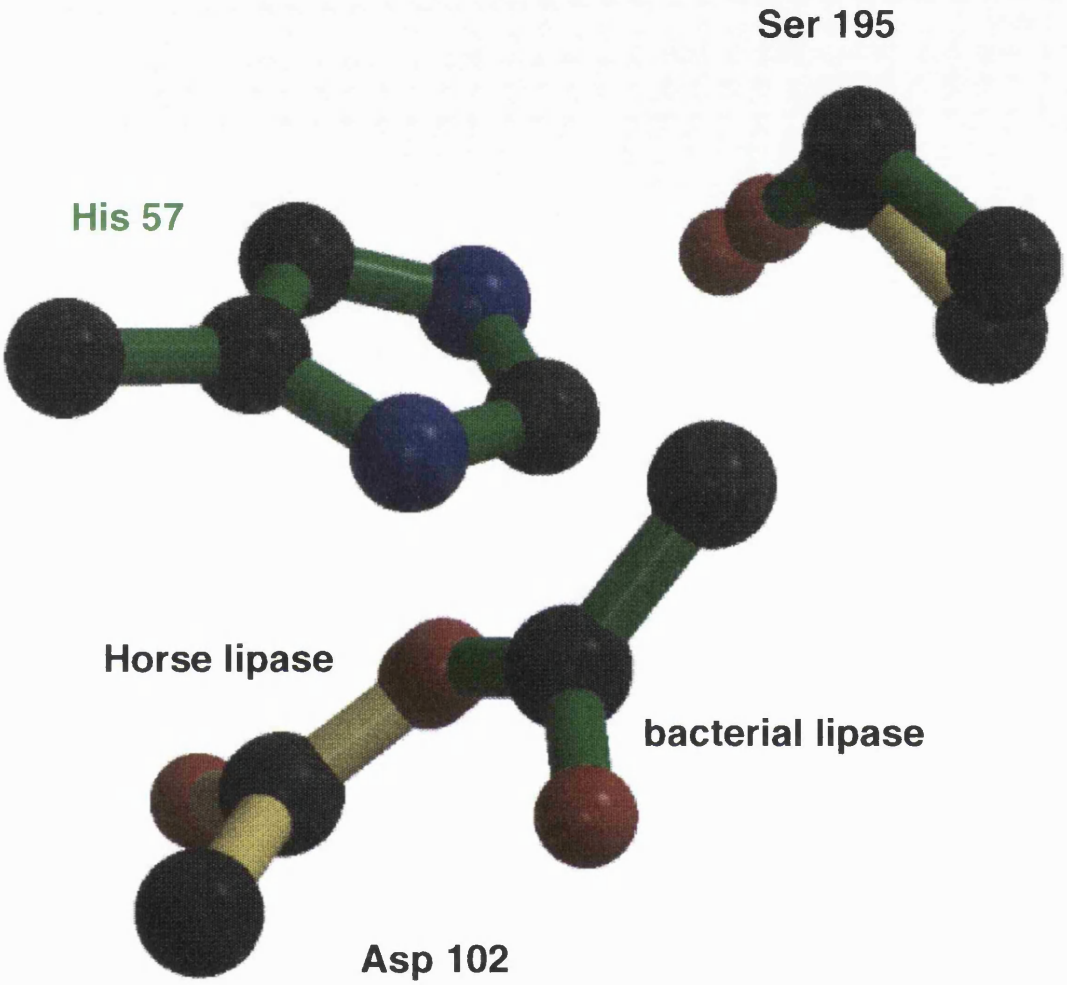


Figure 3.16: Comparison of the catalytic triad of horse lipase 1*hpl* (Bourne *et al.*, 1993) and bacterial lipase 3*tgl* (Brady *et al.*, 1990; Noble *et al.*, 1991)showing the unusual position of the Asp in the former triad. One of the Asp carboxyl oxygen atoms from this triad is still in a position to hydrogen bond to the His ring.

Consensus templates of the His plus these two functional oxygens for all the four structural groups were derived and the *rms* distances of these templates from each other calculated (Table 3.3). In addition, the mean coordinates of these four templates gives us the overall mean consensus template, the coordinates of which are given in Table 3.4. The *rms* distance values of every 'functional oxygen' structural group mean from the consensus template lie between 0.39Å and 0.85Å indicating the positional conservation of the two hydrogen bonded oxygens across the four structural groups. To illustrate this diagrammatically, the mean positions for each of the four structural group functional Asp carboxyl oxygens and Ser O^{γ}s are plotted in Figure 3.17, showing their strong conservation.

The mean distance of the consensus template Asp carboxyl oxygen hydrogen bond acceptor to the His $N^{\delta 1}$ donor is 2.69Å. If a proton is modelled onto this His atom, the angle His $N^{\delta 1}$–His H–Asp O is 177°. Both these criteria are very close to optimal hydrogen bonding geometry (McDonald and Thornton, 1994). The Ser O^{γ} to His $N^{\epsilon 2}$ distance is also close at 2.80Å, but Figure 3.8 shows that the C^{α}–C^{β} bond of the Ser is almost perpendicular to the plane of the His sidechain in all the 4 Groups' Ser residues. A hydrogen was modelled onto the Ser O^{γ} using the program HBPLUS (McDonald & Thornton, 1994) which models hydrogens with the criteria defined by Momany *et al.* (1975). HBPLUS positions the mobile hydrogen atoms of serine, threonine, tyrosine and cysteine sidechains in 2 stages. In the first stage it calculates the range of possible hydrogen bonding positions. In the second stage, it positions the donor's hydrogen separately for each putative hydrogen bond, as close to the acceptor as possible. The largest angle Ser O^{γ}–Ser O^{γ}H–His $N^{\epsilon 2}$ achievable is only about 125°, suggesting a weaker hydrogen bond interaction. Indeed, electrostatic calculations suggest that a deviation of 20° from linear decreases hydrogen bond energy by 10% (Pimental and McClellan, 1960). However, in this case the Ser O^{γ}–Ser O^{γ}H–His $N^{\epsilon 2}$ interaction is there for

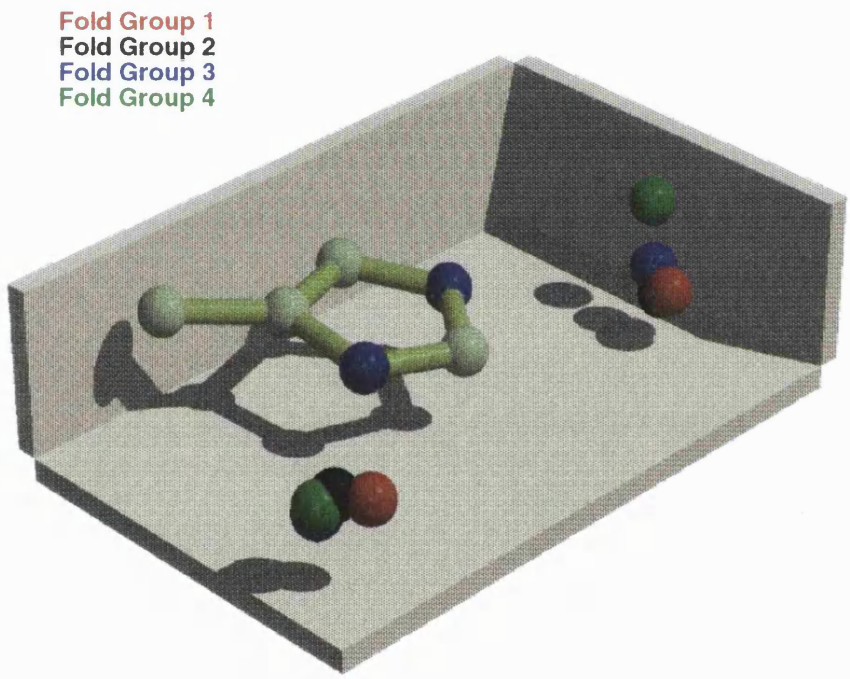


Figure 3.17: A box–plot showing the mean positions of the Ser O^{γ} and the Asp carboxyl oxygen atom for each of the 4 fold groups. These atoms all converge at favourable hydrogen bonding positions relative to the nitrogens of the His ring.

acid/base catalysis and not structural stabilisation and the Ser O^{γ} is placed in a position to perform a nucleophilic role.

3.3.3 Template search through the Enzyme Dataset

Given the combined mean consensus template of Table 3.4, the question arises whether it can be used to distinguish the genuine catalytic triads from the ordinary, non–catalytic associations between Ser, His and Asp residues. Figure 3.18 shows the results for all the Ser–His–Asp triplets in the enzyme dataset. It shows the *rms* distance of the Ser O^{γ} and the closest of the Asp carboxyl oxygens from the overall mean consensus template. The catalytic triads are shaded in black and can be clearly seen to lie within 2.0Å of the mean coordinates. This would appear to define a cut–off region within which these atoms need to lie for the triad to be catalytically active. Table 3.3 indicates that the functional oxygen template of fold Group 1 lies 0.47Å from the mean catalytic triad of the four structural groups and there is indeed a peak at 0.5Å in the histogram in Figure 3.18. It is noticeable that most of the catalytic triads have an *rms* distance of 1.4Å or less and this suggests that the cut–off could be lower as those structures above this *rms* value are those which are distorted by inhibitors bound to their active sites as shown in Figures 3.12, 3.13 and 3.14.

The other triplets in Figure 3.18 which are above the 2.0Å cut–off, are non–catalytic Ser–His–Asp associations and are colour coded according to their structural Groups. There is a sharp peak at 5.0Å and this corresponds to the non–catalytic triad Ser 214–His 57–Asp 102 found in the enzymes of fold Group 1. Notice there is a near 1:1 ratio of triads at the peaks of 0.5Å and 5.0Å because Ser 214 is found in a structurally conserved position in all fold Group 1 active sites. Indeed, these four residues, the catalytic triad plus Ser 214, have been described as the catalytic quartet (Barth *et al.*, 1994).

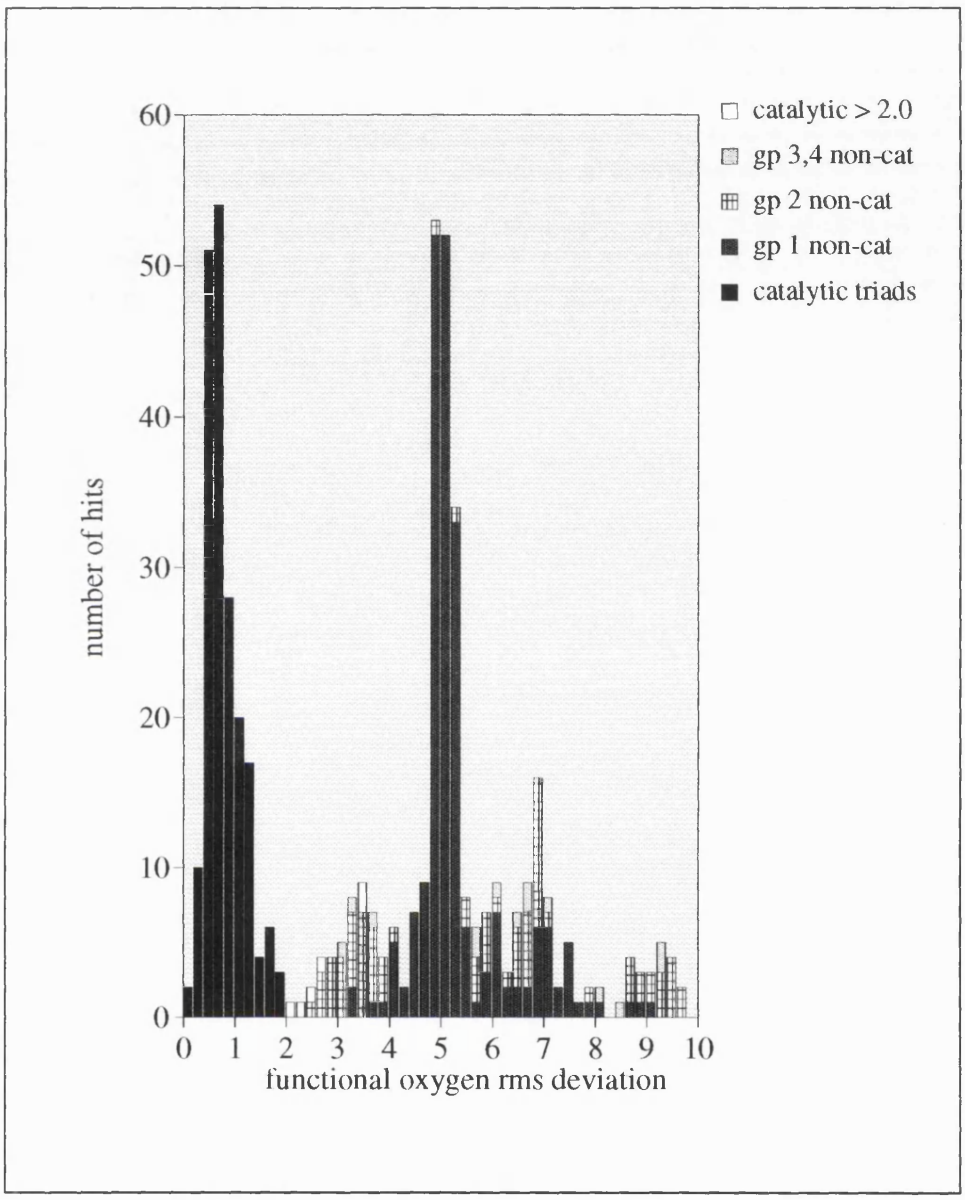


Figure 3.18: A histogram of the *rms* distance of the Ser O^{γ} and Asp carboxyl oxygen atom from the overall mean consensus template position for all Ser, His and Asp associations in the enzyme dataset. This histogram shows how the majority of catalytic triads, in black, are within 2.0Å of the consensus template and can be separated from the non–catalytic interactions. The catalytic triads that lie beyond this cut–off (in white) are those whose conformation has been perturbed by the binding of an inhibitor. In addition, the triads at *rmsd* 2–5.5Å in the histogram represent structurally conserved non–catalytic triads that play a role in hydrogen bonding to and orientating the catalytic triads.

To check the extent of this structural conservation, we derived a consensus template for the Ser 214–His 57–Asp 102 triplet in Group 1 enzymes, using the same method as for the catalytic triads. We find 152 Group 1 Ser 195–His 57–Asp 102 triads compared to 141 Ser 214–His 57–Asp 102 triads. Three of the missing cases are lysyl endopeptidase which does not have the Ser 214 residue; its Asp is hydrogen bonded to two water molecules instead. The others are again caused by inhibitors which push the *rms* deviations above the 2.0Å cut–off. We found that the Ser 214–His 57–Asp 102 triplets, like the Ser 195–His 57–Asp 102 triads all have an *rms* distance of less than 1.0Å from their overall mean. In other words, the position of the non–catalytic Ser 214 is just as tightly defined as the catalytic Ser 195. It appears that Ser 214, together with the peptide backbone of the catalytic His 57, form a network of hydrogen bonds that enable the Asp 102 to be presented in the optimal position for interaction with the His 57 imidazole ring (Figure 3.9). Indeed, the Ser 214 residue has been implicated in performing an electrostatic stabilisation role and mutation of this residue produced decreases in free energy of catalysis which were in agreement with electrostatic calculations (McGrath *et al.*, 1992).

Corey *et al.* (1992) investigated the effect of swapping this non–catalytic serine with the functional Asp. They formed the double mutation of D102S and S214D into the gene coding for rat anionic trypsin, expressed this in *Escherichia coli* and then solved the X-ray structure. The Asp in this mutant was in a totally different position to that in the catalytic triad, though it still formed a hydrogen bond to the His imidazole ring. There was catalytic activity but the k_{cat} was reduced 100 fold indicating that, although a charged Asp in the vicinity is sufficient for low catalytic activity, the position and hydrogen bond interactions of the Ser 214 residue are important for efficient catalysis.

Returning to Figure 3.18, there is another peak at *rms* distance 2–4.5Å which

corresponds to a structurally conserved triplet of Ser 125–His 64–Asp 32 in Group 2. We again derived a consensus template for these triplets. As for the Group 1 non–catalytic Ser 214–His 57–Asp 102 triplet, the Group 2 non–catalytic triad is structurally conserved. There are 29 Ser 125–His 64–Asp 32 triplets identified from a total of 35 chains with the *rms* distance values all being well below 1.00Å, suggesting that the position of Ser 125 is also well conserved. In Figure 3.15, the Ser 125 N is hydrogen bonded via a water to Asp 32. In contrast, Figure 3.9 shows that Ser 214 is directly hydrogen bonded to Asp 102. As mentioned above, Ser 214 has been implicated in stabilising the charge on the buried Asp 102 (McGrath *et al.*, 1992). As yet, there have been no experiments to investigate whether Ser 125 plays a similar role.

The serine–type carboxypeptidases of fold Group 3 do not have a conserved non–catalytic serine. However, it has been suggested that Asn 176 plays a similar role (Liao *et al.*, 1992), but, unlike the two serines, the Asn sidechain is out of the plane of the imidazole ring, which calls into question the significance of this residue .

There is a further peak in Figure 3.18 at 6–7Å *rms* distance, which corresponds to non–catalytic triplets where the Ser O^{γ} hydrogen bonds to His N^{δ_1}, while the Asp carboxyl O hydrogen bonds to His N^{ϵ_2}. This is the opposite hydrogen bonding conformation to that of a Ser–His–Asp catalytic triad (*i.e* Ser O^{γ} hydrogen bonding to His N^{ϵ_2}) hydrogen bonds to and is catalytically inactive because the Ser lies close to the histidine backbone and so would cause steric hindrance to a substrate.

3.3.4 Template search through PDB

It is interesting to see how often the Ser–His–Asp catalytic triad conformation, in terms of the functional oxygens from Asp and Ser, relative to a His residue,

occurs in other protein structures. Having derived the appropriate cut–offs, we can now use the template to search through our dataset of representative protein structures in the January 1995 PDB. This dataset contains some of the structures in the enzyme dataset, the missing ones having been excluded on the basis of having higher than 95% sequence identity.

Figure 3.19 shows us that most of the Ser–His–Asp triplets in the non–enzyme structures have an *rms* distance of > 2.0Å from the consensus template. However, there are two proteins, which are neither serine proteinases nor lipases, but which have a Ser–His–Asp triad with an *rms* deviation below the 2.0Å cut–off. These are the Ser 99–His 92–Asp 123 triad of cyclophilin A, 2*cpl* (Ke, 1992) with *rms* 1.38Å and Ser 191–His 225–Asp 222 of chain H of immunoglobulin G1 2*ig*2 with *rms* 1.57Å (Marquart *et al*, 1980).

The first of these, cyclophilin A, is a binding protein for the immunosuppressive drug cyclosporin A and is also an enzyme with peptidyl–prolyl *cis–trans* isomerase activity. Figure 3.20 shows the identified Ser 99–His 92–Asp 123 triad (red bonds). Although the mechanism of action of this enzyme has yet to be fully elucidated, various residues have been identified as possibly important for catalysis. One of these, His 126, is shown in green in Figure 3.20. Cyclophilin A is a β–barrel structure with eight antiparallel β–strands wrapping around the surface of the barrel and two α–helices sitting on the top and bottom of the barrel. The triad is in the vicinity of the catalytic His 126, lending weight to the theory that this protein has proteinase activity. There has, however, been no proteinase like activity reported for this protein. Since there is no conservation of gross topology with any of the fold groups in Table 3.1, it is unlikely to be evolutionarily related. However, since enzymes are highly specific for substrate, cyclophilin A may not have been sufficiently assayed for proteinase activity.

Figure 3.21 shows a close up of the cyclophilin 'catalytic' triad. The Ser O^{γ}

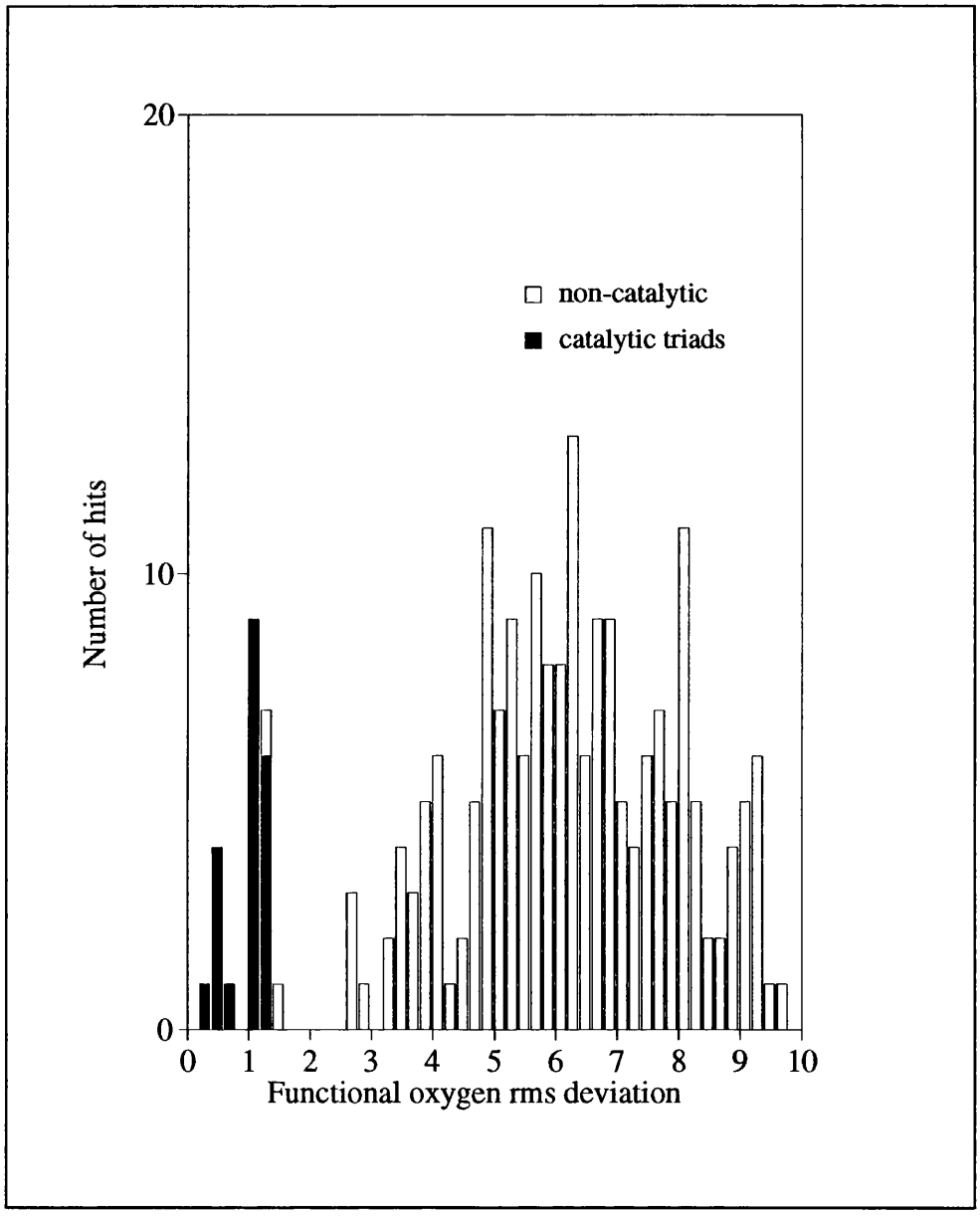


Figure 3.19: Histogram showing the *rms* deviation from the 'functional' consensus template of all Ser–His–Asp interactions extracted from a dataset of non–homologous proteins. The serine proteinases in the protein dataset are shown in black and these are clearly separated from the other, non–catalytic associations. There are, however, two proteins that are not serine proteinases, cyclophilin and immunoglobulin, shown in white, that appear to have a Ser–His–Asp triad in the catalytic conformation.

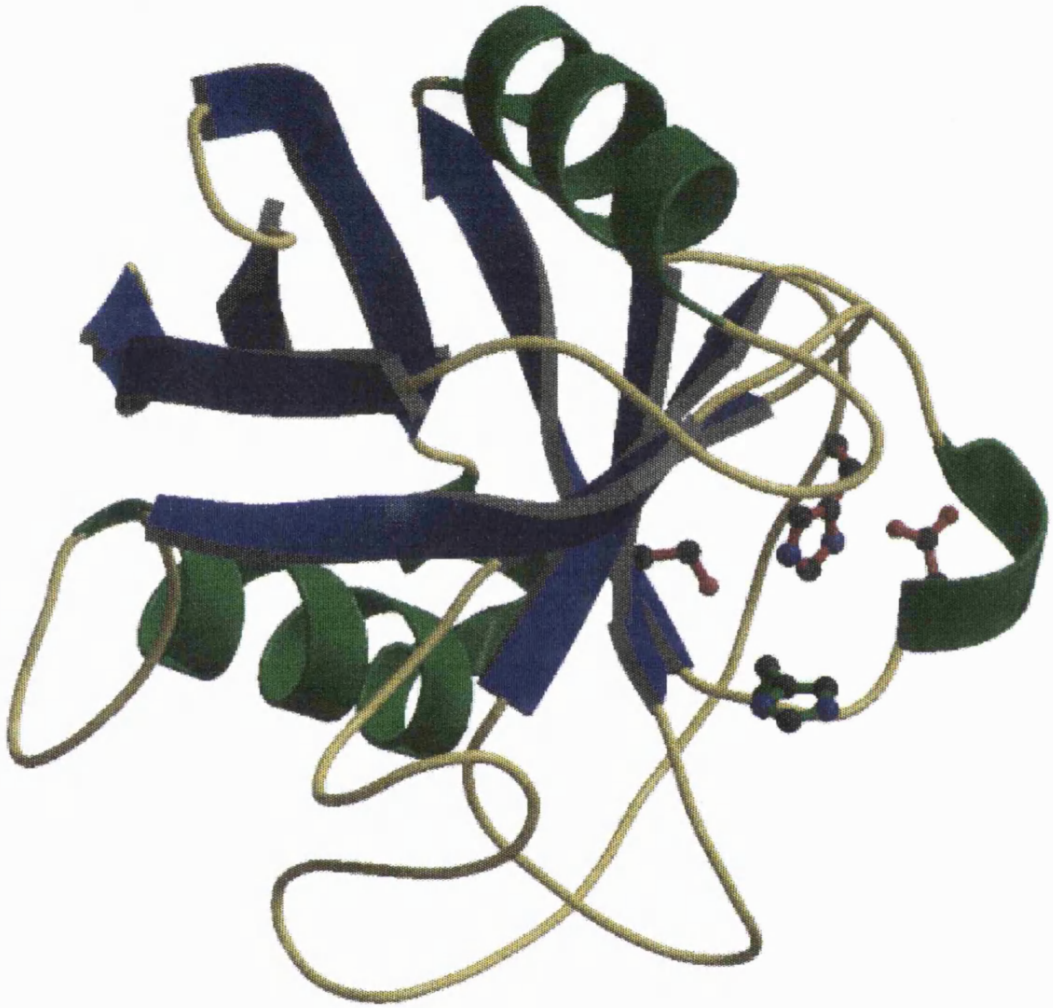


Figure 3.20: A MOLSCRIPT diagram of cyclophilin A in which His 126, shown in green bonds, is thought to be involved in the peptidyl–prolyl *cis–trans* isomerase activity. Also shown is the Ser 99–His 92–Asp 123 triad (red bonds) that may enable cylcophilin A to exhibit protease activity.

and Asp carboxyl oxygen appear to be in an optimal hydrogen bonding position. The Asp carboxyl oxygens form a network of stabilising hydrogen bonds with the surrounding residues making the Asp very similar in nature to a real catalytic aspartate. However, the triad is surrounded by 3 hydrophobic residues, Val 128 and Leu 122 above and below the His 92 ring as well as Phe 113 directly below the catalytic Ser 99 O^{γ}, suggesting that steric hinderance would inhibit any substrate binding. There is the possibility that the binding of a substrate could cause a conformational change in the enzyme which might enable the Phe 113 to move out of the way of the Ser, enabling catalysis to occur. To test the accessiblility and reactivity of the Ser to substrate, diisopropylphosphofluoridate should form a irreversible covalent adduct with the Ser, as it does in all serine proteinases and lipases (Hayashi *et al*, 1973).

The second non–enzyme with an apparent catalytic triad is immunoglobulin G1, 2*ig*2 with an *rms* deviation of 1.57Å. Figure 3.22 shows the Ser 191-His 225-Asp 222 'catalytic' triad surrounded by mostly hydrophobic residues. The triad looks rather different, since the Asp sidechains approach from a different orientation so that the sidechain *rms* is 4.20Å. It lies on the surface of the heavy chain of the immunoglobulin molecule with the Ser O^{γ} pointing out towards the surface. There is also the Ser 192 residue that is in the vicinity of the triad which could be compared to the Ser 214 or Ser 125 residues of the Group 1 or Group 2 enzymes respectively (Table 3.1). However, the immunoglobulin Ser 192 position, though next to the 'catalytic' Ser 191, is in a different position to the Ser 214 or Ser 125 enzyme residues with respect to the His. Figure 3.23 shows the position of the 'catalytic' triad with repect to the whole Ig fragment. The triad lies at the C-terminus of the molecule, near the hinge region, and the triad would probably be buried if the X-ray structure of the whole Ig molecule were available. In addition, this is far away from the hapten binding site, which is

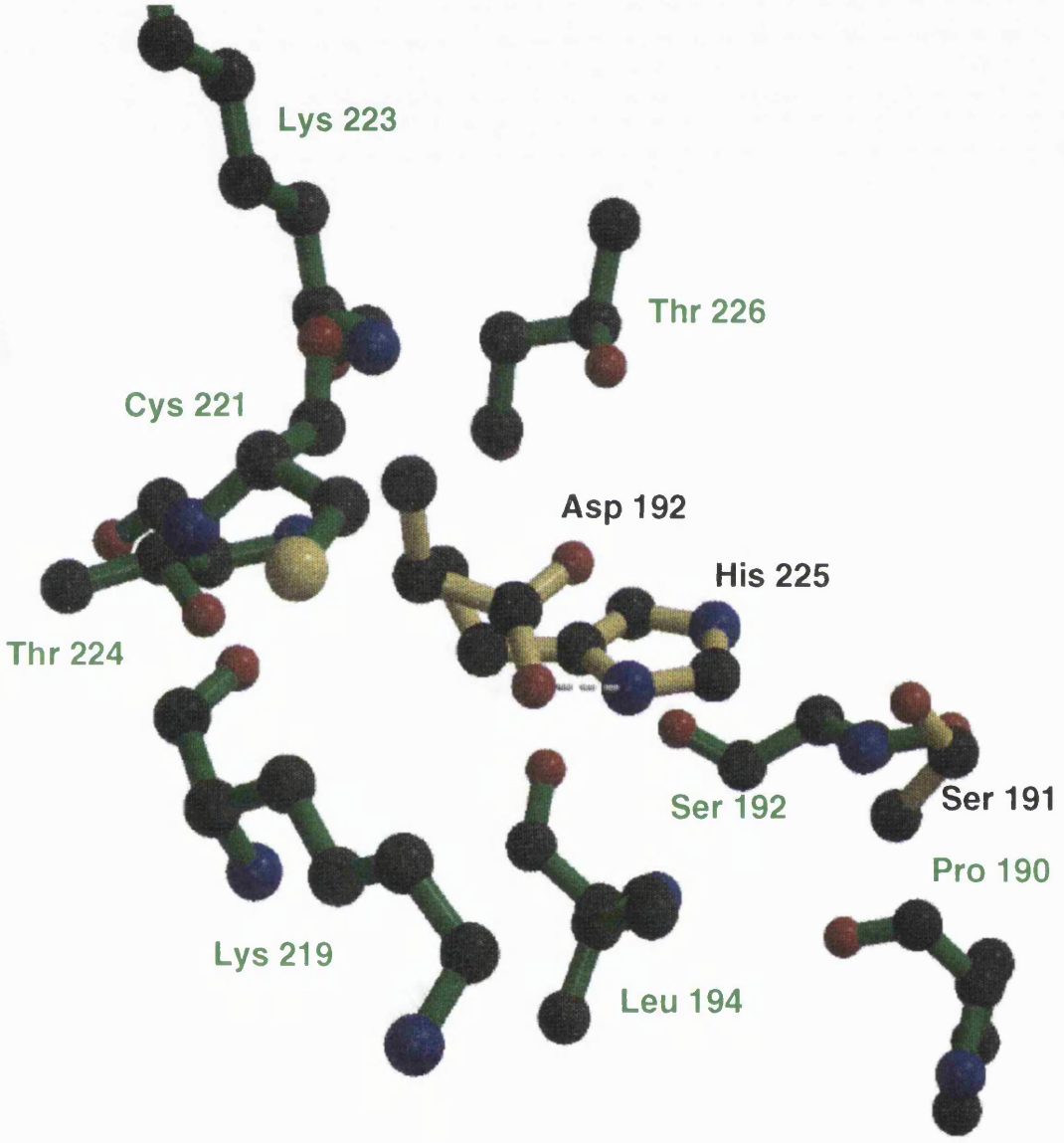


Figure 3.22: Ser-His-Asp triad found in the immunoglobulin molecule G-1 (2*ig*2, Marquart *et al.*, 1980).

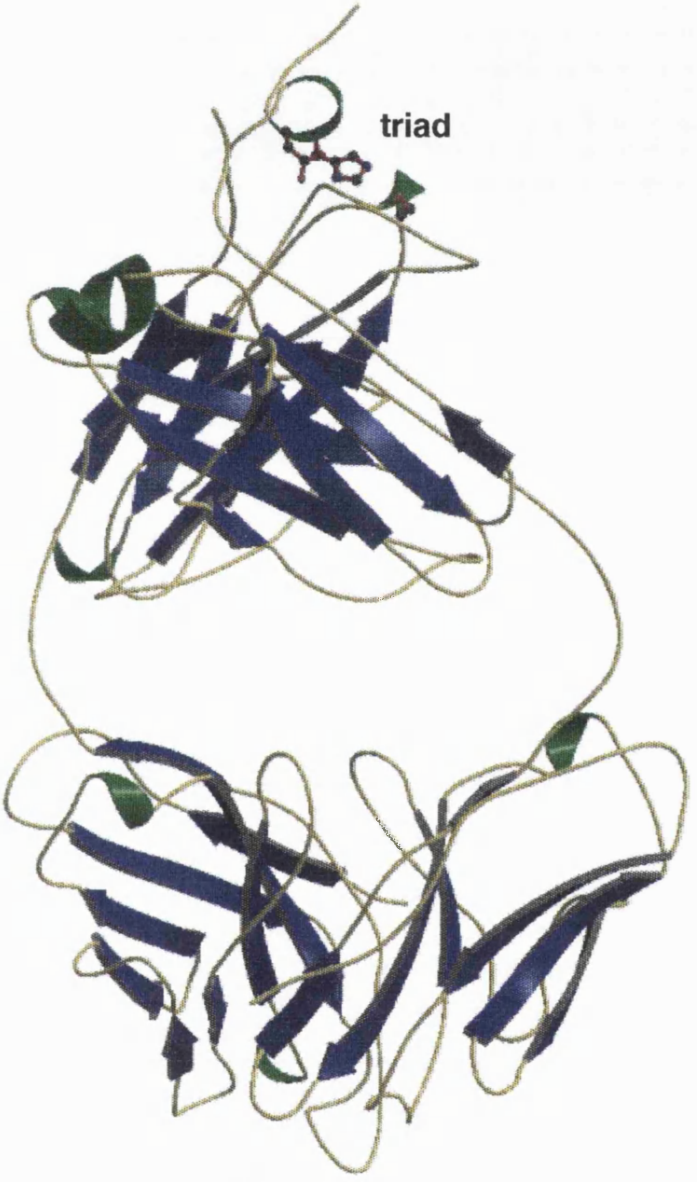


Figure 3.23: A MOLSCRIPT diagram of the intact immunoglobulin fragment with the position of the 'catalytic' triad also shown. The triad lies at the C-terminus of the molecule whereas the hapten binding site is at the N-terminus.

situated at the N-terminus.

3.3.5 Other catalytic triads

This paper has only dealt with the Ser–His–Asp catalytic triad, but the question arises as to whether the template triad we have derived can be used to identify other catalytic or structural triads which employ residues other than Ser, His and Asp. Wei *et al.* (1995) have identified a novel catalytic triad in *Streptomyces scabies* that employs the mainchain carbonyl of a Trp residue to hydrogen bond to the His $N^{\delta 1}$ instead of the Asp carboxyl oxygen, the triad in this case being Ser 144–His 283–Trp 280. We have applied our template to this novel catalytic triad and have found that the *rms* deviation of the Ser O^{γ} and mainchain Trp carbonyl oxygen from our template is 1.92Å. This is larger than the 1.9Å seen for 'native' enzymes, but is still below the 2Å cut off defined above. Upon inspection, we found that the Ser O^{γ} is distorted away from the mean template position by 2.64Å because the structure contains the covalent inhibitor *bis–p–*nitrophenylmethylphosphonate. We have already noted that the Ser–His–Asp catalytic triad is sometimes distorted by binding of unusual or covalent inhibitors.

Of course, there are other enzyme structures in the PDB which employ a His residue as part of their catalytic machinery, for example papain (Cys, His, Asn) and malate dehydrogenase (His, Asp). An extension of this work would be to investigate the structural similarities in catalytic centres of such enzymes with our template triad. This work is currently in progress.

3.4 Conclusion

We have compared the conformation of the Ser–His–Asp catalytic triads in the serine proteinases, serine–type carboxypeptidase and triacylglycerol lipases.

There are significant differences in the conformations of the Asp and Ser sidechain atoms relative to the His, but the positions of the crucial oxygens across all these enzyme families is well conserved. The only exceptions occur where binding of inhibitors has significantly perturbed the catalytic residues.

From these data, we have computed and evaluated a consensus template that enables automatic searching for possible Ser–His–Asp catalytic triad conformations. When tested against the current dataset of all protein structures, the template correctly identified the known Ser–His–Asp catalytic triads and also located a few putative triads of interest.

3.5 References

Alden R.A., Birktoft J.J., Kraut J., Robertus J.D. & Wright C.S. (1971) Atomic coordinates for subtilisin novo. *Biochem. Biophys. Res. Comm.* **45** p337-342

Bacon D. J. & Anderson, W. F. (1988) A fast algorithm for rendering space-filling molecular pictures. *J. Mol. Graph.* **6** 219–220

Bairoch A. The ENZYME data bank (1993) *Nucleic Acid Res.* **33** 3155–3156.

Bairoch A. & Boeckmann B. (1994) The SWISS–PROT protein sequence data bank: current status. *Nucleic Acid Research* **22** 3578–3580

Bajorath J., Raghunathan S., Hinrichs W. & Saenger W. (1989) Long-range structural changes in Proteinase K triggered by calcium removal. *Nature* **337** 481–484

Barth A., Wahab M., Brandt W. & Frost K (1993) Classification of serine–proteases derived from steric comparisons of their active sites. *Drug design and discovery* **10** 297–317

Barth A., Frost K., Wahab M., Brandt W., Schlader H–D. & Franke R. (1994) Classification of serine proteases derived from steric comparisons of their active site geometry, part II: Ser, His, Asp arrangements in proteolytic and non–proteolytic proteins. *Drug Design and Discovery* **12** 89–111

Bender M.L., Clement G.E., Kezdy F.L., & Heck H.A. (1964) The correlation of the pH dependence and the stepwise mchanism of α–chymotrypsin catalysed reactions. *J. Am. Chem. Soc.* **86** 3680–3689

Bernhard S.A., & Gutfreund H. (1965) The optical detection of transients in trypsin and chymotrypsin–catalysed reactions *Proc. Natl. Acad. Sci. USA* **53** 1238–1243

Bernstein F.C., Koetzle T.F., Williams G.J.B., Meyer E.F.Jr., Brice M.D., Rogers J.R., Kennard O., Shimanouchi T. & Tasumi M. (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112** 535–542

Bielka H., Dixon H. B. F., Karlson P., Liebecq C., Sharon N., Van Lenten E. J., Velick S. F., Vliegenthart J. F. G. & Webb E.C. (1992) *E. C. Enzyme nomenclature 1992: recommendations of the Nomenclature Committee of the International Union Of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes. Nomenclature Committee of the International Union of Biochemistry* Academic Press, Inc., (London) Ltd.

Blow D. M. (1976) Structure and mechanism of chymotrypsin. *Acc. Chem. Res.* **9** 145–152

Blow D. M. (1990) More of the catalytic triad. *Nature* **343** 694–695

Blow D. M., Birktoft J. J. & Hartley B. S. (1969) Role of a buried acid group in the mechanism of action of chymotrypsin. *Nature* **221** 337–340

Bone R., Fujishige A., Kettner C.A. & Agard D.A. (1991*a*) Structural basis for broad specificity in α–lytic protease mutants *Biochemistry* **30** 10388-10398

Bone R., Sampson N. S., Bartlett P. A. & Agard (1991*b*) Crystal structure of α–lytic protease complexes with irreversibly bound phosphonate esters *Biochemistry* **30** 2263–2272

Bourne Y., Martinez C., Kerfelec B., Lombardo D., Chapus C. & Cambillau C. (1994) Horse pancreatic lipase *J. Mol. Biol.* **238** 709–732

Brady L., Brzozowski A. M., Derewenda Z. S., Dodson E., Dodson G., Tolley S., Turkenburg J. P., Christianson L., Huge Jensen B., Norskov L., Thim L. & Menge U. (1990) A serine protease triad forms the catalytic centre of triacylglycerol lipase *Nature* **343** 767–770

Braxton S & Wells J.S. (1991) The importance of a distal hydrogen bonding group in stabilising the transition state in subtilisin BPN' *J. Biol. Chem* **266** 11797–11800

Brooks B., Bruccoleri R., Olafson B., States D., Swaminathan S. & Karplus, M. (1983) CHARMM: A Program for Macromolecular Energy, Minimization, and Molecular Dynamics Calculations *J. Comp. Chem.* **4** 187–217

Brunger A.T., Brooks C.L. & Karplus M. (1984) Stochastic Boundary-Conditions for Molecular-Dynamics Simulations of ST2 Water *Chem. Phys. Lett.* **105** 495–500.

Brzozowski A. M., Derewenda U., Derewenda Z. S., Dodson G. G., Lawson D. M., Turkenburg J. P., Bjorkling F., Huge Jensen B., Patkar S. A. & Thim

L. (1991) A model for interfacial activation in lipases from the structure of a fungal lipase-inhibitor complex *Nature* **351** 491–497

Carter P., Abrahmsen L., Wells J.A. (1991) Probing the mechanism and improving the rate of substrate assisted catalysis in subtilisin BPN' *Biochemistry* **30** 6142–6148

Corey D.R. & Craik C.S. (1992) An investigation into the minimum requirements for peptide hydrolysis by mutation of the catalytic triad of trypsin *J. Am. Chem. Soc.* **114** 1784–1790

Corey D. R., McGrath M. E., Vasquez J. R., Fletterick R. J. & Craik C. S. (1992) An alternative geometry for the catalytic triad of serine proteases *J. Am. Chem. Soc.* **114** 4905–4907

Derewenda U., Brzozowski A. M., Lawson D. M. & Derewenda Z. S. (1992) Catalysis at the interface: the anatomy of a conformational change in a triglyceride lipase *Biochemistry* **31** 1532–1541

Derewenda Z. S., Derewenda U. & Dodson G. G. (1992) The crystal and molecular structure of the *Rhizomucor miehei* triacylglyceride lipase at 1.9Å resolution *J. Mol. Biol.* **227** 818–839

Dixon M. (1953) The effect of pH on the affinities of enzymes for substrates and inhibitors *Biochem J.* **55** 161–170

Drenth J., Hol W.G.J., Jansonius J.N. & Koekoek R. (1972) A comparison of the structures of subtilisin and subtilisin novo *Cold Spring Harbor Symp.* **36** 107–134

Epand R.M. & Wilson I.B. (1962) Evidence for the formation of hippuryl chmotrypsin during the hydrolysis of hippuric acid esters *J. Biol. Chem.* **238** 1718–1723

Fersht A.R. (1972) Conformational equilibria in α– and δ–chymotrypsin *J. Mol. Biol.* **64** 497–509

Fischer D., Wolfson H., Lin S. L. & Nussinov R. (1994) Three–dimensional, sequence order–independent comparison of a serine protease against the crystallographic database reveals active site similarities: Potential implications to evolution and to protein folding *Protein Science* **3** 769–778

Fujinaga M. & James M. N. G (1987) Rat submaxillary gland serine–protease, Tonin. Structure solution and refinement at 1.8Å resolution *J. Mol. Biol.* **195** 373–396

Fujinaga M., Sielecki A.R., Read R.J., Ardelt W., Laskowski M. & James M.N.G. (1987) Crystal and molecular structures of the complex of α–chymotrypsin with its inhibitor turkey ovomucoid domain at 1.8Å resolution. *J. Mol.Biol.* **195** 397–418

Frey P. A., Whitt S. A. & Tobin J. B. (1994) A low–barrier hydrogen bond in the catalytic triad of serine proteases *Science* **264** 1927–1930

James M. N. G., Sielecki A. R., Brayer G. D., Delbaere L. T. J. & Bauer C. A. (1980) Structures of product and inhibitor complexes of *Streptomyces griseus* protease at 1.8Åresolution. A model for serine protease catalysis. *J. Mol. Biol.* **144** 43–52

Gallagher T., Bryan P. & Gilliland G.L. (1993) Calcium–independent subtilisin design. *Proteins. Struct. Funct. Genet.* **16** 205–213

Gutfreund H. & Hammond B.R. (1959) Steps in the reactions of chymotrypsin with tyrosine derivatives *Biochem. J.* **73** 526–530

Harel M., Sussman J. L. & Silman I. (1991) γ–chymotrypsin is a complex of α–chymotrypsin with its own autolysis products *Biochemistry* **30** 5217–5225

Hartley B.S. & Kilby B.A. (1954) The reaction of *p*–nitrophenyl esters with chymotrypsin and insulin *Biochem J.* **56** 288–297

Hayashi R., Moore S. & Stein W. H. (1973) Serine at the active center of Yeast carboxypeptidase II *J. Biol. Chem.* **248** 8366–8369

Ke H. (1992) Similarities and differences between human cyclophilin A and other β–barrel structures. *J. Mol. Biol.* **228** 539–550

Kossiakoff A. A., Chambers J. L., Kay L. M. & Stroud R. M. (1977) Structure of bovine trypsinogen at 1.9Å resolution *Biochemistry* **16** 654–696

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

Kraut (1977) Serine proteases: structure and mechanism of catalysis *Ann. Rev. Biochem.* **46** 331–358

Liao D-I., Breddam K., Sweet R., Bullock T. & Remington S. J. (1992) Refined atomic model of wheat serine carboxypeptidase II at 2.2Å resolution *Biochemistry* **31** 9796–9812

Li De La I., Papamichael E., Sakarelos C., Dimicoli J. L. & Prange T. (1990) Interaction of the peptide CF3–LEU–ALA–NH–C6H4–CF3 with porcine pancreatic elastase. X–ray studies at 1.8Å. *J. Mol. Recog.* **3** 36–44

Mann K. G. (1987) The assembly of blood clotting complexes on membranes *Trends Biochem. Sci.* **12** 229–233

Marquart M., Deisenhofer J., Huber R. & Palm W. (1980) Crystallographic refinement and atomic models of the intact immunoglobulin molecule kol and its antigen–binding fragment at 3.0Å and 1.9Å resolution *J. Mol. Biol.* **141** 369–391

Marquart M., Walter J., Deisenhofer J., Bode W. & Huber R. (1983) The geometry of the reactive site and of the peptide groups in trypsin, trypsinogen and its complexes with inhibitors *Acta Crystallogr., sect. B* **39** 480–490

Matthews B.W., Sigler P.B., Henderson R. & Blow D.M. (1967) Three–dimensional structure of tosyl–α–chymotrypsin *Nature* **214** 652–656

Mathews I., Padmanabhan K., Ganesh V. & Tulinsky A. (1994) Crystal structures of thrombin complexed with thrombin receptor peptides: existence of expected and novel binding modes *Biochemistry* **33** 3266–3279.

McDonald I. K. & Thornton J. M. (1994) Satisfying hydrogen potentials in proteins *J. Mol. Biol.*, **238** 777–793

McGrath M. E., Vasquez J. R., Craik C. S., Yahg A. S., Honig B. & Fletterick R. J. (1992) Perturbing the polar environment of Asp 102 in trypsin: consequences of replacing the conserved Ser 214. *Biochemistry* **31** 3059–3064

Merritt E. A. & Murphy M. E. P. (1994) Raster3D version 2.0. A program for photorealistic molecular graphics. *Acta Cryst. Sec. D* **50** 869–873

Meyer E. F. Junior, Presta L. G. & Radhakrishnan R. (1985) Stereospecific reaction of 3–methoxy–4–chloro–7–aminoisocoumarin with crystalline porcine elastase. *J. Am. Chem. Soc.* **107** 4091–4093

Momany F.A., McGuire R.F., Burgess A.W. & Scheraga H.A. (1975) Energy parameters in polypeptides VII. Geometric parameters, partial atomic charges, non–bonded interactions, hydrogen bond interactions and intrinsic torsional potentials for the naturally occurring amino acids *J. Phys. Chem.* **79** 2373–2381

Noble M. E. M., Cleasby A., Johnson L. N., Egmond M. R. & Frenken L. G. J. (1993) The crystal structure of triacylglycerol lipase from *Pseudomonas*

glumae reveals a partially redundant catalytic aspartate *Febs Lett.* **331** 123-128

Orengo C. A., Flores T.P., Taylor W.R. & Thornton J.M. (1993) Identification and classification of protein fold families *Protein Engineering* **6**, 485–500

Pantoliano M. W., Whitlow M., Wood J. F., Dodd S. W., Hardman K. D., Rollence M. L. & Bryan P. N. (1989) Large increases in general stability for subtilisin *bpn* through incremental changes in the free energy of unfolding. *Biochemistry* **28** 7205–7213

Perona J. & Craik C. (1995) Structural basis of substrate specificity in the serine proteases *Protein Science* **4** 337–360

Pimental G. C. & McClellan A. L. (1960) *The hydrogen bond* Freeman and Co., London 242, 282–288

Polgar L. (1989) Serine proteinases In *Mechanisms of protease action* Boca Raton CRC press 87–113

Powers J. C., Oleksyszyn J., Narasimhan S. L., Kam C. M., Radhakrishnan R. & Meyer E. F. (1990) Reaction of porcine pancreatic elastase with 7–substituted 3–alkoxy–4–chloroisocoumarins: design of potent inhibitors using the crystal structure of the complex formed with 4–chloro–3–ethoxy–7–guanidio–isocoumarin. *Biochemistry* **29** 3108–3118

Radhakrishnan R., Presta L. G., Meyer E. F. & Wildonger R. (1987) Crystal structures of the complex of porcine pancreatic elastase with two valine-derived benzoxazinone inhibitors *J. Mol. Biol.* **198** 417–424

Renard M. & Fersht A.R. (1973) Anomalous pH dependence of k_{cat}/K_m in enzyme reactions. Rate constants for the association of chymotrypsin with substrates *Biochemistry* **12** 4713–4717

Rydel T., Ravichandran K. G., Tulinsky A., Bode W., Huber R., Roitsch C. & Fenton J. W. (1990) The structure of a complex of recombinant hirudin and human α–thrombin *Science* **249** 277–281

Sawyer L., Shotton D. M., Campbell J. W., Wendell P. L., Muirhead H., Watson H. C., Diamond R. & Ladner R. C. (1978) The atomic structure of crystalline porcine pancreatic elastase at 2.5Å resolution. Comparisons with the structure of α–chymotrypsin *J. Mol. Biol.* **118** 137–208

Sprang S., Standing T., Fletterick R. J., Stroud R. M., Finer–Moore J., Xuong N. H., Hamlin R., Rutter W. J. & Craik C. S. (1987) The three dimensional structure of Asn 102 mutant of trypsin. Role of Asp 102 in serine protease catalysis *Science* **237** 905–909

Syed R., Hogle J. M. & Hilvert D. (1993) Crystal structure of selenosubtilisin at 2.0Å resolution *Biochemistry* **32** 6154–6157

Takahashi L. H., Radhakrishnan R., Rosenfield R. E., Meyer E. F. & Trainor D.A. (1989) Crystal structure of the covalent complex formed by a peptidyl α, α–difluoro–β–keto amide with porcine pancreatic elastase at 1.78Å resolution *J. Am. Chem. Soc.* **111** 3368–3374

Takeuchi Y., Nonaka T., Nakamura K. T., Kojima S., Miura K–I. & Mitsui Y. (1992) Crystal structure of an engineered subtilisin inhibitor complexed with bovine trypsin *Proc. Natl. Acad. Sci.* **89** 4407–4411

Takeuchi Y., Satow Y., Nakamura K.T. & Mitsui Y. (1991) The refined crystal structure of the complex of subtilisin and *Streptomyces* subtilisin inhibitor at 1.8Å resolution. *J. Mol. Biol.* **221** 309–325

Tsunasawa S., Masaki T., Hirose M., Soejima M. & Sakiyama F. (1989) The primary structure and structural characteristics of *Achromobacter lyticus*

protease I, a lysine–specific serine proteinase. *J. Biol. Chem.*. **264** 3832–3839

Van Tilbeurgh H., Egloff M. P., Martinez C., Rugani N., Verger R. & Cambillau C. (1993) Interfacial activation of the lipase–procolipase complex by mixed micelles revealed by X-ray crystallography *Nature* **362** 814–820

Wallace A. C., Laskowski R. A. & Thornton J. M. (1995) LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions *Protein Engineering* **8** 127–134

Schoellman G. & Shaw E. (1953) Direct evidence for the role of Histidine in the active site of chymotrypsin *Biochemistry* **2** 252–255

Wei Y., Schottel J. L., Derewenda U., Swenson L., Patkar S. & Derewenda Z. S. (1995) A novel variant of the catalytic triad in the *Streptomyces scabies* esterase. *Nature Structural Biology* **2** 218–223

Wright C. S., Alden R. A. & Kraut J. (1969) Structure of subtilisin *bpn* at 2.5Å resolution. *Nature* **221** 235–242

Xiayang Q., Padmanabhan K., Carperos V., Tulinsky A., Kline T., Maraganore J. & Fenton J. (1992) Structure of the hirulog 3–thrombin complex and nature of the S' subsites of substrates and inhibitors. *Biochemistry* **31** 11689–11697

Chapter 4

TESS: an algorithm for automatically deriving 3D templates for enzyme active sites

4.1 Introduction

The detection of recurring structural motifs or templates in proteins is already well documented, indeed they exist in all levels of protein structure, from primary to tertiary. At the primary level, there are now comprehensive protein sequence databases such as SWISS–PROT (Bairoch & Boeckmann, 1994) and OWL (Bleasby *et al.*, 1994) that have been analysed, using both automatic and manual pattern–matching and sequence alignment techniques, to produce databases of recurring sequence motifs or templates such as PROSITE (Bairoch & Bucher, 1994) and PRINTS (Attwood *et al.*, 1994). At the tertiary level, protein structure analysis, using both automatic (Orengo *et al.*, 1993) and manual techniques, has enabled the creation of databases such as CATH and SCOP (Murzin *et al.*, 1995). These databases are useful for the identification of biological role and prediction of

tertiary structure (see reviews by Taylor, 1988; Hodgman, 1989; Taylor & Jones, 1991).

Many investigations have taken place at the substructural level of proteins, for example analysis of 3D topologies of metal–binding sites in proteins and small molecules (see reviews by Glusker, 1991; Jernigan, 1994). In addition, there are already various algorithms to detect similar 3D arrangements of secondary structure in proteins. Some are comparison techniques that require the linear order of the amino acid sequences to be conserved (Matthews & Rossman, 1985); others allow some degree of insertion/deletion in the protein sequence (Alexandrov *et al.*, 1992) while others match by secondary structure elements (Mitchel *et al.*, 1992). PROMOTIF (Hutchinson & Thornton, 1996) specificially identifies and analyses structural motifs in proteins such as secondary structure, β– and γ–turns and disulphide bridges. Artymiuk *et al.* (1994) have used a graph-theoretic approach for the identification of 3D patterns of amino–acid side–chains in protein structures. As an example, they constructed a search template from the side–chain atoms of the Ser 195–His 57–Asp 102 catalytic triad of chymotrypsin and, depending on the allowed inter–atomic distance tolerances, different numbers of catalytic triads were identified from their dataset. In their method they represent each amino–acid sidechain by two 'pseudo–atoms'. For example the Asp pseudo–atoms are equivalently positioned on C^{β} and the mean of the $O^{\delta 1}$ and $O^{\delta 2}$ atoms. Though this allows orientational factors about two residues to be compared directly, the method is not applicable to our serine proteinase and lipase 3D consensus template example of Ser O^{γ}–His sidechain–Asp $O^{\delta 1}$ as the Ser and Asp consist of only one atom.

A different structural comparison of the serine proteinases, using a less specific technique, has been performed by Fischer *et al.*, (1994). Their method, derived from geometric hashing methods first described by Lamdan *et al.* (1988) for use

in computer vision research, treats all C^{α} atoms in a protein as points in space and compares proteins purely on the geometrical relationships between these points. It can detect recurring substructural 3D motifs and was able to identify the structural similarities of the active sites of the trypsin–like and subtilisin–like serine proteases based solely on the similarities of the C^{α} geometries of their constituent residues.

However, there is not a database of recurring 3D templates or motifs in proteins; these can be thought of as the 3D equivalent of the 1D templates found in the PRINTS and PROSITE databases. The number of protein 3D structures being solved by X–ray crystallography and NMR spectroscopy techniques is increasing rapidly; there are expected to be around 30000 by the turn of the century (see PDB world wide web page http://www.pdb.bnl.gov/statistics.html). This suggests that the need for a 3D equivalent of PROSITE is also growing; this would enable us to suggest functions of proteins whose roles are unknown as well as allow us to locate functional regions and catalytic residues within the protein structure. Such databases could address many different substructural aspects of proteins, such as enzyme active sites, ligand binding sites, loop conformation and metal binding sites. Here, we concentrate on enzyme active sites.

In chapter 3 we showed that a Ser–His–Asp 3D enzyme active site template can be defined that will identify all the serine proteinases and lipase active sites in a database of PDB structures with the exclusion of all other non–catalytic Ser, His and Asp interactions (Wallace *et al.*, 1996). This suggests that enzyme active sites in general may have a unique conformation when compared to non–catalytic regions of a protein; indicating that we could create a database of unique enzyme active site templates. To test this, we need a fast generalised 3D template database search tool that, in analogy to the method that was used to create the Ser–His–Asp 3D template, is able to take a 'seed' 3D template and automati-

cally search through a database of PDB structures for residues with the same 3D conformation.

The method we used to generate the Ser–His–Asp 3D template is not as suitable as a generalised 3D search method for several reasons. Firstly, it required all Ser, His and Asp interactions to be extracted by means of a user defined distance cut–off, in this case at least one interatomic contact is less than the sum of the van der Waals radii of the two atoms plus 1Å. This provided an initial filter that removed all Ser, His and Asp interactions that are not within a reasonable contacting distance of one another. However, it is not applicable to the case where the catalytic residues are lying far apart in the enzyme active site as the run–time and number of interactions output by DISTRIB would increase greatly. For example, the Glu 11–Asp 20 catalytic residues of T4 lysozyme (Weaver *et al.*, 1987) lie around 9Å apart and are bisected by the substrate in the active site. In addition, the method is not automatic to run as the output of the program DISTRIB also requires a filtering routine to extract the relevant catalytic triads and then calculate the coordinates of the 3D consensus template.

Here, an algorithm called TESS (**te**mplate **s**earch and **s**uperposition) will be described that overcomes these methodological shortcomings. TESS is a program that allows the user to search through a given dataset of PDB structures for any combination of residues or atoms in 3D space irrespective of the position of those residues in the protein's sequence. TESS will allow us to generate any enzyme active site 3D consensus template from the structures in the PDB, as long as the catalytic residues have been identified. The run time of TESS is fast; it takes around 0.25 CPU seconds to search through a typical PDB structure on an SGI Challenge and the run time proportional to the order, $O(nh)$, where n is the number of atoms present in the template and h the number of hits. TESS is also easy to use; it requires only a user defined query template which can be extracted

directly from a PDB file, and a dataset of PDB files to be searched.

TESS is similar in method to that of Fischer *et al.* (1994) as it is also based on the geometric hashing paradigm. The difference is that TESS is not confined to just the C^{α} atoms in a protein as it can search any user defined combination of atoms in space from single atoms to multiple residues. TESS is more flexible than the graph–theoretic approach used by Artymiuk *et al.* (1994) as it is also able to search for single atoms in space and not the whole sidechain.

The TESS algorithm comprises two major steps. Firstly, it is necessary to calculate the geometric relationships between all the atoms in each PDB file; this information is stored in a so called TESS table and therefore needs only to be calculated once. The TESS table is designed to use the minimum amount of computer memory possible combined with fast access and processing. The second part of the algorithm compares the geometric relationship between atoms in a user defined 3D template and the PDB structures stored in the tables. The 3D template is a list describing the atom types and their geometric positions that are to be searched for in the TESS table. In addition the user is able to define the allowed distance deviation between the template atoms and the atoms in the search structure. In order to assess the accuracy and efficiency of TESS we have tested it against a large number of typical structure comparisons. The algorithm is described in detail in the following sections.

4.2 Method

The TESS algorithm needs to deal with the following problem: given the 3D coordinates of two molecules, one the 3D query template and the other the protein structure, find the transformation, if any, that will best superimpose the 3D query template onto the protein molecule. This means that the atoms, residue types

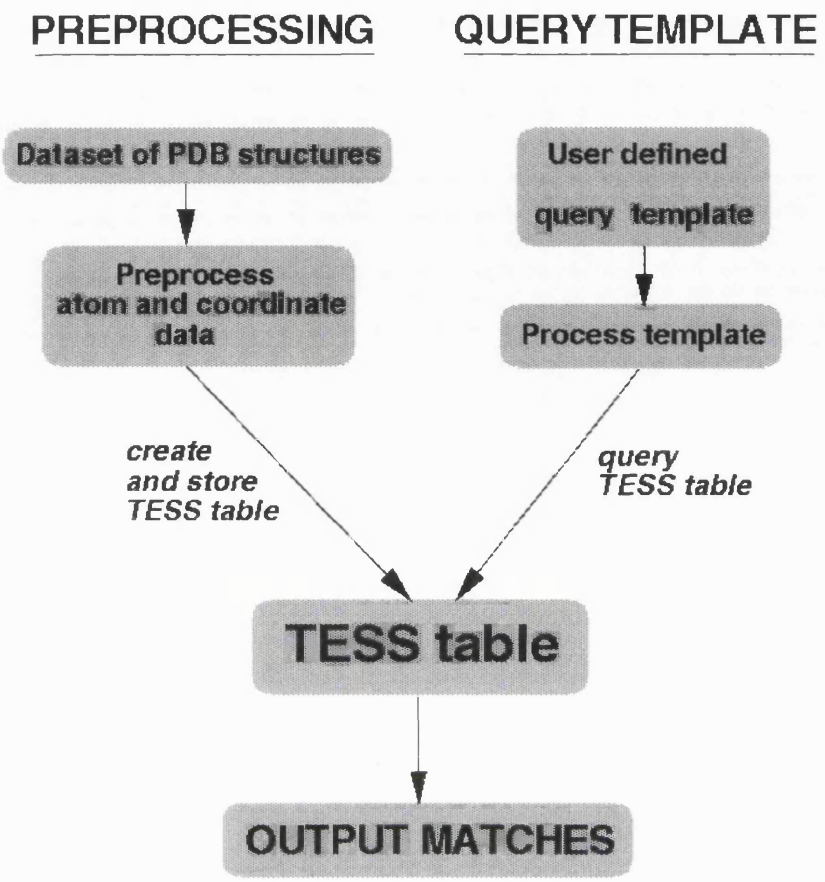


Figure 4.1: A summary of the process involved when TESS searches through a dataset of PDB structures for a user defined 3D template.

and the 3D coordinates of the 3D query template need a partial overlap (defined by a distance cut-off) with those of the stored PDB structure. The method we have devised to solve this problem is summarised in Figure 4.1.

Firstly, the preprocessing stage extracts all the relevant information from the dataset of PDB structures and stores it in the TESS table. Secondly, the user defines a query template which is processed and compared with the information stored in the TESS table. TESS then automatically outputs any matches that occur. The details of these two stages are described in the following sections.

4.2.1 Stage 1

Pre-processing the structures in the PDB

In the preprocessing stage, we represent information about the atoms of all the structures in the PDB in a TESS table. Relevant information about a protein structure in the PDB would be readily available and does not have to recalculated each time TESS is run; this speeds up the comparison process considerably.

To locate the position of a given 3D template, for example the Ser O^{γ}–His sidechain–Asp O^{δ} consensus template of the serine proteinases and lipases (Wallace *et al.*, 1996), in a dataset of PDB structures we need to represent the atoms of the template and the PDB structures with respect to a reference frame. If we pick the same reference frame for both the template and the PDB structures, then a match will occur when a PDB structure has a substructure matching the 3D template. The reference frames we use are the sidechain atoms of the 20 standard amino acids. Specifically, we use three atoms for each residue as listed in Table 4.1 and defined by Singh & Thornton (1992).

Of course, using one of the amino acid sidechains as a reference frame means that any 3D template must consist of at least one amino acid sidechain with the other atoms/residues surrounding it. This may appear to be a major constraint in defining any potential enzyme active site template. However, we saw in Chapter 3 that the Ser–His–Asp 3D template is defined as the position of the Ser O^{γ} and Asp O^{δ} with respect to the His sidechain reference frame (Wallace *et al.*, 1996).

A summary of the process involved in creating a TESS table is shown in Figure 4.2 using a His residue as an example. The first stage involves reading in the 3D coordinates of the first protein in the PDB dataset. For each of the His residues in the protein structure, all atoms within 18Å are identified. The tranformation matrix is calculated that places each His residue C^{γ} at the origin, with the C^{δ_2}

RESIDUE	1	2	3	RESIDUE	1	2	3	RESIDUE	1	2	3
ALA (A)	N	C^{α}	C^{β}	GLY (G)	N	C^{α}	C	PRO (P)	N	C^{α}	C^{β}
ARG (R)	$N^{\eta 1}$	N^{ϵ}	$N^{\eta 2}$	HIS (H)	$C^{\delta 2}$	C^{γ}	$N^{\delta 1}$	SER (S)	C^{α}	C^{β}	O^{γ}
ASN (N)	$O^{\delta 1}$	C^{γ}	$N^{\delta 2}$	ILE (I)	$C^{\gamma 1}$	C^{β}	$C^{\gamma 2}$	THR (T)	$O^{\gamma 1}$	C^{β}	$C^{\gamma 2}$
ASP (D)	$O^{\delta 1}$	C^{γ}	$O^{\delta 2}$	LEU (L)	$C^{\delta 1}$	C^{γ}	$C^{\delta 2}$	TRP (W)	$C^{\delta 1}$	C^{γ}	$C^{\delta 2}$
CYS (C)	C^{α}	C^{β}	S^{γ}	LYS (K)	C^{ϵ}	C^{δ}	N^{ζ}	TYR (Y)	$C^{\delta 1}$	O^{η}	$C^{\delta 2}$
GLN (Q)	$O^{\epsilon 1}$	C^{δ}	$N^{\epsilon 2}$	MET (M)	C^{ϵ}	C^{γ}	S^{δ}	VAL (V)	$C^{\gamma 1}$	C^{β}	$C^{\gamma 2}$
GLU (E)	$O^{\epsilon 2}$	C^{δ}	$O^{\epsilon 1}$	PHE (F)	$C^{\epsilon 1}$	C^{γ}	$C^{\epsilon 2}$				

Table 4.1: Sidechain atoms used to define the reference frames for each standard amino-acid, as defined by Singh & Thornton (1992). The atoms in column 2 are transformed to the origin with the atoms in column 1 and 3 either side of the positive x direction.

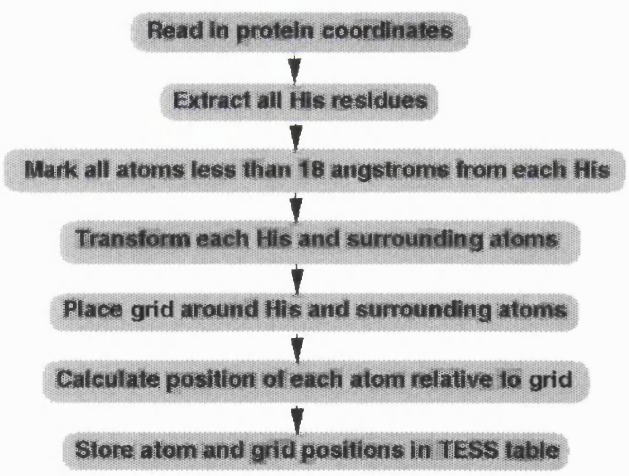


Figure 4.2: A flow diagram showing the steps required in producing a TESS table in the pre-processing stage. The example given is for generation of templates involving His sidechains. This process is repeated for each of the proteins in the PDB which results in a TESS table for the His reference residue.

and $N^{\delta 1}$ either side of the positive x direction. The same transformation matrix is applied to the atoms surrounding each His residue, giving us the reference His sidechain at the origin, surrounded by its neighbouring atoms. We now need a method to describe the relative position of the atoms around each His. This is achieved by placing a grid around each His and its neighbouring atoms and calculating the box position within the grid that each atom occupies. The atoms in each grid are then reordered according to their box number and assigned an 'atom identifier' according to the residue and atom type. This information, along with the PDB file atom numbers, is stored in the TESS table.

The whole process is repeated for each protein structure in the PDB and a marker for the position of each PDB structure in the TESS table is also stored, which allows a quick call by reference parsing for any given protein structure in the TESS table. This, of course, only deals with His environments; it is repeated for each of the 20 standard amino acids that are listed in Table 4.1, giving us 20 TESS tables, one for each of the standard amino acids.

4.2.2 Stage 2

Defining, Processing and Comparing the Query 3D template

Defining the Query Template

We need a method to compare a 'query template' with the position of the atoms of the PDB structures stored in the TESS tables created in stage 1. The query template is transformed to the appropriate standard reference frame (e.g. His) and then compared with the relevant TESS table. Both the geometric positions and atomic labels (atom and residue type) of the query template and a PDB structure's atoms are compared; a 'hit' occurs when all the atoms of the query

SEARCH OPTION	ATOM NUMBER
TEMPLATE ATOM	-1
SEARCH BY ATOM TYPE i.e. $O^{\delta 1}$, $O^{\delta 2}$, O^{γ}	0
SEARCH BY NON–CARBON SIDECHAIN ATOM	1
SEARCH BY NON–CARBON ATOM	2
SEARCH BY SPECIFIED ATOM i.e. C, O, N	3
SEARCH BY NON–CARBON MAINCHAIN ATOM I.E. ' O ', ' N '	4
SEARCH BY ANY MAINCHAIN ATOM	5
SEARCH BY ANY SIDECHAIN ATOM	6
SEARCH BY ANY ATOM TYPE	7

Table 4.2: Search parameter numbers placed in the atom number column of the query PDB format file. One of these numbers is placed against each of the atoms in the query template. This defines which atom types are to be searched for at the corresponding atom position. To search for different residue types at a given atom point requires the one letter code of that amino acid to be placed after the coordinates in the query template file.

template have a corresponding match with the atoms of a PDB structure.

To perform such a search, we set up a query template which is a slightly modified PDB format file constituting the atoms or residues which are to be searched for in the TESS table. For a given atom in the query template, it is possible to define any combination of both atom and residue types which are to be searched for at that point. To define which residues are to be located at a given template atom position requires the 1 letter amino acid code corresponding to that amino acid to be placed after the coordinates of the query template. The residue in the residue column of the file is searched for by default. The atom type to be searched for is defined by one of the numbers in Table 4.2.

For example, it has been observed that different enzymes have similar catalytic machinery in their active sites even though they catalyse reactions on different substrates. We saw in Chapter 3 that both the serine proteinases and lipases have a Ser–His–Asp catalytic triad. In addition, haloalkane dehalogenase (Franken *et*

Search Option	Residue	Residue Number	Atom	x	y	z	Residue Search
1	Ser	195	O^{γ}	16.3	30.6	14.7	D
1	Asp	102	O^{δ_2}	18.2	31.5	20.8	E
0	His	57	C^{β}	14.5	28.8	20.9	
-1	His	57	C^{γ}	15.0	29.3	19.5	
-1	His	57	N^{δ_1}	16.2	30.0	19.3	
-1	His	57	C^{δ_2}	14.3	29.1	18.3	
0	His	57	C^{ϵ_1}	16.2	30.3	18.0	
0	His	57	N^{ϵ_2}	15.1	29.8	17.4	

Table 4.3: An example of a typical query template which is taken from the active site of α–lytic proteinase, *1lpr*.

al., 1991) has a catalytic triad consisting of Asp–His–Asp and acetylcholinesterase a Ser–His–Glu triad (Harel *et al.*, 1993). These 3 triads are similar in that they all comprise a His–Asp or His–Glu acid/base catalyst (sidechain oxygen electrostatic atom) and a nucleophilic Ser or Asp residue (sidechain oxygen nucleophilic atom), so the triad we wish to search for can be thought of as nucleophilic O–His–electrostatic O.

To investigate the structural similarity of these 3 triads, we use the coordinates of one of the triads as a query template and compare the other two triads with it. In this case, the query template is derived from the serine proteinase α–lytic proteinase, *1lpr* (Bone *et al.*, 1991), as shown in Table 4.3. A query template needs a reference frame residue, in this case a His, by putting a '-1' in the atom number column corresponding to, C^{δ_2}, C^{γ} and N^{δ_1}. This enables comparison of the relative positions of the atoms of the query template and the PDB structures in the TESS table with respect to the same reference frame.

If we want to search for any of the 3 catalytic triads, as defined by the nucle-

ophilic O–His–electrostatic O template, we use this Ser–His–Asp triad but put a '1' in the atom number column corresponding to the Ser O^{γ} and Asp $O^{\delta 2}$. This number corresponds to a 'search parameter' listed in Table 4.1. The '1' next to the Asp $O^{\delta 2}$ means there will be a search for any non–carbon sidechain atom of residue type Asp (i.e. the atoms $O^{\delta 1}$ and $O^{\delta 2}$) at that coordinate point. Secondly, a D (Asp 1 letter code) is put after the Ser O^{γ} coordinates, enabling a search for nucleophilic Asp of haloalkane dehalogenase. Finally, E (Glu 1 letter code) is placed after the Asp $O^{\delta 2}$, which allows a search for the electrostatic Glu of acetylcholinesterase. It is also possible to define a distance cut–off for matching two atoms; the default is 2Å.

Processing Query Template

The 3D query template is processed so its atoms can be compared to the TESS table created in the preprocessing stage. The method is similar to that used to create the TESS tables, though considerably quicker due to the relatively small number of atoms involved.

As an example, Table 4.3 shows the coordinates and atom identifiers of the Ser O^{γ}–His–Asp $O^{\delta 2}$ 3D query template. TESS calculates the transformation matrix that places His reference atoms (from Table 4.1, $C^{\delta 2}$, C^{γ} and $N^{\delta 1}$) of the 3D query template at the origin. The same matrix is applied to the other atoms in the query file. This means that the His reference frame is in the same position as the His reference residues for all the PDB files in the table. The same transformation matrix is applied to all the other atoms in the query template, giving us the His reference frame at the origin surrounded by all the other template atoms. The relative position of the atoms around the query His residue are also calculated in the same way as the preprocessing stage whereby a grid is placed around the 3D template atoms and the box numbers of each of the atoms in the grid are

calculated.

Comparing the Query Template with the PDB structures

TESS now compares the labels and positions of the atoms in the query template with those stored in the TESS tables. Figure 4.3 illustrates the comparison process between the query template and a PDB structure stored in the TESS table. The protein is shown to have a Ser–His–Asp triad. The relative positions of these Ser, His and Asp atoms in the grid have been calculated and stored in the TESS table as described in the preprocessing section. This protein has 3 other His residues and these will also have been preprocessed and their environments stored in the TESS table. The grid positions of the query template atoms are compared with those of the TESS table. In Figure 4.3, a match has occured between the query template and the PDB structure in terms of grid positions and search parameters (atom, residue types and user defined distance cut–off). For each match, the matching PDB file is opened and the relevant atomic coordinates are transformed to the same reference frame. We can now calculate the root mean square (rms) distance of the 3D query template and the transformed atoms of the PDB file. The atoms of the central sidechain are not included in this calculation because they are superimposed. If any of the distances between the equivalent atoms of the 'seed' template and the PDB structure are greater than the user–defined distance cut–off then the match is discarded.

4.3 Performance of the TESS algorithm

The number of X–ray and NMR structures is expected to increase to about 30000 by the turn of the century. It is therefore important that any search algorithm such as TESS should have a search time as near as possible to $O(nh)$, where n

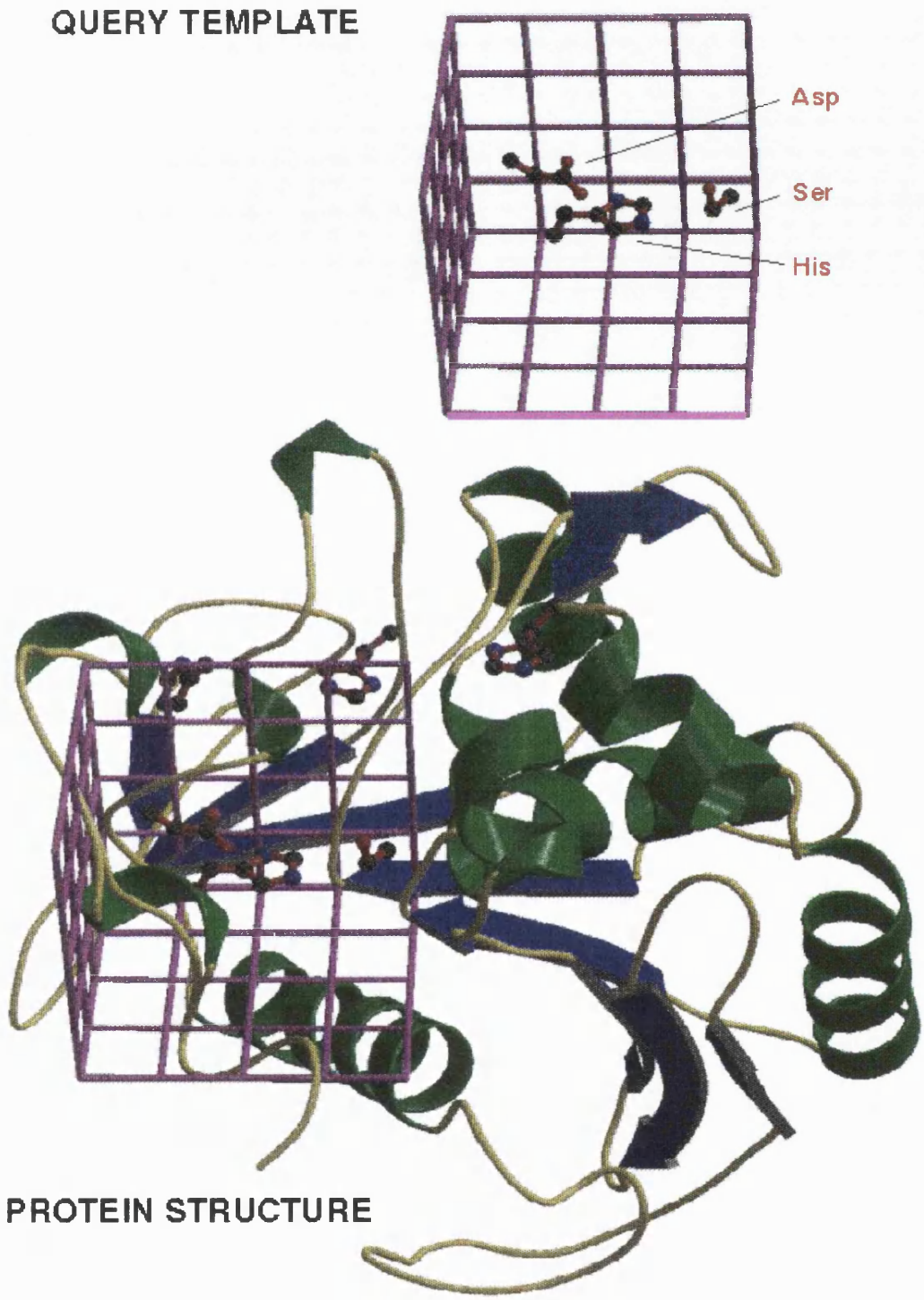


Figure 4.3: A diagram illustrating the comparison process that occurs when a 3D query template is parsed against the TESS table.

is the number of atoms in the query template and h the number of hits located. This will mean that the search time will not be adversely affected by the large increase in data.

There are various parameters that may affect the search speed of TESS; the box size of the grid, the atom or residue search parameters of the query template and the user defined distance cut–off; here I describe how TESS was optimised with respect to these variables.

4.3.1 Optimising the box size

When comparing the atoms of the PDB structures in the TESS table and those of the 3D query template, the situation is slightly complicated because it is not possible to tell where the 3D template atoms are lying in the query boxes; they may be very near to one of the sides. Therefore, if the given distance cut–off is equal to grid box size, we in fact have to search for 27 box numbers in the TESS table per atom of the query template (i.e. all neighbouring boxes to the central box). Of course, if the distance cut–off is slightly larger than the box size we would have to check two neighbouring layers or 125 box numbers per query template atom. Therefore, the grid box size needs to be optimised so that the smallest number of boxes will normally be searched and each box should only contain a few atoms. This suggests that the box size needs to be about the same size as the distance cut–off.

To find the optimum box size, the run time was measured for TESS with different box sizes and search parameters. The template used was the Ser 195 O^{γ}–His 57–Asp 102 O^{δ_2} from α–lytic proteinase with the 228 PDB structures of the serine proteinase and lipase dataset. TESS was run four times for each test box size, each run giving an increasing number of hits, due to different template search parameters. The latter templates are listed in Table 4.4. Each search was

in fact performed first with a cut–off of 2Å and then a 3Å cut–off; the results are listed in Table 4.5.

The results show that if the box size is a factor of the *rms* distance cut–off, then the runs times are quicker (bold figures). When this occurs, TESS searches a distance which is exactly equal to the *rms* cut–off. A 1.9Å box size is therefore particularly slow because TESS in fact searches a 3.8Å distance even when the distance cut–off is 2Å. This increases the multicombinatorial search time needed to filter the atoms in the boxes and produce the template hits. In summary, a box size of 1Å is best for 2Å and 3Å distance cut–offs; however, at different distance cut–offs, other box sizes may be better.

4.3.2 The run–time of TESS

We now investigate how the run time of TESS depends on the number of atoms in the 3D query template or the number of matches located in the TESS table.

Figure 4.4 is a plot of run–time of TESS against number of hits located, with the data taken from the 1.0Å box size of Table 4.2. The best–fit line through the points has a correlation coefficient of 0.79, indicating the run–time is close to $O(h)$, where h is the number of query template hits.

Finally, the run time of TESS was tested against the number of atoms in the query template. The atoms in Table 4.6 constitute the query template used, it was taken from the active site region of the chymotrypsin structure *4gch* (Stoddard *et al.*, 1990). The dataset of proteins used to parse the table comprised 22 chymotrypsin PDB structures, giving a total of 25 chains and therefore 25 catalytic triads. The user distance cut–off was set at 3Å. Figure 4.5 is a graph of the CPU time in seconds against the number of template atoms. The best–fit line through the points has a correlation coefficient of 0.97 indicating the run–time is near to n, the number of template atoms.

Search Option	Residue	Residue Number	Atom	x y z	Residue Search
RUN 1					
0	Ser	195	O^{γ}	16.3 30.6 14.7	
1	Asp	102	O^{δ_2}	18.1 31.5 20.8	
RUN 2					
0	Ser	195	O^{γ}	16.3 30.6 14.7	
5	Asp	102	O^{δ_2}	18.1 31.5 20.8	ARNDCQGHILMFPSTWYV
RUN 3					
5	Ser	195	O^{γ}	16.3 30.6 14.7	
5	Asp	102	O^{δ_2}	18.1 31.5 20.8	ARNDCQGHILMFPSTWYV
RUN 4					
5	Ser	195	O^{γ}	16.3 30.6 14.7	ARNDCQGHILMFPSTWYV
5	Asp	102	O^{δ_2}	18.1 31.5 20.8	ARNDCQGHILMFPSTWYV
TEMPLATE RESIDUE					
0	His	57	C^{β}	14.5 28.8 20.9	
-1	His	57	C^{γ}	15.0 29.3 19.5	
-1	His	57	N^{δ_1}	16.2 30.0 19.2	
-1	His	57	C^{δ_2}	14.3 29.2 18.3	
0	His	57	C^{ϵ_1}	16.2 30.3 17.9	
0	His	57	N^{ϵ_2}	15.1 29.8 17.4	

Table 4.4: The four templates used for optimisation of the TESS box size. Each run had the same His template residue but the Ser and Asp atoms had different search parameters, giving different numbers of hits.

2.0Å distance cut–off														
		Box size (Å)												
Template	No. hits	0.8	0.9	**1.0**	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	**2.0**
1	238	45.8	44.9	**42.2**	43.1	42.3	45.9	44.8	46.9	47.7	49.6	50.4	52.4	**43.5**
2	460	49.6	50.3	**45.0**	46.1	47.3	51.5	52.3	58.4	58.0	79.9	90.8	384.9	**58.9**
3	1080	58.5	64.0	**50.1**	52.8	56.9	69.5	77.1	93.8	146.5	312.3	401.7	-	**58.4**
4	1840	96.9	157.5	**72.9**	73.8	108.9	179.5	302.1	-	-	-	-	-	**107.2**
3.0Å distance cut–off														
		Box size (Å)												
Template	No. hits	0.8	0.9	**1.0**	1.1	1.2	1.3	1.4	**1.5**	1.6	1.7	1.8	1.9	2.0
1	417	50.9	50.2	**47.9**	51.1	69.8	53.4	55.4	**51.4**	50.6	51.4	55.8	52.6	54.0
2	1347	65.6	79.4	**54.2**	76.3	123.0	466.6	668.0	**59.1**	58.6	62.9	61.7	389.3	880.7
3	5734	323.8	-	**85.3**	408.3	557.4	-	-	146.8	**163.7**	-	-	-	-
4	17217	-	-	**654.7**	-	-	-	-	-	-	-	-	-	-

Table 4.5: Run times (CPU seconds) to find the optimum box size for TESS. Those runs left blank take over 1000 CPU seconds; the figures on bold are the quickest runs

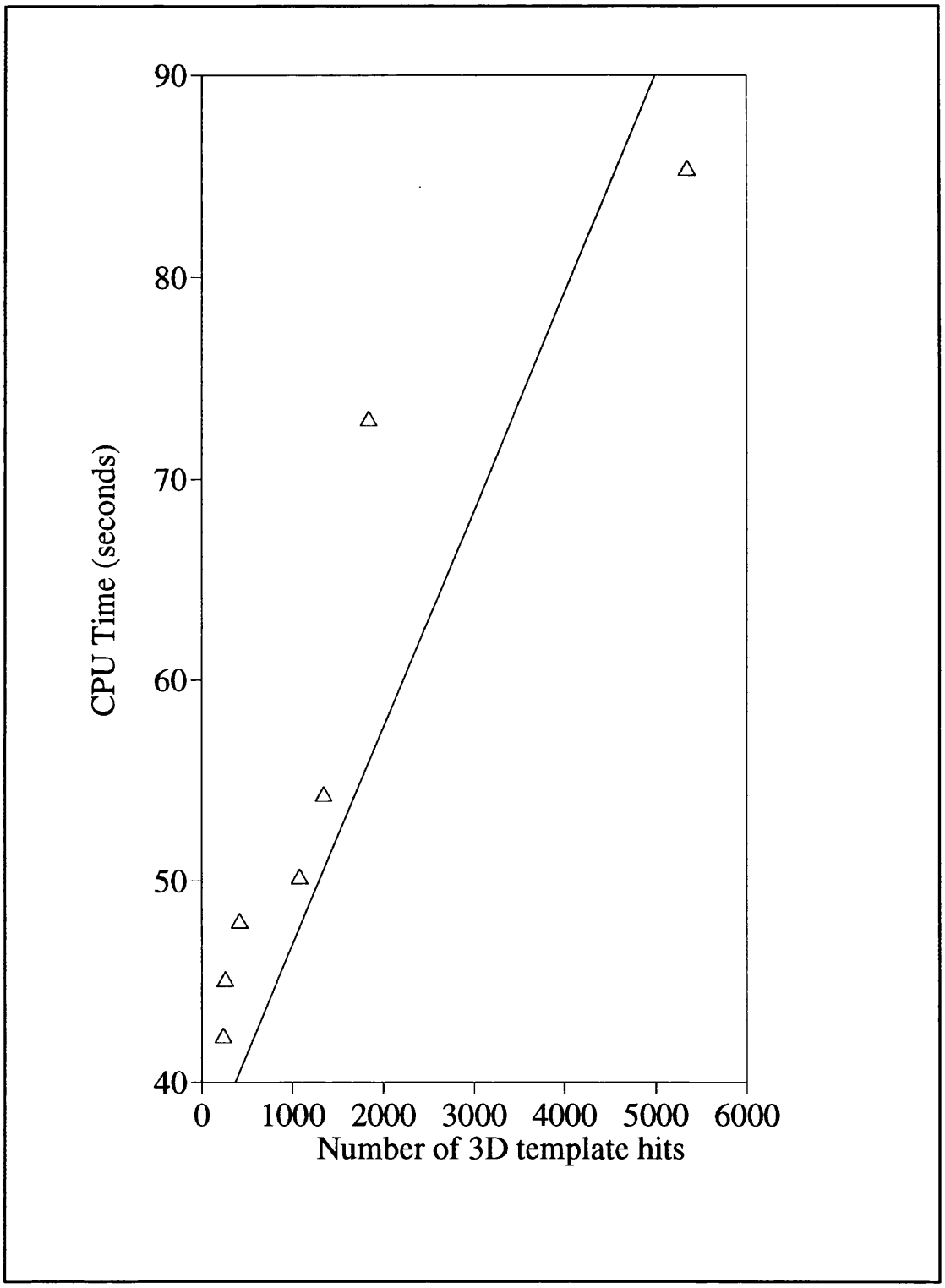


Figure 4.4: A plot of run time in CPU seconds against number of hits showing the run time of TESS is near to Oh, where h is the number of hits

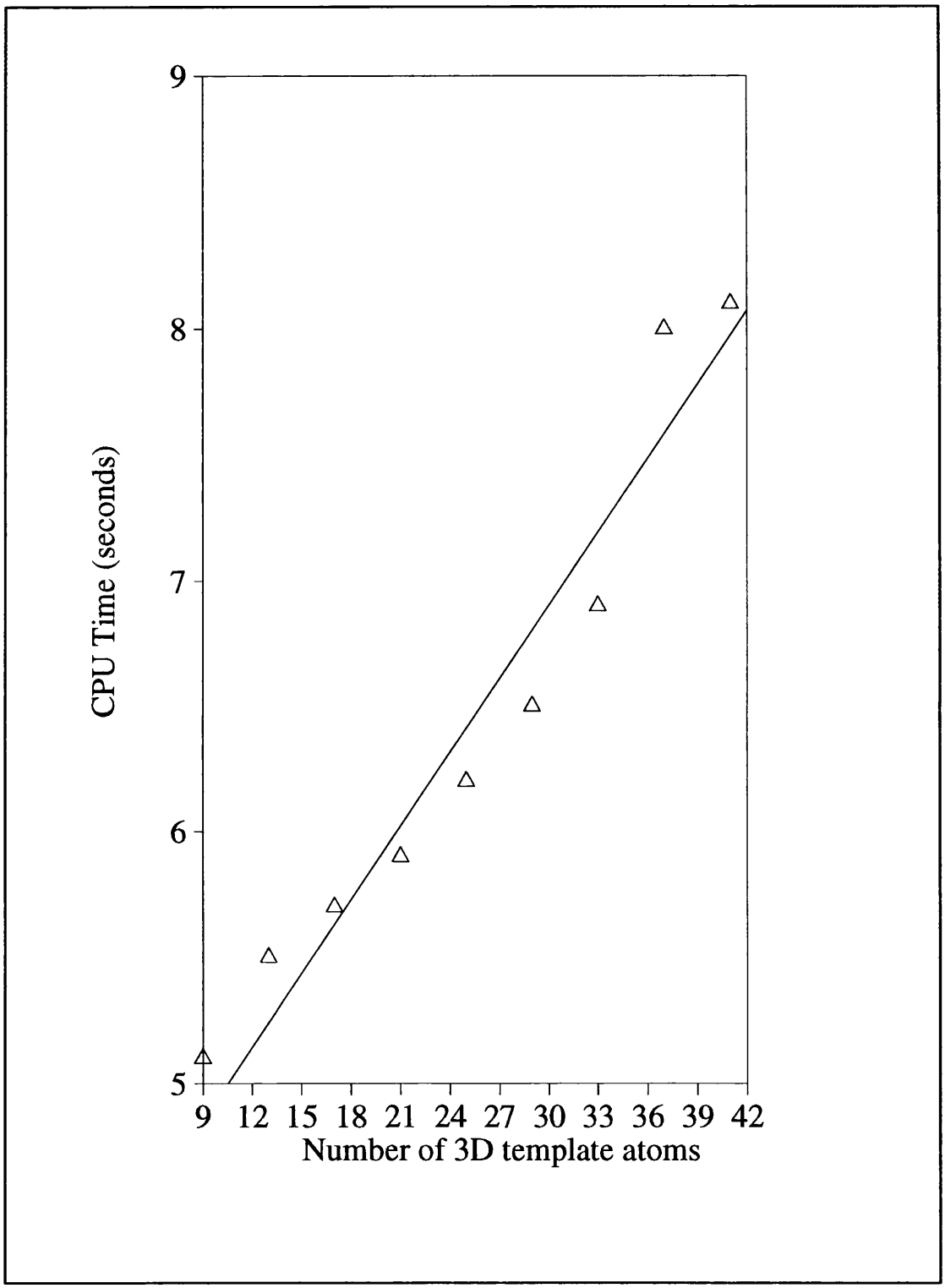


Figure 4.5: A plot of run–time in CPU seconds against number of template atoms.

Atom number	Atom	Residue	Residue number	x	y	z	Atom number	Atom	Residue	Residue number	x	y	z
6	N	ALA	56	41.6	72.8	79.0	20	C^{β}	CYS	58	37.0	75.4	78.9
6	C^{α}	ALA	56	41.0	71.6	78.4	21	S^{γ}	CYS	58	37.4	76.7	80.1
6	C	ALA	56	39.6	71.3	79.0	22	N	ASP	102	45.8	70.7	81.5
6	O	ALA	56	38.7	70.9	78.1	23	C^{α}	ASP	102	44.6	71.4	81.2
6	C^{β}	ALA	56	42.0	70.5	78.3	24	C	ASP	102	44.7	72.7	80.4
6	N	HIS	57	39.5	71.5	80.3	25	O	ASP	102	44.5	73.9	80.9
7	C^{α}	HIS	57	38.2	71.1	80.9	26	C^{β}	ASP	102	44.0	71.6	82.7
8	C	HIS	57	37.2	72.1	80.4	27	C^{γ}	ASP	102	42.6	72.0	82.6
9	O	HIS	57	36.0	71.8	80.6	28	$O^{\delta 1}$	ASP	102	42.2	72.7	83.6
10	C^{β}	HIS	57	38.3	70.9	82.4	29	$O^{\delta 2}$	ASP	102	42.0	71.8	81.6
-1	C^{γ}	HIS	57	38.2	72.3	83.1	30	N	SER	195	37.8	80.1	84.3
-1	$N^{\delta 1}$	HIS	57	39.4	72.9	83.5	31	C^{α}	SER	195	38.6	79.1	83.7
-1	$C^{\delta 2}$	HIS	57	37.2	73.1	83.3	32	C	SER	195	38.9	79.3	82.2
14	$C^{\epsilon 1}$	HIS	57	39.1	74.1	83.7	33	O	SER	195	38.2	79.7	81.4
15	$N^{\epsilon 2}$	HIS	57	37.	74.3	83.7	34	C^{β}	SER	195	37.8	77.8	83.8
16	N	CYS	58	37.5	73.2	79.8	35	O^{γ}	SER	195	38.7	76.9	84.5
17	C^{α}	CYS	58	36.4	74.1	79.4	36	N	SER	214	42.5	76.8	86.3
18	C	CYS	58	35.5	73.3	78.4	37	C^{α}	SER	214	42.0	75.5	86.0
19	O	CYS	58	34.3	73.6	78.4	38	C	SER	214	41.1	75.1	87.0
39	O	SER	214	39.9	74.7	86.7	40	C^{β}	SER	214	43.2	74.6	85.7
							41	O^{γ}	SER	214	43.1	73.3	86.2

Table 4.6: The template used to investigate how the run–time of TESS depends on the number of atoms in the query template.

4.3.3 Memory usage and the TESS tables

Each TESS table uses a relatively large amount of computer memory; around 70 mega bytes for the 3019 structures found in the January 1995 PDB. The amount of memory used is proportional to l^3, where l is the length of the grid side, in our case 36Å. This is an example of a time–space trade–off. If there is unlimited memory then every piece of information about an atom could be stored at a unique address making the algorithm extremely fast; alternatively, if there is little memory then a program would have to re–calculate all the information about the PDB structure atom positions, wasting time.

When searching enzyme active sites, it is not really necessary to generate all 20 amino acid TESS tables. Zvelebil & Sternberg (1988) analysed the constituent amino acids in the active sites of enzymes and found that His is present in around 30% of them with Asp, Glu, Asn and Arg also relatively common. This means that for enzyme searches TESS tables only have to be generated for these reference sidechains. In addition, if we try, where possible, to use the same reference

sidechain amino acid it allows us to compare the active site consensus templates from different enzymes.

4.3.4 Creating a mean 3D consensus template

In Chapter 3 we showed (Wallace *et al.*, 1996) that the derivation of the Ser–His–Asp 3D consensus template initially uses a 'seed' template. In that case the Ser 195 O^{γ}–His 57–Asp 102 $O^{\delta 2}$ from α–lytic proteinase (Bone *et al.*, 1991) was used to extract other catalytic Ser–His–Asp interactions from the output of DISTRIB. This was followed by an iterative procedure that calculated the mean consensus template from all the extracted templates that was then tested against a general dataset of non–identical proteins (compiled so they had a sequence identity greater than 95%) to see if the resultant Ser–His–Asp catalytic template is unique to the serine proteinases and lipases.

This procedure has now been generalised into a procedure called TESSPLATE. TESSPLATE enables a seed enzyme active site template from any given structure in the PDB to be evaluated automatically for its potential as a 3D template. The seed template and its associated dataset of structures (for example the *1lpr* Ser 195 O^{γ}–His 57–Asp 102 $O^{\delta 2}$ seed template and the dataset of serine proteinases and lipases) are processed by TESS, outputing a list of matches. These matches are averaged, creating a 3D consensus template which is then tested automatically against the 95% by sequence non–homologous protein dataset.

To test the validity of the TESSPLATE procedure, the results of the Ser 195 O^{γ}–His 57–Asp 102 $O^{\delta 2}$ α–lytic proteinase 'seed' template used to create the serine proteinase and lipase 3D consensus template (Wallace *et al*, 1996), can be compared with the output automatically from TESSPLATE. The similarity obtained is illustrated in Figure 4.6.

The His residues from the two templates have been superimposed and the

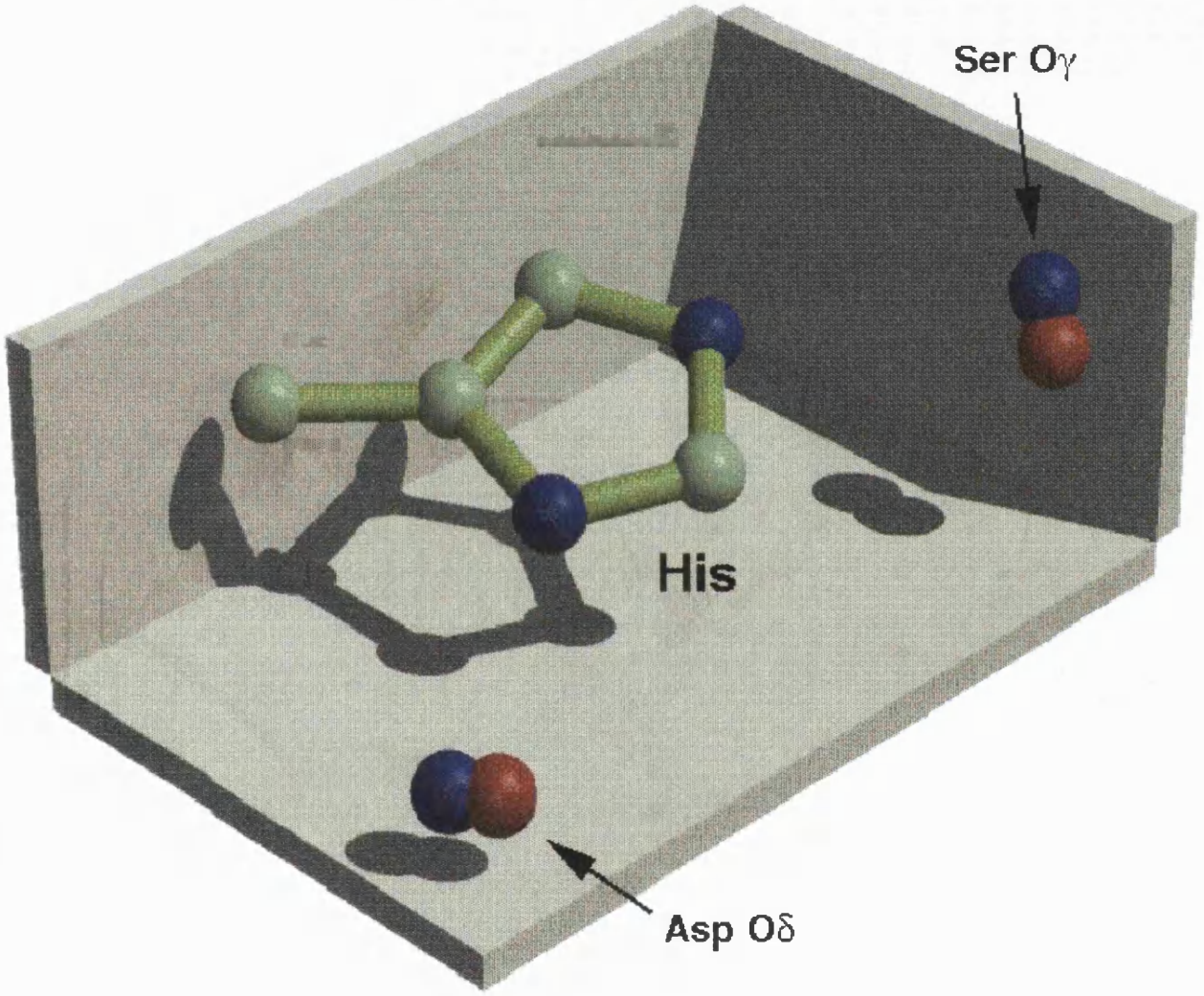


Figure 4.3: A diagram illustrating the similarity in the 3D consensus templates produced in the TESSPLATE (red) and method of Chapter 3 (Wallace *et al.*, 1996) (blue).

rms distance between them is 0.22Å, which is below the estimated errors of an average X-ray crystal structure. The slight discrepancy observed occurs because the method in Chapter 3 divided the enzymes into groups according to their tertiary fold, whereas here the mean consensus template is calculated without taking this into consideration.

The 3D consensus template produced by TESSPLATE is automatically checked against the 95% by sequence non-identical protein dataset. Several hits were serine proteinases or lipase structures with the Ser-His-Asp catalytic triad. In addition, there were 2 non-catalytic hits located; immunoglobulin $2ig2$ (Marquart *et al.*, 1980) with *rms* 1.49 and cyclophilin $2cpl$ (Ke, 1992) with *rms* 1.72. These hits were also located in the previous method, and have already been discussed in detail.

4.4 Discussion

We have developed an algorithm called TESS based on the geometric hashing paradigm that enables us to search through a dataset of 3D PDB structures for any user defined sequence-order independent 3D template. The 3D template consists of atoms or residues extracted directly from a PDB file, and it is possible to define explicitly which atoms or residue types at every 3D template atom point the user wishes to search for. This is a very useful feature as it allows TESS to locate geometric similarities between the conformation of the active sites of enzymes with diverse biological functions.

The TESS algorithm works by first storing relevant information about the protein structures in the PDB in TESS tables. These tables are then queried by a given 3D template and a match occurs if the 3D coordinates and parameters match a substructure of one of the PDB structures stored in the tables. Since

the number of proteins structures is expected to increase to around 30000 by the turn of the century, it is important that TESS is quick. We have found the search time is linearly dependent on the size of the database, which will enable TESS to quickly search through increasingly large amounts of structural data.

Previously, it has been established that it is possible to derive a 3D consensus template for the serine proteinases and lipases (Wallace *et al.*, 1996). TESS allows us to extend the work by creating more 'generic' templates so that we can investigate the structure of other enzyme active sites. This will eventually lead to a database of 3D enzyme active site or functional templates, that, in analogy to the 1D templates present in protein sequence motif databases such as PROSITE or PRINTS, will allow the assessment of the biological function and evolutionary origins of a new protein structure. Of course, TESS could be used to produce databases of other recurring 3D templates such as metal or non–enzyme ligand binding sites.

4.5 References

Alexandrov N.N., Takahashi K. & Go N. (1992) Common spatial arrangements of backbone fragments in homologous and non–homologous proteins *J. Mol. Biol.* **225** 5–9

Artymiuk P.J., Poirette A.R., Grindley H.M., Rice D.W. & Willett P. (1994) Graph Theoretic approach to the identification of three-dimensional patterns of amino–acid side–chains in protein structures *J. Mol. Biol.* **243** 327–344

Attwood T.K., Beck M.E., Bleasby A.J. & Parry-Smith D.J. (1994) PRINTS – a database of protein motif fingerprints *Nucleic Acids Research* **22** 3590–3696

Bairoch A. & Boeckmann B. (1994) The SWISS–PROT protein sequence databank: current status *Nucleic Acid Research* **22** 3578–3589

Bairoch A. & Bucher P. (1994) PROSITE: recent deveopments *Nucleic Acids Research* **22** 3583–3582

Bernstein F.C., Koetzle T.F., Williams G.J., Meyer E.F. Jr., Brice M.D., Rogers J.R., Kennard O., Shimanouchi T. & Tasumi M. (1977) The Protein Databank: a computer–based archival file for macromolecular structures *J. Mol. Biol.* **112** 535–542

Bleasby A.J., Akrigg D., Attwood T.K. (1994) OWL – A non–redundant, composite protein sequence database *Nucleic Acid Research* **22** 3574–3577

Bone R., Fujishige A., Kettner C.A., Agard D.A. (1991) Structural basis for broad specificity in α–lytic protease mutants *Biochemistry* **30** 10388-10398

Brady L., Brzozowski A. M., Derewenda Z. S., Dodson E., Dodson G., Tolley S., Turkenburg J. P., Christianson L., Huge Jensen B., Norskov L., Thim L. & Menge U. (1990) A serine protease triad forms the catalytic centre of triacylglycerol lipase *Nature* **343** 767–770

Derewenda U., Brzozowski A.M., Lawson D. & Derewenda Z.S. (1992) Catalysis at the interface: the anatomy of a conformational change in a triacylglycerol lipase *Biochemistry* **31** 1532–1541

Drenth J., Jansonius J.N., Koekoek R., Swen H.M. & Wolthers B. G. (1968) Structure of Papain *Nature* **218** 929–934

Fischer D., Wolfson H., Lin S.L. & Nussinov R. (1994) Three–dimensional, sequence order–independant comparison of a serine–proteinase against the crystallographic database reveals active site similarities: Potential implications to evolution and to protein folding *Protein Sci.* **3** 769–778

Franken S.M., Rozeboom H.J., Kalk K.H., Dijkstra B.W. (1991) Crystal structure of haloalkane dehalogenase: an enzyme to detoxify halogenated alkanes *Embo J.* **10** 1297–1309

Glusker J. P. (1991) Structural aspects of metal liganding to functional groups in proteins *Advances in Protein Chemistry* **42** 1–76

Harel M., Schalk I., Ehret–Sabattier L., Bouet F., Goeldner M., Hirth C., Axelson P., Silman I., Sussman J. (1993) Quaternary ligand binding to aromatic residues in the active–site gorge of Acetylcholinesterase *Proc. Nat. Acad. Sci. (USA)* **90** 9031–9042

Hodgman T. C. (1989) The elucidation of protein function by sequence motif analysis *CABIOS* **5** 1–13

Hutchinson E.G. & Thornton J.M. (1996) PROMOTIF – A program to identify and analyse structural motifs in proteins *Protein Science* **5** 212–220

Jernigan R., Raghunathan G. & Bahar I. (1994) Characterization of interactions and metal-ion binding-sites in proteins *Curr. Opinion in Struct. Biol.* **4** 256–263

Ke H. (1992) Similarities and differences between human cyclophilin A and other β–barrel structures. *J. Mol. Biol.* **228** 539–550

Lamdan Y., Schwartz J.T. & Wolfson H.J. (1988) On recognising 3D objects from 2D images. *Proceedings of IEEE Int. Conf. on Robotics and Automation, Philadelphia, Pa.* 1407–1413

Matthews B.W. & Rossman M.G. (1985) Comparison of protein structures *Methods Enzymol* **115** 397–420

Marquart M., Deisenhofer J., Huber R. & Palm W. (1980) Crystallographic refinement and atomic models of the intact immunoglobulin molecule kol and its antigen–binding fragment at 3.0Å and 1.9Å resolution *J. Mol. Biol.* **141** 369–391

Mitchel E.M., Artymuick P.J., Rice D.W. & Willet P. (1990) Use of techniques derived from graph theory to compare secondary structure motifs in proteins *J. Mol. Biol.* **212** 151–166

Murzin A.G., Brenner S.E., Hubbard T. & Chothia C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures *J. Mol. Biol.* **247** 536–540

Nussinov R. & Wolfson H.J. (1991) Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques *Proc. Natl. Acad. Sci. USA* **88** 10495–10499

Ollis D.L., Cheah E., Miroslaw C., Dijkstra B., Frolow F., Franken S.M., Harel M., Remington S.J., Silman I., Schrag J., Sussman J.L., Verschueren K.H.G & Goldman A. (1992) The α/β hydrolase fold *Protein Engineering* **5** 197–211

Orengo C.A., Flores T.P., Taylor W.R. & Thornton J.M (1993) Identification and classification of protein fold families *Protein Engineering* **6** 485–500

Pathak D. & Ollis D.J. (1990) Refined crystal structure of dienelactone hydrolase at 1.8Å*J. Mol. Biol* **214** 497–525

Singh J.S. & Thornton J.M. (1992) Protein side–chain interactions *IRL press* at Oxford University Press **1** 6–9

Stoddard B.L., Bruhkne J., Porter N., Ringe D. & Petsko G.A (1990) Structure and activity of two photoreversible cinnamates bound to chymotrypsin *Biochemistry* **29** 4871–4879

Taylor W.R. (1988) Pattern matching in protein sequence comparison and structure prediction *Protein Engineering* **2** 77–86

Taylor W.R. & Jones D.T. (1991) Templates, consensus patterns and motifs *Curr. Opinion in Struct. Biol.* **1** 327–323

Wallace A.C., Laskowski R.A. & Thornton J.M. (1996) Derivation of 3D coordinate templates for searching structural databases: application to the Ser–His–Asp catalytic triads in the serine proteinases and lipases *Protein Science* **5** 1001–1013

Weaver L.H. & Matthews B.W. (1987) Structure of bacteriophage T4 at 1.7Å resolution. *J. Mol. Biol*, **193** 189–199

Zvelebil M.J.J.M & Sternberg M.J.E (1988) Analysis and prediction of catalytic residues in enzymes *Protein Engineering* **2** 127–138

Chapter 5

The catalytic triad

5.1 Introduction

This chapter will describe the role of histidine as part of a generalised 'catalytic triad'; a specific example is the Ser–His–Asp catalytic triad (Wright *et al.*, 1969; Blow *et al.*, 1969) described in chapter 3.

The generalised triad can be thought of as Nu:–His–ELEC, where Nu: is a nucleophilic group and ELEC, the electrostatic group, acts to perturb the pK_a of the acid/base His. In chapter 3 we saw that a 3D template in the form of Ser O^{γ}–His sidechain–Asp O^{δ} can be defined that will identify all serine proteinase and lipase active sites with the exclusion of all other Ser, His and Asp interactions. Two other groups of enzymes in the PDB, the cysteine proteinases and the α/β hydrolase fold enzymes (Ollis *et al.*, 1992), have a catalytic triad. In this chapter we will compare and contrast the catalytic triad conformations of all these enzymes. With the exception of the cysteine proteinases, the catalytic triad, in terms of its functional atoms, is structurally conserved. This has enabled us to construct one consensus template that can describe the active site of more than one unique enzyme by E.C. number.

5.2 The Nu:–His–ELEC catalytic triad

The enzymes with a Nu:–His–ELEC triad have been divided into 4 classes according to the residues corresponding to Nu: and ELEC; these are listed in Table 5.1. The α/β hydrolase fold (Ollis *et al.*, 1992) occurs in 3 of the 4 classes. Despite a low sequence identity, it suggests that the enzymes in this fold group have evolved from a common ancestor so as to preserve the positions of the key catalytic components.

The class 1 Ser–His–Asp catalytic triad of Table 5.1 has been discussed in detail in the previous chapters; the structural aspects of the other 3 classes' catalytic triads will be described in detail in the following sections.

5.3 class 2: The Ser–His–Glu catalytic triad

The Ser–His–Glu triad has been found in both triacylglycerol lipase from *Candida rugosa*, 1*trh* (Grochulski *et al.*, 1993) and acetylcholinesterase, 1*ace* (Sussman *et al.*, 1991). Lipase is also present in class 1 where Glu is replaced by an Asp; it hydrolyses triacylglycerides into diacylglycerides and subsequently monoacylglycerides and free fatty acids. Acetylcholinesterase is responsible for termination of impulse transmission at cholinergic receptors by hydrolysis of the neurotransmitter acetylcholine.

A 3D representation of these two catalytic triads is shown in Figure 5.1. The His residue of these two triads have been superimposed showing that the relative conformations of the corresponding Ser and Glu sidechains are similar. The catalytic Ser 200, His 440 and Glu 327 residues of 1*ace* were used to create sidechain and functional (Ser O^{γ}, His sidechain and Glu $O^{\epsilon 1}$) consensus templates. The *rms* distance of both the sidechain and functional Ser–His–Glu coordinate sets

class 1: Ser–His–Asp catalytic triad		
β–sandwich trypsin–like fold		
trypsin family	E.C.3.4.21.x	Blow *et al.*, 1969
alternating α/β subtilisin–like fold		
subtilisin family	E.C.3.4.21.x	Wright *et al.*, 1969
α/β hydrolase fold		
serine–type carboxypeptidase	E.C.3.4.16.5	Liao *et al.*, 1992
lipase	E.C.3.1.1.3	Brady *et al.*, 1990
class 2: Ser–His–Glu catalytic triad		
α/β hydrolase fold		
Candida rugosa lipase	E.C.3.1.1.3	Grochulski *et al.*, 1993
acetylcholinesterase	E.C.3.1.1.4	Sussman *et al.*, 1991
class 3: Asp–His–Asp catalytic triad		
α/β hydrolase fold		
haloalkane dehalogenase	E.C.3.8.1.5	Verscheuren *et al.*, 1993
class 4: Cys–His–Asn catalytic triad		
$\alpha + \beta$ cysteine proteinase fold		
papain	E.C.3.4.22.2	Drenth *et al.*, 1968
actinidin	E.C.3.4.22.14	Baker, 1980
caricain	E.C.3.4.22.30	Pickersgill *et al.*, 1993

Table 5.1: The four different Nu:–His–ELEC catalytic triads found in the PDB, where ELEC acts to perturb the pK_a of the acid/base His and Nu: is a nucleophilic group.

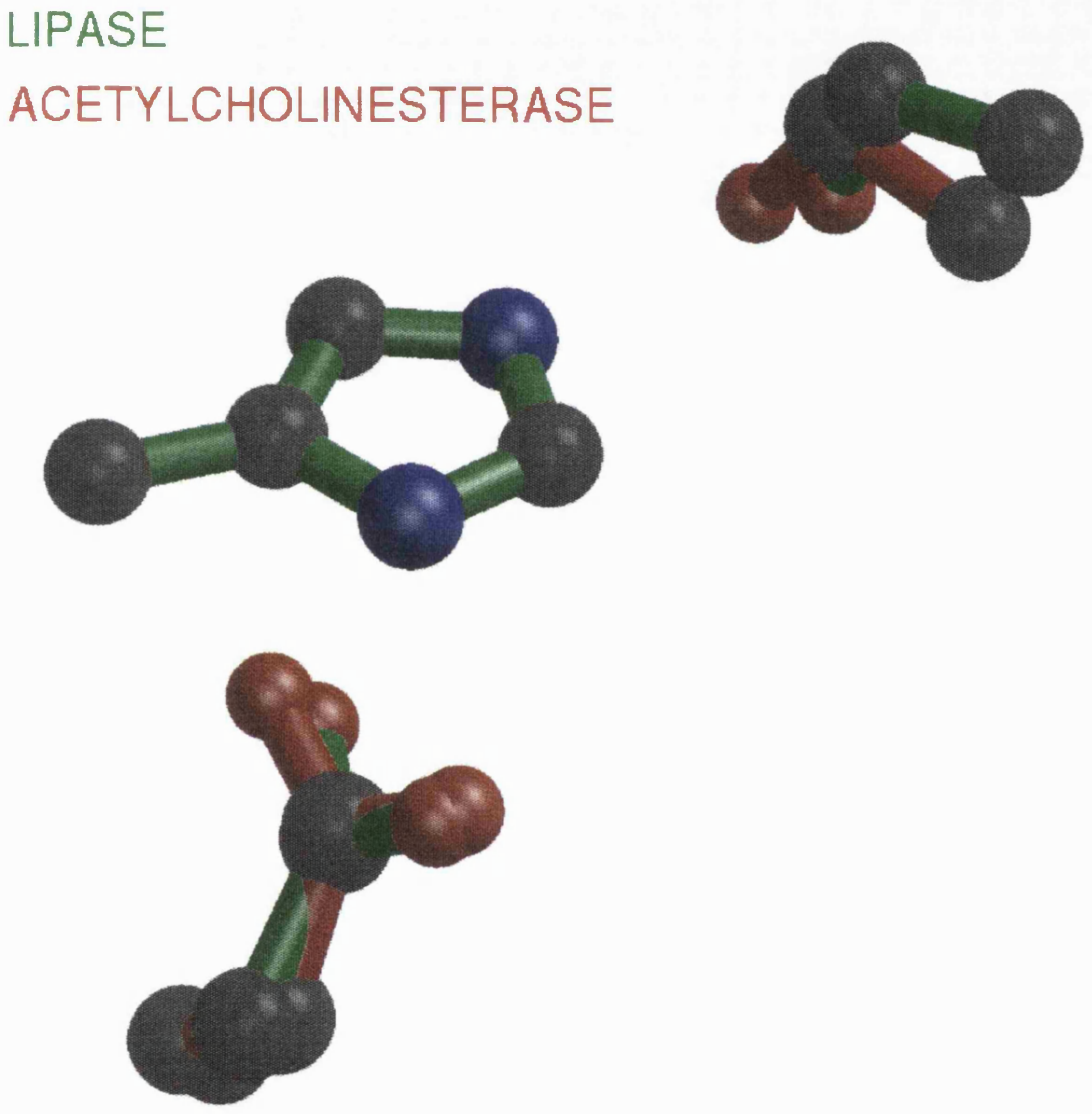


Figure 5.1: A 3D representation of the Ser–His–Glu catalytic triads from acetylcholinesterase 1*ace* (Sussman *et al.*, 1991) in green and triacylglycerol lipase 1*trh* (Grochulski *et al.*, 1993) in red.

class 2: Ser–His–Glu catalytic triad		
acetylcholinesterase E.C.3.1.1.7		
PDB code	*rms* from functional consensus template	*rms* from sidechain consensus template
1*ace*	0.23	0.26
1*acj*	0.27	0.37
1*acl*	0.40	0.82
1*ack*	-	-
lipase E.C.3.1.1.3		
PDB code	*rms* from functional consensus template	*rms* from sidechain consensus template
1*crl*	0.14	0.40
1*thg*	0.25	0.49
1*trh*	0.30	0.49

Table 5.2: The *rms* deviations from the mean functional and sidechain consensus templates for the Ser–His–Glu catalytic triad present in acetylcholinesterase and lipase X–ray crystal structures. The results show that the catalytic triad is structurally conserved in these two enzyme types.

of all lipase and acetylcholinesterase structures were measured against the appropriate consensus template; the results are given in Table 5.2. The catalytic triads of these two enzymes are the same, with no sidechain coordinate having an *rms* greater than 0.9Å from the mean sidechain consensus template. The exception to this is the acetylcholinesterase structure 1*ack* (Harel *et al.*, 1993), which has the cyclic compound 'edrophonium' (ethyl(3-hydroxyphenyl)dimethylammonium) in its active site. Covalent or non–native inhibitors bound to the active site of the serine proteinases perturb the Ser–His–Asp catalytic triad geometry; this acetylcholinesterase structure is another example.

Lipase and and acetylcholinesterase have a high sequence identity (Shimada *et al.*, 1990), the same α/β hydrolase fold (Ollis *et al.*, 1992) and similar active

site geometry suggesting that the two enzymes evolved from a common ancestor.

5.4 class 3: The Asp–His–Asp catalytic triad

The nitrogen–fixing hydrogen bacteria *Xanthobacter autotrophicus* can grow in a medium of 1,2–dichloroethane or 2–chloroethanol as its sole carbon energy source. These compounds are initially metabolised by haloalkane dehalogenase which converts 1–haloalkanes into primary alcohols and a halide ion by hydrolytic cleavage of the carbon–halide bond. The crystal structure of this enzyme (PDB code 2*dhc*) has been determined to 1.9Å resolution (Verschueren *et al.*, 1993); it is a member of the α/β hydrolase–fold family with a catalytic triad consisting of residues Asp 124, His 289 and Asp 260. In haloalkane dehalogenase the nucleophilic group is an Asp residue and the bond cleaved is carbon–halide as opposed to an ester group in the other classes.

Figure 5.2 is a 3D representation of the catalytic centre of haloalkane dehalogenase, 2*dhc* (Verschueren *et al.*, 1993). It shows the substrate 1,2 dichloroethane lying below the plane of the ring of the acid/base catalyst His 289. The His 289 $N^{\delta 1}$ and Asp 260 $O^{\delta 2}$ atoms are hydrogen bonded to each other; these two residues constitute the His–Asp acid/base catalyst in analogy to the His 57–Asp 102 of the serine proteinases. The nucleophilic Asp 124 $O^{\delta 1}$ is poised in a position to attack the C1 carbon of the substrate. Figure 5.3 is a LIGPLOT picture of the resultant complex, taken from the X–ray structure 2*dhd* (Verschueren *et al.* (1993)). It shows the Asp 124 residue covalently bound to the substrate (Mce 124) and the CL1 chlorine has been displaced. The His 289–Asp 260 acid/base catalyst is in hydrogen bonding vicinity. A water molecule attacks this acyl–enzyme intermediate resulting in the hydrolytic cleavage of the C–CL substrate bond to yield the primary alcohol product.

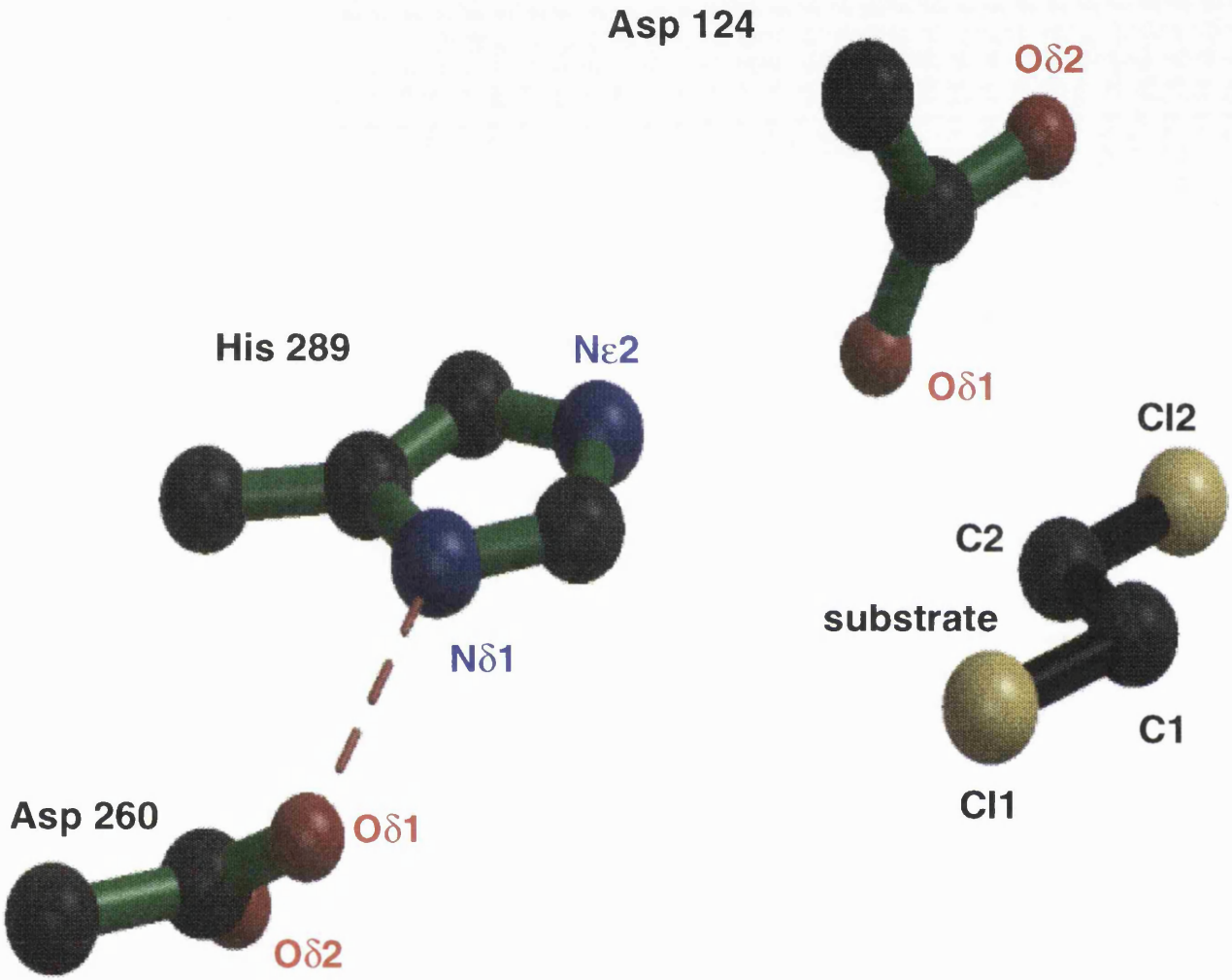


Figure 5.2: A 3D representation of the active site of haloalkane dehalogenase, 2*dhc* (Verschueren *et al.*, 1993). The His 289–Asp 260 residues hydrogen bond to each other and constitute the acid/base catalyst. Asp 124 is the nucleophilic group and attacks the C1 of the 1,2 dichloroethane substrate, forming the acyl–enzyme intermediate.

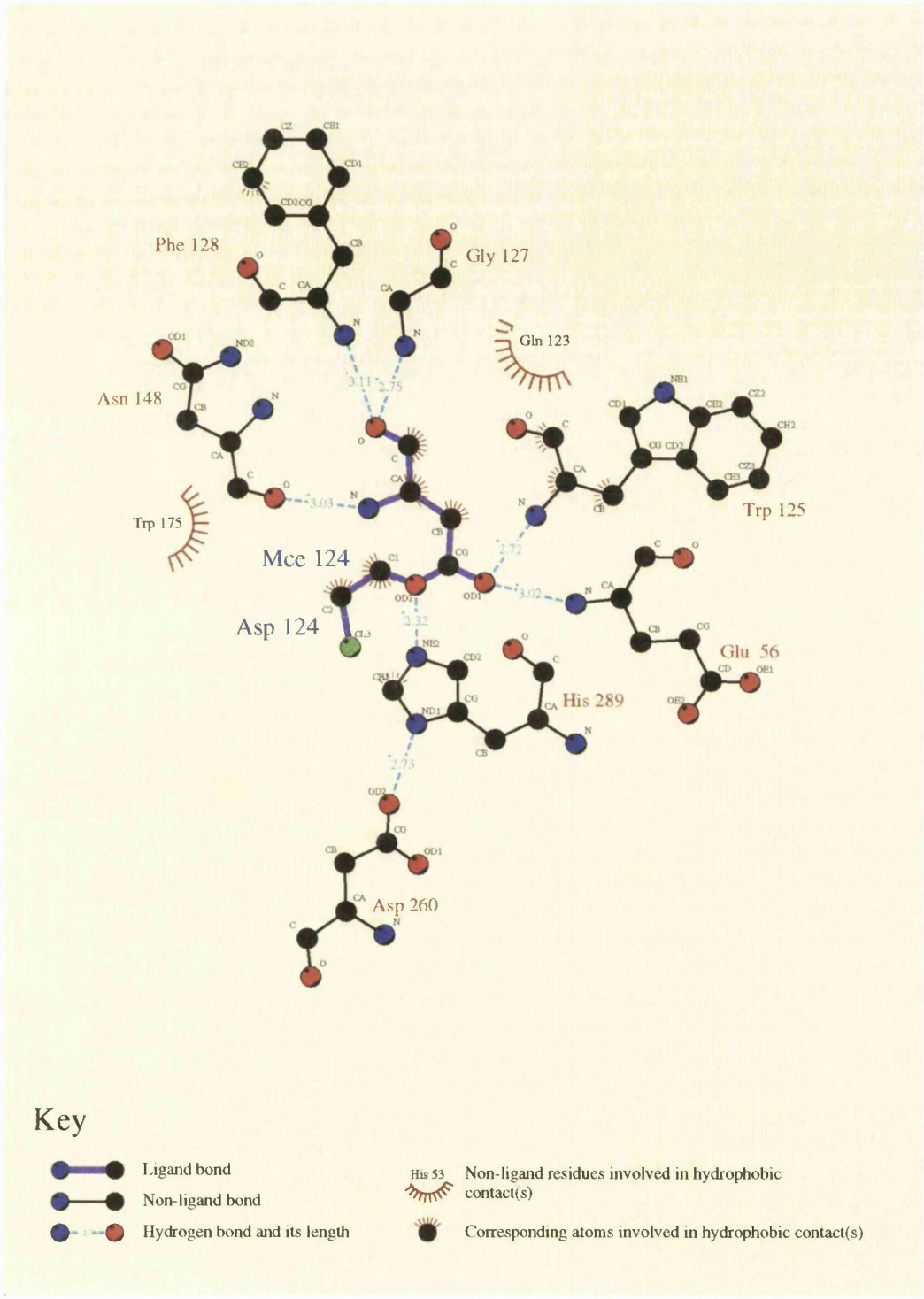


Figure 5.3: A LIGPLOT diagram representing the acyl–enzyme intermediate of haloalkane dehalogenase 2*dhd* (Verschueren *et al.*, 1993) formed after the nucleophilic Asp 124 attacks the 1,2 dichloroethane substrate. A water molecule would attack this intermediate, forming the primary alcohol product.

class 3: Asp–His–Asp catalytic triad		
haloalkane dehalogenase E.C.3.8.1.5		
PDB code	*rms* from functional consensus template	*rms* from sidechain consensus template
1edb	0.15	0.17
1edd	0.13	0.15
1ede	0.75	0.88
2dhc	0.40	0.47
2dhd	0.26	0.34
2dhe	0.11	0.17
2eda	0.38	0.56
2edc	0.11	0.16
2had	0.25	0.32

Table 5.3: The *rms* deviations from the functional and sidechain templates for the dataset of haloalkane dehalogenase X–ray crystal structures.

All the structures in the haloalkane dehalogenase dataset are from *Xanthobacter autotrophicus*; the differences being the crystallisation conditions and binding of inhibitors to the active site. Sidechain and functional (Asp 124 O^{δ_1}, His sidechain and Asp O^{δ_2}) templates using the seed atoms from *2had* (Franken *et al.*, 1991) have been calculated; all the sidechain and functional templates have an *rms* deviation below 1Å from their respective consensus templates (Table 5.3); even when Asp 124 is part of an acyl–enzyme intermediate, the catalytic triad does not change conformation; for example, the *2dhd* (Verschueren *et al.*, 1993) sidechain template is only 0.34Å from the mean sidechain consensus template. This suggests that there is limited conformational change of the catalytic triad during the reaction course.

5.4.1 class 4: The Cys–His–Asn catalytic triad

The cysteine proteinases are widely distributed in nature and the X–ray structures have been solved for papain E.C.3.4.22.2 (Drenth *et al.*, 1968) and cari-

cain E.C.3.4.22.30 (Pickersgill *et al.*, 1991) from the papaya plant and actinidin E.C.3.4.22.14 from kiwi fruit (Varughese *et al.*, 1992). These structures all have the same $\alpha + \beta$ cysteine proteinase fold.

Papain is the best understood of the cysteine proteinases; the His 159–Asn 175 pair constitutes the acid/base catalyst while Cys 25 is the nucleophilic group. The reaction mechanism is analogous to that of the serine–proteinases whereby an acyl–enzyme intermediate (Baker & Drenth, 1987) is formed between the protein substrate and the cysteine group and is subsequently hydrolysed by water to form product. There is, however, controversy about the precise protonation states of the intermediates during the reaction course and whether other residues are involved in the reaction mechanism (Wang *et al.*, 1994).

The sidechain and functional (Cys S^{γ}, His sidechain and Asn O^{δ_1}) consensus templates have been calculated using the Cys 25–His 159–Asn 175 catalytic triad from the papain structure 1*ppp* (Kim *et al.*, 1992) as a seed template; the results are shown in Table 5.4. All triads in the dataset are conserved in structure with the maximum *rms* distance for the sidechain template being 0.86Å. Figure 5.4 is a 3D representation of the catalytic triads from the 3 cysteine proteinases. The catalytic triad conformations are very similar and they, like the Ser–His–Glu catalytic triad from acetylcholinesterase and lipase, must have evolved from a common ancestor. However, the conformation of the Nu:–His–ELEC atoms with respect to the His sidechain is opposite to classes 1, 2 and 3; the Cys S^{γ} nucleophile is interacting with the His N^{δ_1} and the Asn O^{δ_1} with the N^{ϵ_2} of the His ring.

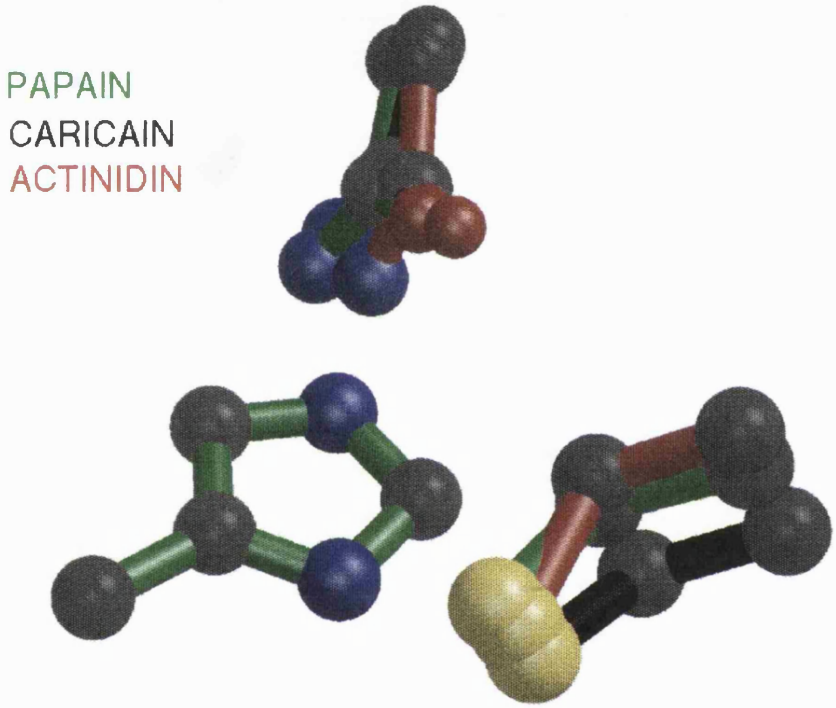


Figure 5.4: A 3D representation of the Cys–His–Asn catalytic triads from the cysteine proteinases papain (Drenth *et al.*, 1968), actinidin (Varughese, 1992) and caricain (Pickersgill *et al.*, 1991). The triads are very similar reflecting the high sequential and structural similarities of the 3 proteinases.

class 4: Cys–His–Asn catalytic triad

papain E.C.3.4.22.2

PDB code	*rms* from functional consensus template	*rms* from sidechain consensus template
1*aec*	0.48	0.55
1*edd*	0.55	0.78
1*pe6*	0.28	0.37
1*pip*	0.28	0.41
1*pop*	0.26	0.61
1*ppd*	0.46	0.68
1*ppn*	0.32	0.37
1*ppp*	0.27	0.29
1*stf*	0.39	0.38
1*pad*	0.55	0.78
2*pad*	0.69	0.86
4*pad*	0.76	0.86
5*pad*	0.55	0.78
6*pad*	0.53	0.78
9*pap*	0.48	0.80

actinidin E.C.3.4.22.14

PDB code	*rms* from functional consensus template	*rms* from sidechain consensus template
2*act*	0.54	0.85
1*aec*	0.48	0.55

caricain E.C.3.4.22.30

PDB code	*rms* from functional consensus template	*rms* from sidechain consensus template
1*ppo*	0.44	0.66

Table 5.4: The *rms* deviations from the functional and sidechain templates for the dataset of X–ray crystal structures for the thiol proteinases papain, actinidin and caricain.

5.5 Comparison of the 4 Nu:–His–ELEC catalytic triads

The class 1 catalytic triads have had their sidechain and functional triads compared extensively in chapter 3. In summary, we found that though the sidechain templates varied extensively among the different protein folds of this class, the functional templates (Ser O^{γ}, His sidechain, Asp $O^{\delta 1}$) adopted the same conformation. This suggests that convergent evolution has drawn the nucleophilic Ser O^{γ} into a position that enables it to interact catalytically with the His–Asp acid/base pair, forming the catalytic triad.

Figure 5.5 is a 3D representation of the sidechain templates from classes 1, 2 and 3. The His sidechains have been superimposed enabling us to compare the relative position of the nucleophilic and electrostatic sidechains. Chymotrypsin 1*cho* (Fujinaga *et al.*, 1987) represents class 1; the sidechain templates of the subtilisin, serine–type carboxypeptidase and lipase have been left out for clarity. We have already noted that the catalytic triad of the cysteine proteinases is markedly different from classes 1, 2 and 3 so this has also been omitted. The diagram shows that the sidechains originate from different orientations across classes 1, 2 and 3. However, we can see that there is clustering of the functional atoms for each of these classes; the nucleophilic oxygens are all in proximity of the acid/base catalyst His $N^{\epsilon 2}$ and the electrostatic residues are in a hydrogen bonding position with the His $N^{\delta 1}$. This suggests that we can create a single consensus template that will enable us to describe the active sites of classes 1, 2 and 3 (the class 1–2–3 consensus template).

Functional consensus templates were created for each of the classes 1, 2 and 3 and were averaged to create the class 1–2–3 consensus template whose coordinates are given in Table 5.5. The *rms* deviation of all the PDB structures in the class 1,

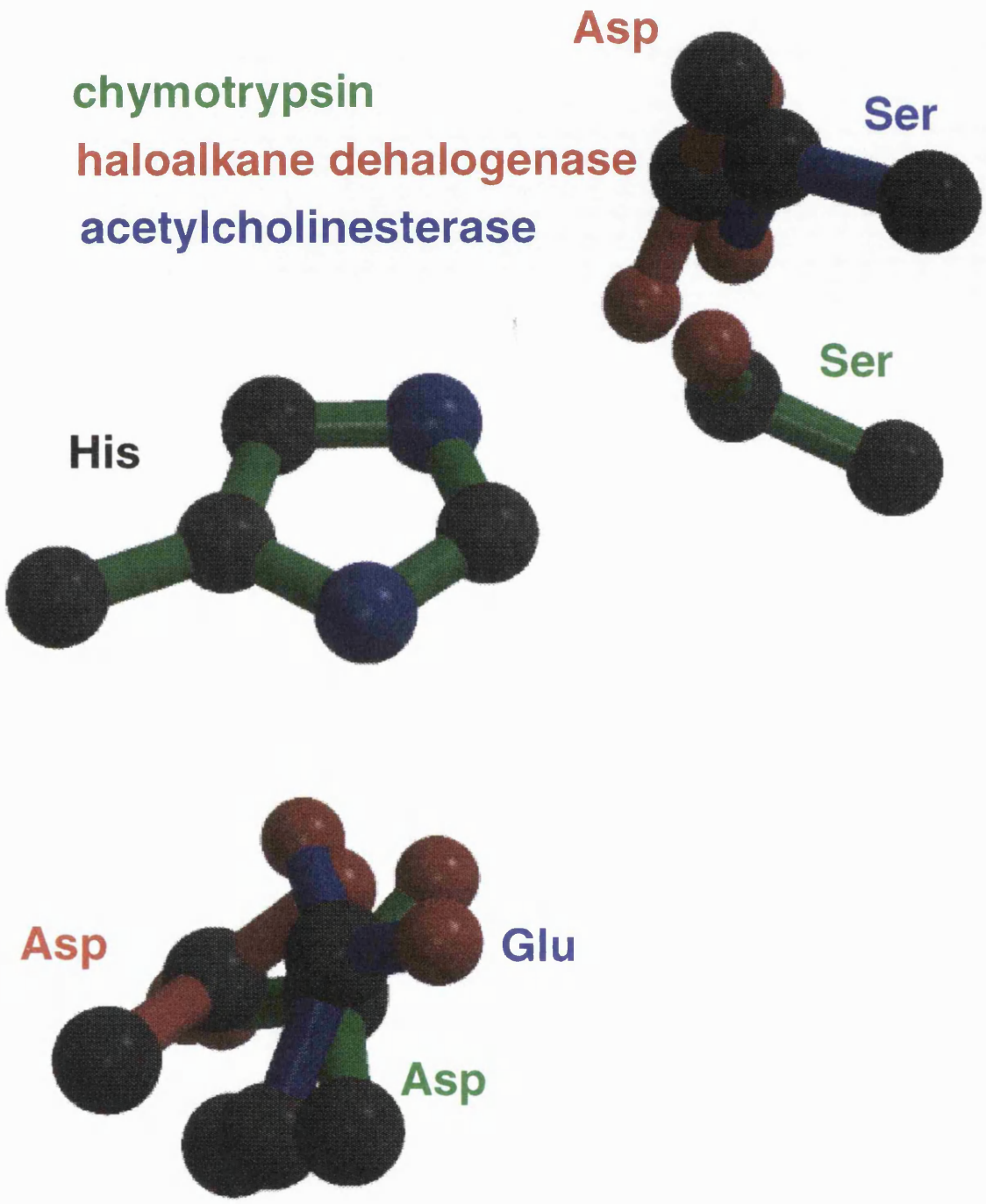


Figure 5.5: A comparison of the catalytic triads from chymotrypsin 1*cho* (Fujinaga *et al.*, 1987), haloalkane dehalogenase 2*dhc* (Verschueren *et al.*, 1993) and acetylcholinesterase 1*ace* (Sussman *et al.*, 1991). All the triads His residues have been superimposed allowing us to compare the relative conformations of the nucleophilic and electrostatic sidechains.

class 1–2–3 functional consensus template					
Residue	Res. Number	Atom	x	y	z
Ser/Asp	1	O^{γ}/O^{δ_1}	4.8	1.1	-0.0
Glu/Asp	2	$O^{\epsilon_1}/O^{\delta_1}$	-0.3	-3.5	0.2
His template residue					
Residue	Res. Number	Atom	x	y	z
His	3	C^{β}	-1.4	-0.1	-0.0
His	3	C^{γ}	0.0	0.0	0.0
His	3	N^{δ_1}	0.8	-1.1	0.0
His	3	C^{δ_2}	0.8	1.1	0.0
His	3	C^{ϵ_1}	2.1	-0.7	-0.0
His	3	N^{ϵ_2}	2.1	0.6	-0.0

Table 5.5: The coordinates of the functional consensus template that describes the active sites of the serine proteinases, lipases, acetylcholinesterase and haloalkane dehalogenase enzymes.

2 and 3 datasets were measured against the class 1–2–3 template. Figure 5.6 is a histogram of number of hits against *rms* deviation from the class 1–2–3 template. The majority of triads have an *rms* deviation between 0.3Å and 1.3Å, those triads above this value are the class 1 triads with inhibitors bound to their active sites.

The consensus templates derived for each of the 4 classes were measured against the class 1–2–3 template; the coordinates and *rms* values of these templates are given in Table 5.6. Apart from class 4, the cysteine proteinases, the *rms* values are below 0.6Å, showing that the functional atoms are structurally conserved. This can be seen clearly in Figure 5.7 where the functional oxygen atoms cluster in positions that allow them to interact with the His N^{ϵ_2} and His N^{δ_1}. The cysteine proteinase thiol nucleophile and electrostatic oxygen are in completely different positions reflecting the different active site geometry for this enzyme.

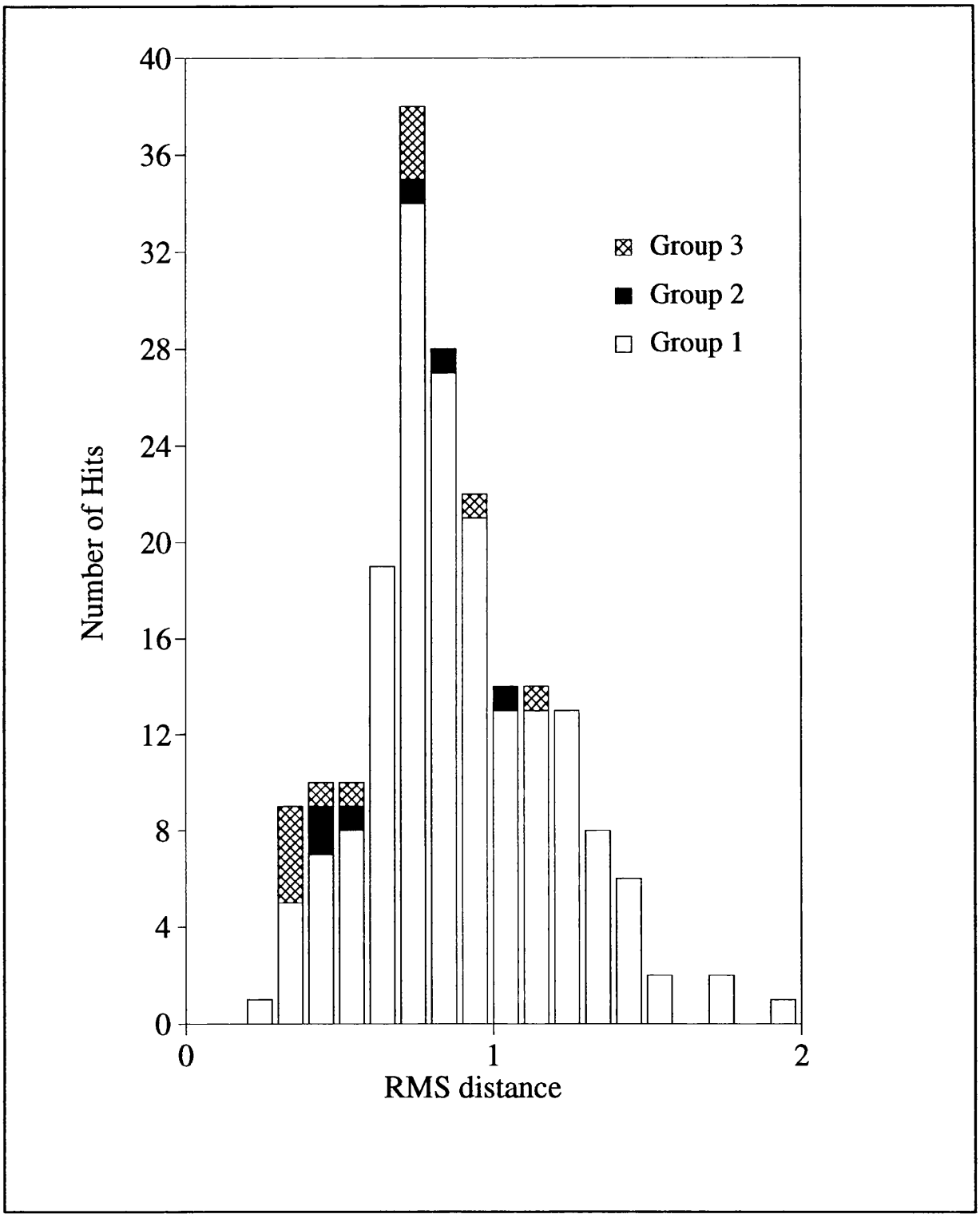


Figure 5.6: A histogram of number of hits against *rms* distance from the class 1–2–3 template for all the 95% non–identical PDB structures present in classes 1, 2 and 3.

class 1: Ser–His–Asp functional consensus template
rms from class 1–2–3 template 0.59Å

Residue	Res. Number	Atom	x	y	z
Ser	194	O^{γ}	4.9	0.8	-0.3
Asp	91	O^{δ_2}	0.4	-3.7	0.1

class 2: Ser–His–Glu functional consensus template
rms from class 1–2–3 template 0.59Å

Residue	Res. Number	Atom	x	y	z
Ser	200	O^{γ}	5.0	1.3	0.3
Glu	327	O^{ϵ_1}	-0.9	3.2	0.4

class 3: Asp–His–Asp functional consensus template
rms from class 1–2–3 template 0.24Å

Residue	Res. Number	Atom	x	y	z
Asp	124	O^{δ_1}	4.5	1.1	-0.1
Asp	260	O^{δ_2}	-0.3	-3.7	0.1

class 4: Cys–His–Asn functional consensus template
rms from class 1–2–3 template 7.20Å

Residue	Res. Number	Atom	x	y	z
Cys	25	S^{γ}	0.5	-4.0	2.1
Asp	260	O^{δ_2}	4.4	2.3	0.5

His template residue

Residue	Res. Number	Atom	x	y	z
His	3	C^{β}	-1.4	-0.1	-0.0
His	3	C^{γ}	0.0	0.0	0.0
His	3	N^{δ_1}	0.8	-1.1	0.0
His	3	C^{δ_2}	0.8	1.1	0.0
His	3	C^{ϵ_1}	2.1	-0.7	-0.0
His	3	N^{ϵ_2}	2.1	0.6	-0.0

Table 5.6: The consensus templates for each of the 4 classes. Each template is superimposed onto the same His template residue. Their *rms* distances from the class 1–2–3 template is also given.

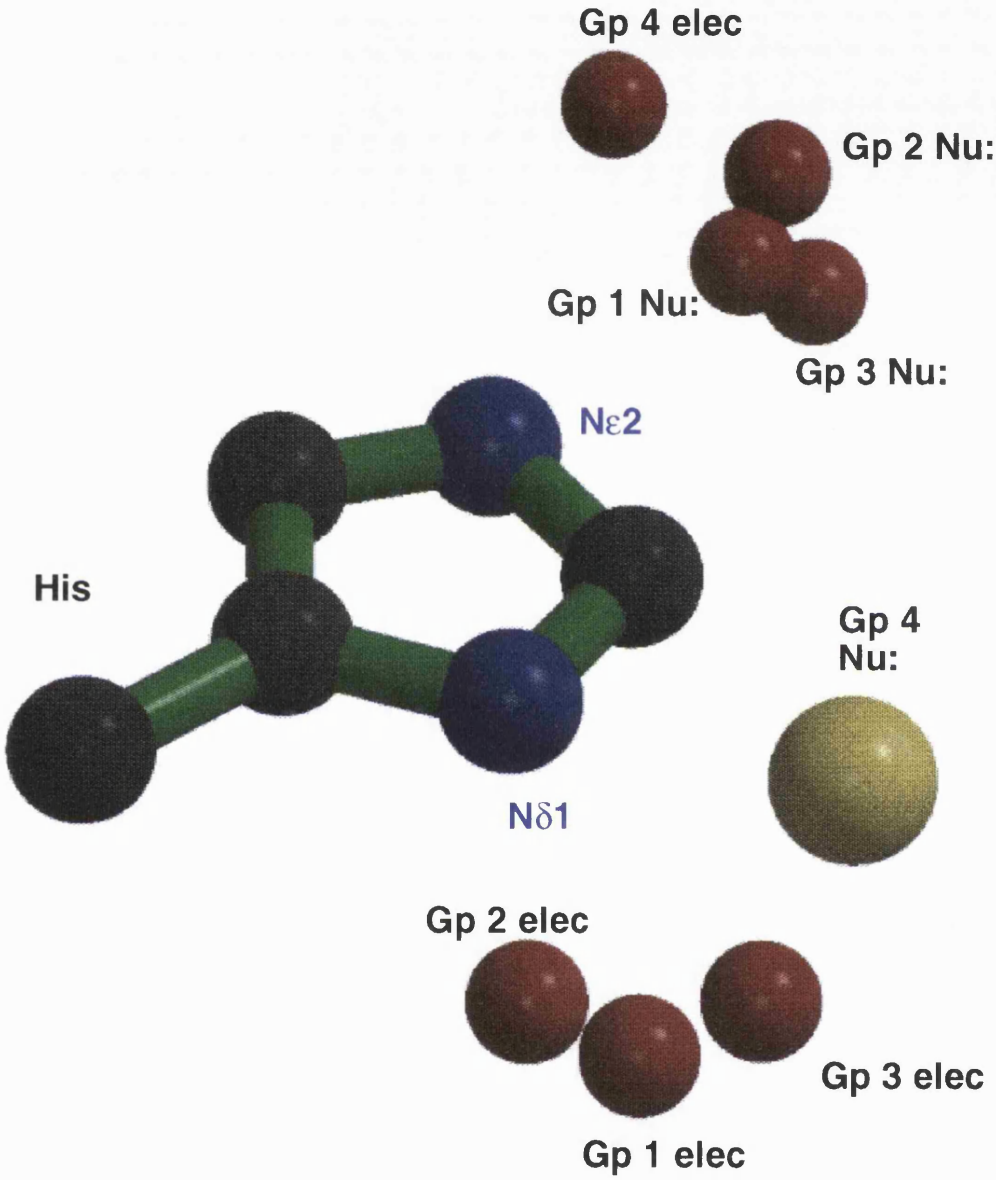


Figure 5.7: A 3D representation of the mean position of the functional atoms with respect to the His sidechain for classes 1 to 4.

5.5.1 class 1–2–3 template search through the PDB

It is interesting to see if the arrangement of the atoms in the class 1–2–3 consensus template, in terms of the functional oxygens surrounding the His sidechain, occurs elsewhere in the PDB. We searched through a dataset of representative protein structures in the January 1995 PDB which contains some of the structures from classes 1, 2 and 3, the others having been excluded on the basis of having greater than 95% sequence identity. A triad located in the non–homologous dataset is considered as interesting if its *rms* deviation is less than 2.0Å from the Group 1–2–3 consensus template and there is an Asp or Ser sidechain atom in the position equivalent to Nu:, the nucleophilic group and an Asp or Glu sidechain atom in the position equivalent ELEC, the electrostatic group.

Figure 5.8 shows that there are some proteins which are not members of classes 1, 2 and 3 but which have the characteristics of a Nu:–His–ELEC triad. There were 2 triads found with a potential Ser–His–Asp catalytic triad; cyclophilin 2*cpl*, immunoglobulin 2*ig*2. These triads have already been discussed in detail in Chapter 3. There are no triads located in the PDB with a Ser–His–Glu triad (class 2) other than the acetylcholinesterase and lipase PDB structures. Table 5.7 is a list of PDB structures, in order of their *rms* distances, with the Asp–His–Asp triad that are not members of the class 3 PDB dataset but fit the criteria mentioned above; these triads are discussed in detail in the following sections.

5.5.2 nitrogenase molybdenum–iron protein E.C.1.18.6.1

PDB code 1*min* - *rms* distance 0.42Å

This enzyme is part of the nitrogenase enzyme system which provides the biochemical machinery for nitrogen fixation and is essential for maintaining the nitrogen cycle on Earth. The nitrogenase enzyme system consists of two metallopro-

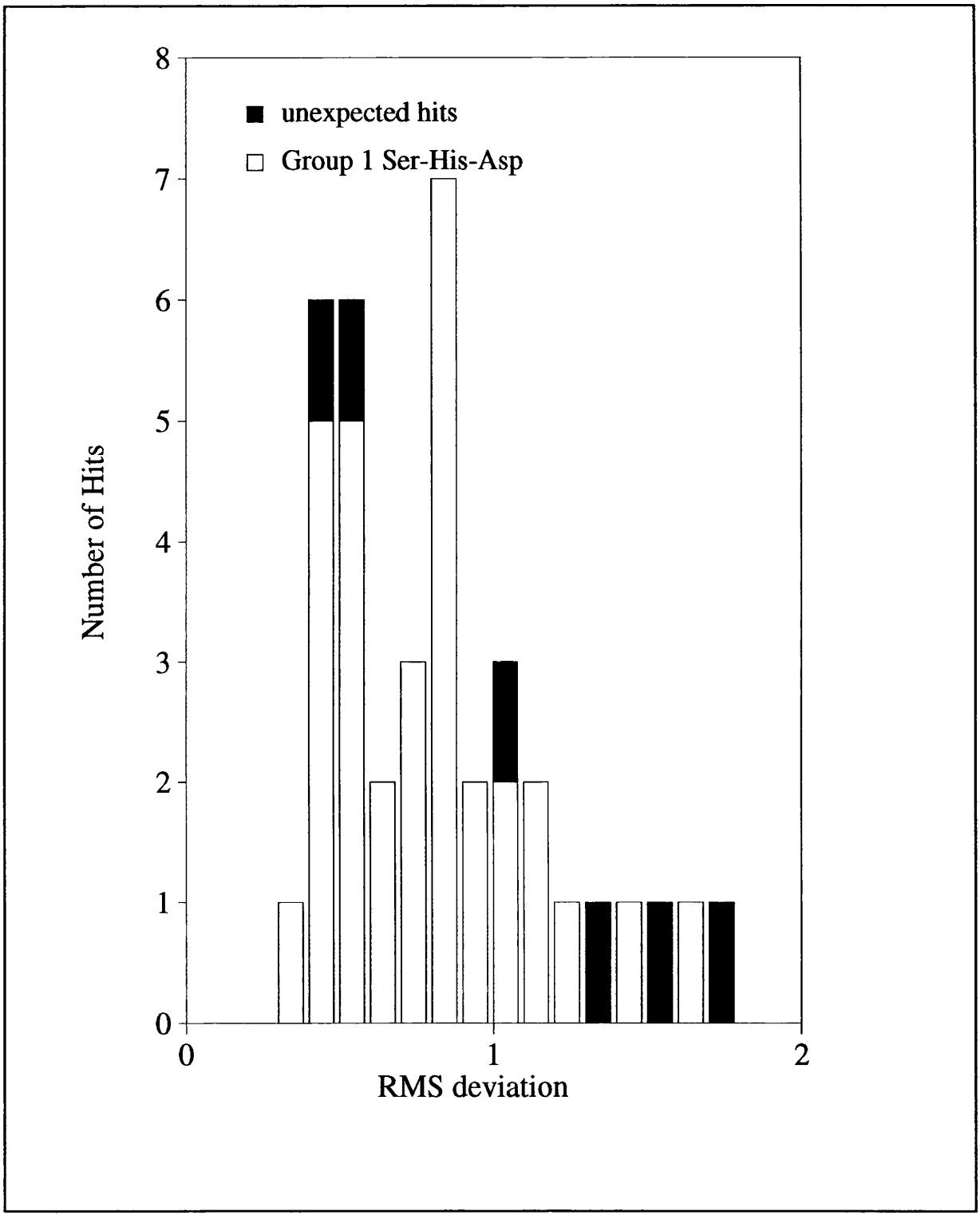


Figure 5.8: Histogram of number of hits against *rms* deviation when the 95% by sequence non–identical PDB dataset was searched using the class 1–2–3 consensus template. There are some triads that are not members of the enzyme datasets of classes 1, 2 or 3 but fit the criteria necessary to be a potential catalytic triad.

class 3: Asp–His–Asp catalytic triad

nitrogenase molybdenum–iron protein *rms* 0.42

Residue	Res. Number	Atom	Chain	x	y	z	PDB code
Asp	160	O^{δ_2}	D	5.0	1.4	-0.1	1*min*
Asp	116	O^{δ_2}	D	-0.5	-3.4	0.6	1*min*
His	90	Sidechain	D	-	-	-	1*min*

pyruvate oxidase E.C.1.2.3.3 *rms* 0.58

Residue	Res. Number	Atom	Chain	x	y	z	PDB code
Asp	69	O^{δ_1}	B	4.7	1.8	-0.4	1*pox*
Asp	27	O^{δ_2}	B	-0.4	-3.6	0.4	1*pox*
His	28	Sidechain	B	-	-	-	1*pox*

macromycin *rms* 1.08

Residue	Res. Number	Atom	Chain	x	y	z	PDB code
Asp	100	O^{δ_2}		5.8	1.9	-0.6	2*mcm*
Asp	53	O^{δ_1}		0.2	-3.5	-0.3	2*mcm*
His	32	Sidechain		-	-	-	2*mcm*

protein R2 of ribonucleotide reductase E.C.1.17.4.1 *rms* 1.37

Residue	Res. Number	Atom	Chain	x	y	z	PDB code
Asp	237	O^{δ_2}	A	4.3	2.1	-0.2	1*rib*
Asp	84	O^{δ_1}	A	-1.4	-4.8	0.7	1*rib*
His	118	Sidechain	A	-	-	-	1*rib*

superoxide dismutase E.C.1.15.1.1 *rms* 1.51

Residue	Res. Number	Atom	Chain	x	y	z	PDB code
Asp	124	O^{δ_1}	D	4.1	2.1	-0.8	1*sos*
Asp	83	O^{δ_1}	D	1.6	-3.7	0.8	1*sos*
His	71	Sidechain	D	-	-	-	1*sos*

D-glyceraldehyde-3-phosphate dehydrogenase E.C.1.2.1.12 *rms* 1.72

Residue	Res. Number	Atom	Chain	x	y	z	PDB code
Asp	312	O^{δ_2}	O	6.2	1.8	-1.0	1*gd*1
Asp	47	O^{δ_1}	O	-1.3	-4.7	0.3	1*gd*1
His	50	Sidechain	O	-	-	-	1*gd*1

Table 5.7: List of the potential catalytic triads found when the PDB was searched with the class 1–2–3 catalytic triad template.

teins, the molybdenum iron (MoFe) protein and the iron Fe–protein. Metabolic redox reactions pass electrons to the Fe–protein which in turn transfers them to the MoFe–protein in a process that is coupled to the hydrolysis of Mg–ATP. The MoFe–protein from *Azotobacter vinelandii* is a α_2/β_2 tetramer and the X–ray structure has been determined to 2.7Å resolution (Jongsun *et al.*, 1992). Both the α and β subunits have a general doubly wound α/β like fold and the MoFe cofactor is found in the α subunit. There is another cofactor, the P–cluster pair, which is located 10Å from the MoFe on the two fold axis that relates the α– and β– subunits. It is thought to transfer electrons between the Fe–protein 4Fe:4S cluster and the MoFe cofactor. The precise catalytic mechanism is unknown, but the N_2 substrate is proposed to bind directly to the MoFe cofactor. The MoFe cofactor has homocitrate bound and is surrounded by water molecules, it may be a source of protons for the formation of the NH_3 product.

A close up of the triad with the MoFe and P–cluster cofactors is shown in Figure 5.9. The Asp 160–His 90–Asp 116 triad and the P–cluster are located within about 8Å of each other, the triad has not been mentioned or implicated in the reaction course and whether it is involved in proton or electron transfer is unknown. Closer inspection reveals that the Asp 116 is in fact accessible to the surface of the protein, indicating it could have access to solvent or ligands.

The sequence of the MoFe–protein has been checked against the SWISS–PROT (Bairoch & Boeckmann, 1994; March 1995 release) database using the automatic sequence alignment program BLAST (Altschul *et al.*, 1990). There were several MoFe–protein sequences extracted from the database, Table 5.8 shows the residues found at the positions equivalent to the Asp 116, His 90 and Asp 160 residues in the MoFe protein structure. The sequences are all derived from nitrogen fixing bacteria, though the sequence identity with $1min$ is as low as 25%. The His residue is conserved in all but one case and there are several instances

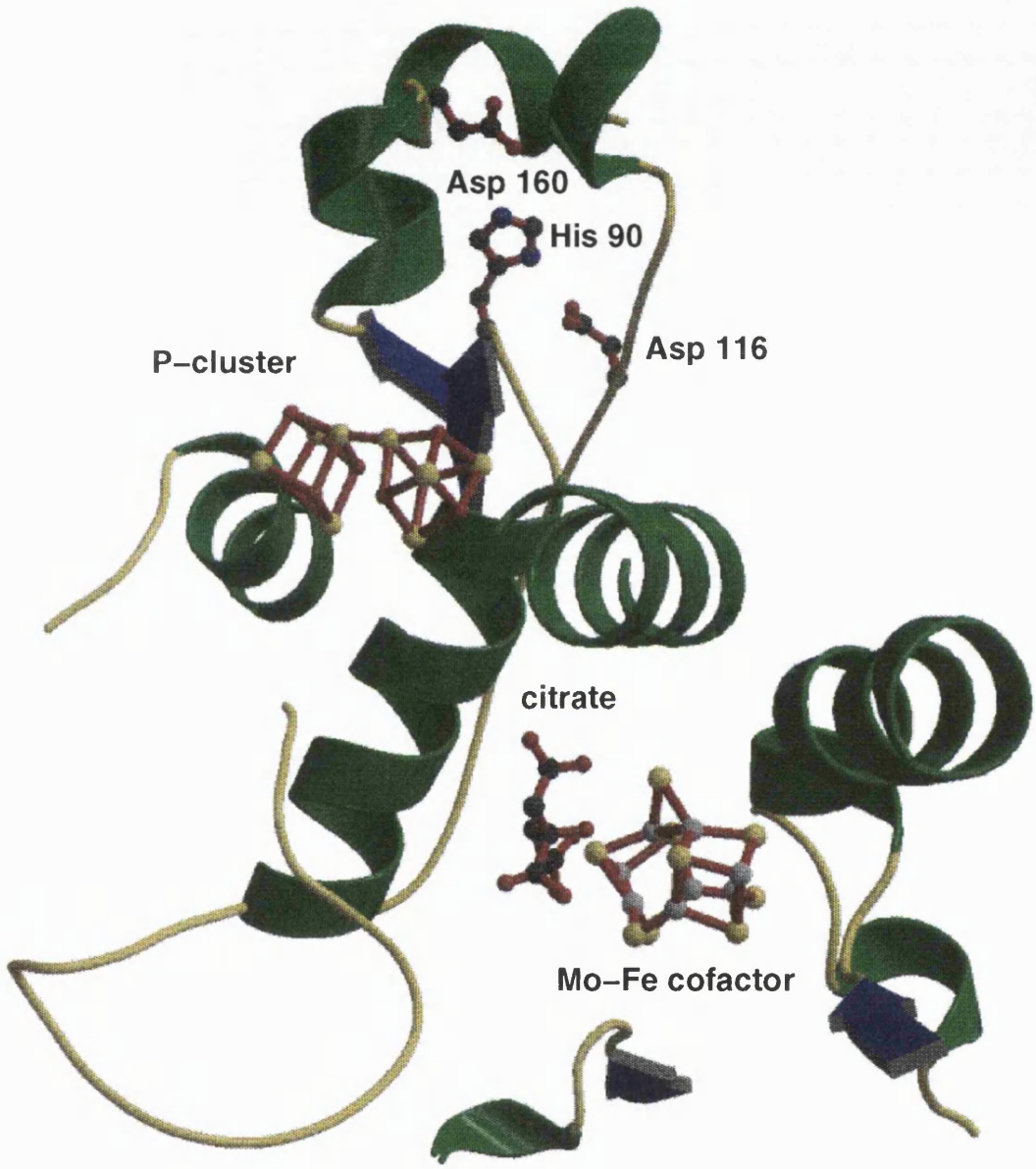


Figure 5.9: A view of the positional relationship of the P-cluster, the Asp–His–Asp triad and the MoFe cofactor associated with the nitrogenase enzyme (Jongsun *et al.*, 1992).

Swiss–Prot entry	Asp 160 $O^{\delta 2}$	– $N^{\epsilon 2}$ His 90 $N^{\delta 1}$	– Asp 116 $O^{\delta 2}$	sequence identity
NIFK-AZOVI	D	H	D	94
NIFK-KLEPN	D	H	D	63
NIFK-FRASP	D	H	D	59
NIFK-ANASP	D	Q	S	50
NIFK-BRAJA	D	H	S	71
NIFK-THIFE	D	H	S	51
NIFK-BRASP	D	H	S	55
NIFK-AZOBR	D	H	S	45
NIFK-CLOPA	D	H	S	35
ANFK-RHOCA	D	H	S	31
NIFN-BRAJA	D	H	T	38
ANFK-AZOVI	D	H	S	32
NIFN-KLEPN	D	H	T	30
NIFK-FRASP	D	H	D	56
VNFK-AZOVI	D	H	S	28
VNFK-AZOCH	D	H	T	26
NIFN-RHIME	D	H	T	28
NIFN-RHOCA	D	H	T	25
NIFN-AZOVI	D	H	T	26

Table 5.8: Results of a BLAST search on the D–chain of nitrogenase I analysing the conservation of Asp 160–His 90–Asp 116 triad from $1min$.

of Ser and Thr in the position of Asp 116. The Asp at position 160 is conserved in all cases.

This presents a possible region for further investigation and may give further insight into the MoFe–proteins catalytic mechanism.

5.5.3 pyruvate oxidase E.C.1.2.3.3

PDB code $1pox$ - rms distance 0.58Å

Lactic acid bacteria grow readily on surfaces exposed to air. During fermentation they degrade carbohydrates, yielding lactic acid. Since they lack cytochromes, they cannot produce ATP in a respiratory pathway so they convert lactic acid to acetate and ATP. Pyruvate oxidase is an enzyme involved in catalysing this reaction; its X–ray crystal structure has been solved from *Lactobacillus plantarum* to 2.1Å resolution by Muller *et al.* (1994). It is a homotetramer and a 3D representation of a monomer is shown is Figure 5.10. The Asp–His–Asp triad is on the opposite side of the monomer when compared to the thiamine pyrophosphate

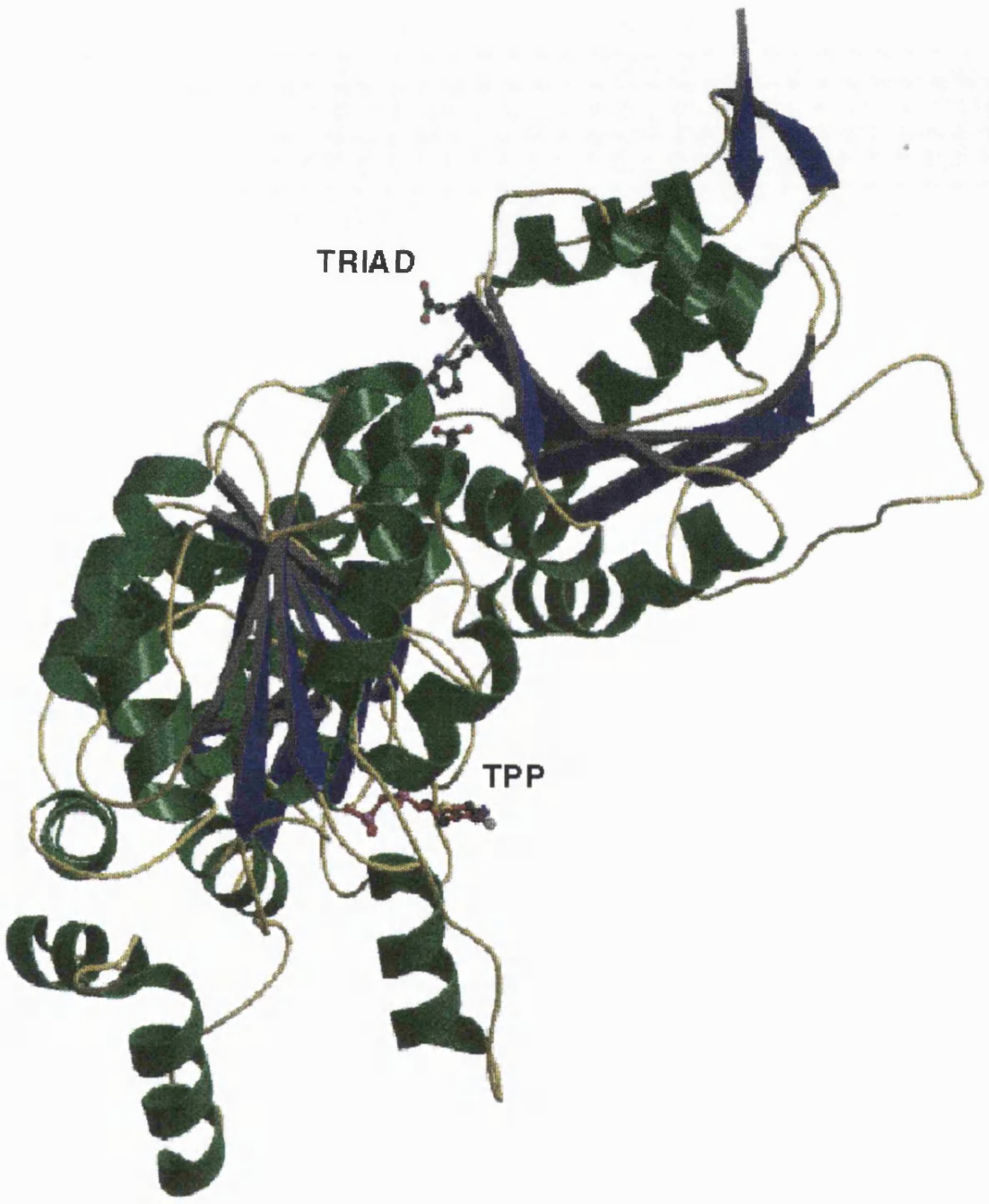


Figure 5.10: A 3D representation of a monomer of the enzyme pyruvate oxidase, 1*pox* (Muller *et al.*, 1994). The positions of the Asp–His–Asp triad and the cofactor thiamine pyrophosphate (TPP) are also shown.

(TPP) cofactor and would be on the outside of the protein in the tetrameric state.

A BLAST search of the SWISS–PROT database (March 1995 release) reveals that pyruvate oxidase has a typical thiamine pyrophosphate–binding sequence signature. Two SWISS–PROT (March 1995 release) identification codes are given for pyruvate oxidase: POXB-ECOLI (*E. coli*) and POXB-LACPL (*Lactobacillus plantarum*). Inspection of these sequences reveals that the POXB-ECOLI has Lys, Arg and Glu in the positions of Asp 27, His 28 and Asp 69 of the POXB-LACPL sequence respectively suggesting that the triad identified in 1*pox* is not functionally important.

5.5.4 macromomycin

PDB code 2*mcm* - *rms* distance 1.08Å

Macromomycin is the apoprotein of the antitumour antibiotic auromomycin which is a member of a large group of *Streptomyces* antibiotics including neocarzinostatin and actinoxanthin. Macromomycin carries and protects a nonprotein chromophore, which is the cytotoxic and mutagenic component of auromomycin. The nonprotein chromophore binds to and causes single and double strand breaks in DNA. The crystal structure of macromomycin has been determined to 1.6Å by Van Roey & Beerman (1989). Figure 5.11 shows that the overall structure of macromomycin is a seven–stranded β–barrel and two antiparallel β–sheet ribbons. The barrel and ribbons define a deep cleft which is occupied by two 2–methyl–2,4–pentanediols in the crystal structure and two chromophores *in vivo*. The Asp–His–Asp triad (red bonds) is shown on the surface of the molecule, away from the two chromophore binding sites.

A sequence search through the SWISS–PROT database (March 1995 release) reveals that there is only one macromomycin sequence present, however the se-

Figure 5.11: A 3D representation of the macromomycin apoprotein which has a seven–stranded β–barrel and two antiparallel β–sheet ribbons (Van Roey & Beerman, 1989). The positions of the Asp–His–Asp triad (red bonds) is also shown. The 2–methyl–2,4–pentanediol ligands in the crystal structure are the binding sites of chromophores *in vivo*.

quences of the related antibiotics (neocarzinostatin and actinoxanthin) are also available for comparison. neocarzinostatin has Ser, Asp and Ala at the positions equivalent to Asp 53, His 32 and Asp 100 in macromomycin, while actinoxanthin has Ser, Tyr and Ser. If the Asp–His–Asp triad in macromomycin does have a functional role, it is not shared by other members of this antibiotic family.

5.5.5 protein R2 of ribonucleotide reductase E.C.1.17.4.1

PDB code 1*rib* - *rms* distance 1.37Å

This enzyme catalyses the production of deoxyribonucleotides by reduction of ribonucleotides. It is a multi–subunit enzyme of type α_2/β_2 and constituted of homodimeric proteins denoted R1 and R2. The larger α_2 protein R1 has the binding site for substrate and allosteric effectors, whereas the smaller β_2 subunit has a dinuclear ferric centre and a stable tyrosyl radical for enzymatic activity. The X–ray structure of the R2 subunit from *E. coli* has been solved to 2.2Å by Nordlund & Ekland (1993).

The iron centre is, in concert with molecular oxygen, responsible for oxidation of Tyr 122 into a stable free–radical which is necessary for enzymatic activity. Figure 5.12 is a close up of the Asp–His–Asp triad (green bonds) with the iron–center (red bonds) and Tyr 122 (yellow bonds). Tyr 122 is clearly within interacting distance of the iron centre and it appears that the iron has displaced Asp 84 and is interacting electrostatically with His 118 N^{δ_1}, this explains the high *rms* distance of 1.37Å. The His 118–Asp 237 and Asp 84 are acting as ligands, binding the iron centre so it can perform its catalytic role. Indeed, the Asp 237 increases the ligand strength of the His 118. A search through the SWISS–PROT sequence database (March 1995 release) using the sequence from 1*rib* reveals that this triad is conserved across all species, from *E. coli* (*e.g.* RIR2_ECOLI) to humans

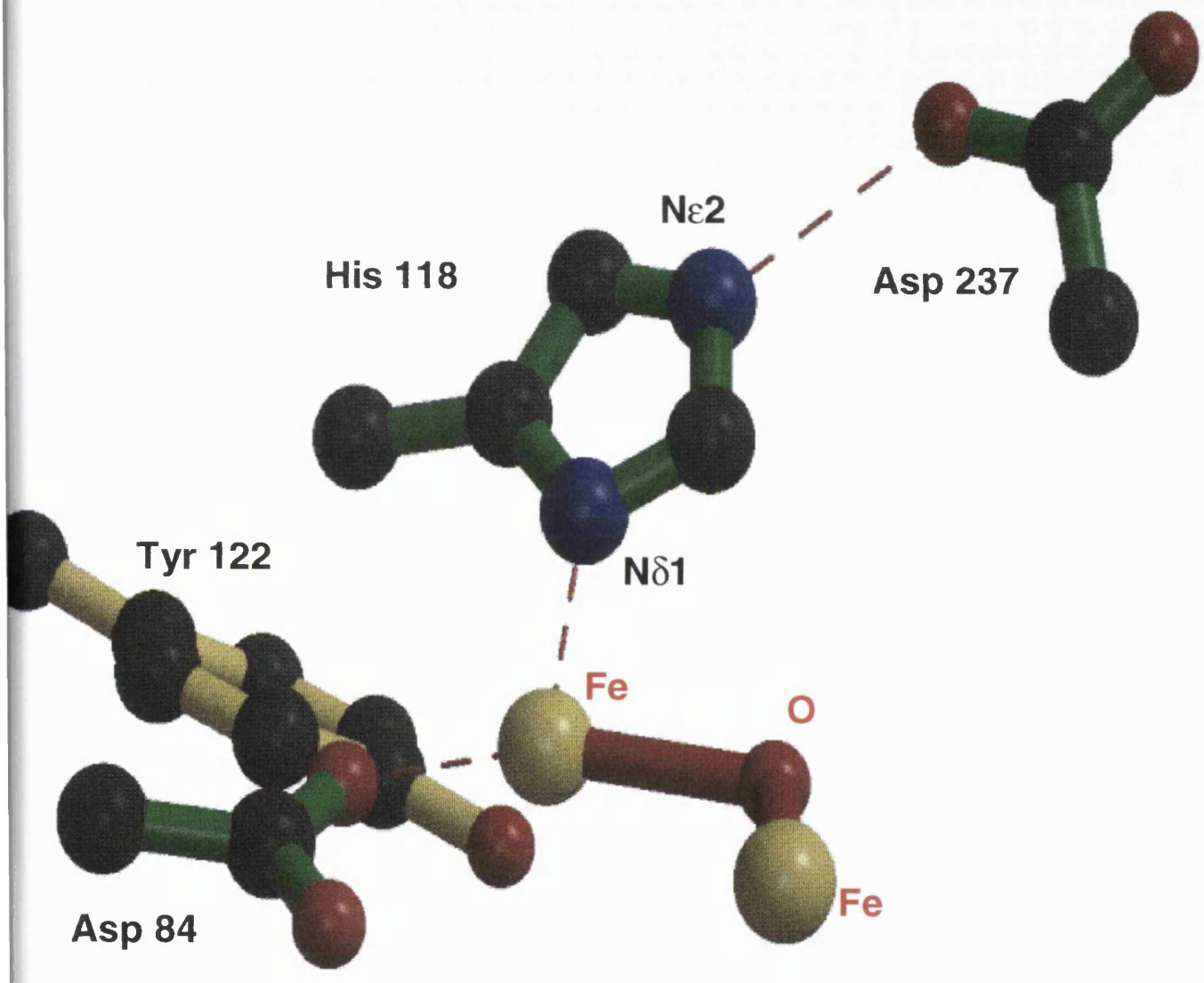


Figure 5.12: A view of the Asp 84–His 118–Asp 237 triad (green bonds) located in ribonuclease reductase (Nordlund & Ekland, 1993). The iron center (red bonds) oxidises Tyr 122 (yellow bonds) which is essential for catalytic activity of the enzyme.

(RIR2_HUMAN).

Christianson & Alexander (1990) first noted the similarity in the His–Asp diad of the serine proteinases and the diad that ligates Zn in carboxypeptidase, thermolysin and carbonic anhydrase. They have also noted that the His–Asp diad of the serine proteinases can selectively bind transition metals, including zinc. They therefore suggest a new catalytic triad of type Asp–His–metal; an example of this is the Asp 237–His 118–Fe of ribonuclease reductase. Therefore, although this is not a catalytic Asp–His–Asp triad of the type found in haloalkane dehalogenase, it does have an important functional role.

5.5.6 superoxide dismutase E.C.1.15.1.1

PDB code 1*sos* - *rms* distance 1.51Å

Superoxide is a by–product of aerobic metabolism and is produced in various reactions including oxidative phosphorylation and photosynthesis. Superoxide dismutase is responsible for catalysing the reduction of superoxide to oxygen and hydrogen peroxide. This enzyme can be thought of as a safeguard against oxygen toxicity and therefore tissue damage. The X–ray structure has been solved from human to 2.5Å resolution (Parge *et al.*, 1992); it has two bound metals, a zinc and a copper.

The Asp 124–His 71–Asp 83 triad located in the structure is illustrated in Figure 5.13, with the triad residues shaded in red. The Asp 124–His 71 forms a diad in a similar manner to the His–Asp acid/base diad of the serine proteinases, except that the His $N^{\epsilon 2}$ is hydrogen bonded to the Asp, rather than the $N^{\delta 1}$. In this case, Asp 124 increases the ligand strength of His 71, enabling stronger binding of the catalytic Zn. The Zn 155 metal is in the ideal hydrogen bonding position relative to the His 71 $N^{\delta 2}$; this is the position occupied by the Asp $O^{\delta 2}$

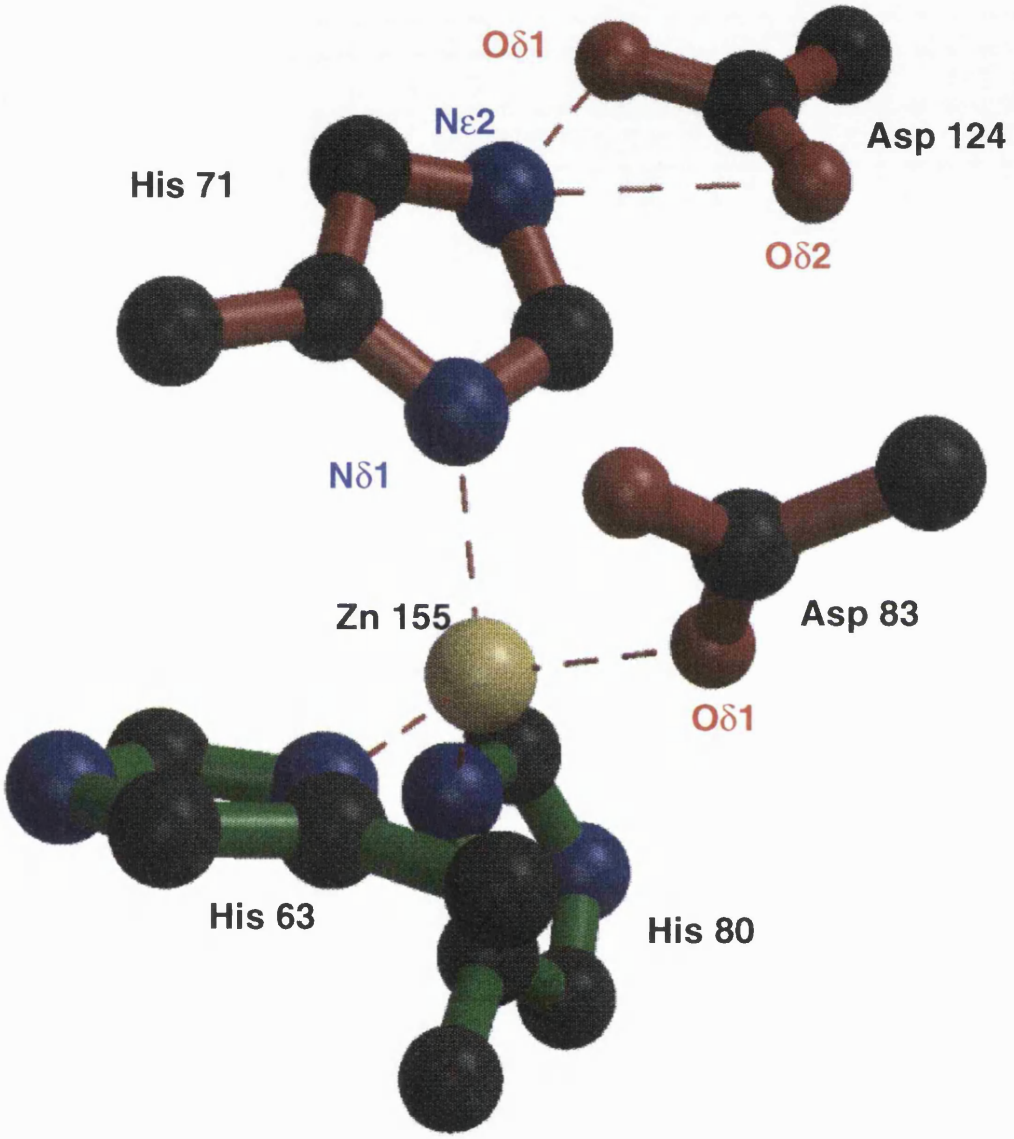


Figure 5.13: A 3D representation of the Asp 124–His 71–Asp 83 triad found in superoxide dismutase (Parge *et al.*, 1992). The triad residues are shaded in red.

in the catalytic Asp–His–Asp triad of haloalkane dehalogenase and explains the relatively high *rms* distance of 1.51Å. When the 1*sos* sequence is parsed against the SWISS–PROT database (March 1995 release), it is found that the Asp–His–Asp triad is conserved in all species present.

This structural and sequence information shows that the Asp–His–Asp triad found in this enzyme has a similar ligand binding function to the triad located in ribonuclease reductase and that noted by Christianson & Alexander (1990).

5.5.7 D–glyceralydehyde–3–phosphate dehydrogenase E.C.1.2.1.12

PDB code 1*gd*1 - *rms* distance 1.72Å

D–glyceralydehyde–3–phosphate dehydrogenase catalyses the oxidative phosphorylation of glyceraldehyde–3–phosphate to 1,3–biphosphoglycerate, yielding NADH in the process. The enzyme is part of the glycolytic pathway whereby glucose is metabolised to pyruvate and its X–ray structure has been determined to 3.0Å resolution, PDB code 1*gd*1 (Buehner *et al.*, 1974). It is a tetramer, with each subunit having two domains; a coenzyme binding domain which has the same α/β Rossmann– fold of the nucleotide binding domains of lactate dehydrogenase, malate dehydrogenase and liver alcohol dehydrogenase. There is also the substrate binding domain which has an $\alpha + \beta$ fold. These domains can be seen in Figure 5.14, also shown is the NADH molecule, the Asp–His–Asp triad (red bonds) and two catalytic residues, Cys 149 and His 176 (black bonds). The Asp–His–Asp triad lies between the two domains; the Asp 47 and His 50 occur in the nucleotide binding domain whereas Asp 312 lies in the binding domain.

Table 5.9 lists the results of a BLAST search of the 1*gd*1 sequence against the SWISS–PROT database (March 1995 release). There are various sequences from

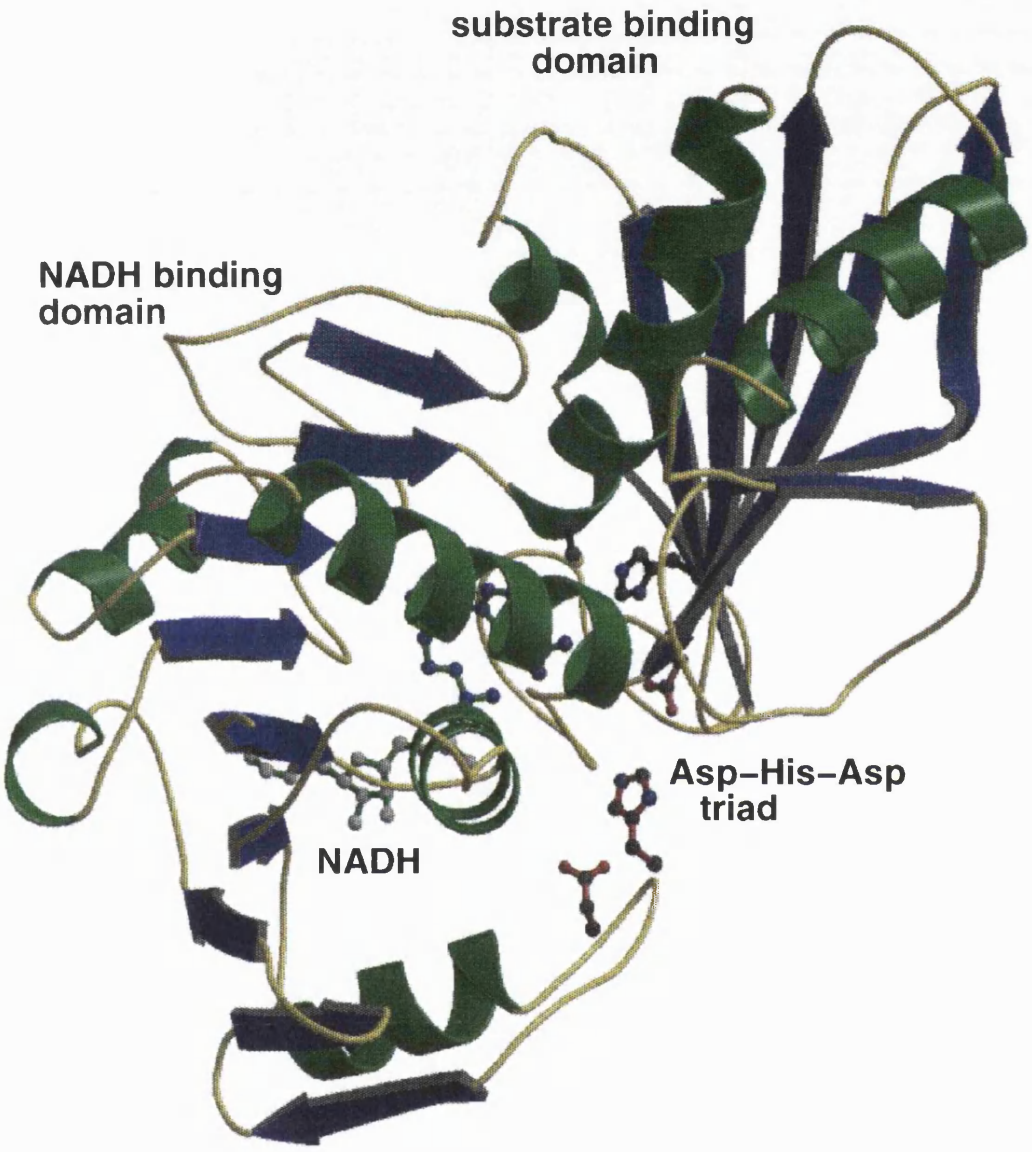


Figure 5.14: A 3D representation of the Asp 47–His 50–Asp 312 triad (red bonds) located in glyceraldehyde–3–phosphate dehydrogenase (Buehner *et al.*, 1974). Also shown are the two catalytic residues of Cys 149 and His 176 (black bonds) and the NADH coenzyme.

Swiss-Prot entry	Asp 47 $O^{\delta 2}$	- $N^{\epsilon 2}$ His 50 $N^{\delta 1}$ -	Asp 312 $O^{\delta 1}$
G3P-BACSU	D	H	D
G3P-BACME	D	H	D
G3P-THEMA	D	H	D
G3P2-ANAVA	D	L	D
G3P-THEAQ	D	Y	D
G3PB-PEA	D	L	D
G3PB-TOBAC	D	L	D
G3PB-ARATH	D	L	D
G3PB-SPIOL	D	L	D
G3P-CORGL	D	M	D
G3P1-ECOLI	D	H	D
G3PA-CHOCR	D	L	D
G3PC-TRYBB	D	H	D
G3PC-LEIME	D	H	D
G3P-ZYMMO	D	H	D
G3P1-ANAVA	D	H	D
G3P-KLULA	D	H	D
G3P3-ANAVA	D	H	D
G3P-EMENI	D	H	D
G3PC-PINSY	D	H	D

Table 5.9: Results of a BLAST search on the O-chain of glyceraldehyde-3-phosphate dehydrogenase to see if the D-H-D catalytic triad from 1*gd*1 (Buehner *et al.*, 1974) is conserved.

different species yet they all have a sequence identity of over 60%. The Asp residues of the triad are conserved, yet the His is replaced by a Met, Tyr or the hydrophobic residue Leu.

5.6 The Ser–His pair

In Chapter 3 we saw that the Ser O^{γ} nucleophilic group of the catalytic triad is not in a position to form an ideal hydrogen bond with the His $N^{\epsilon 2}$; the Ser O^{γ}'s role is to cleave a peptide bond, not to induce structural stability in the protein.

With this in mind, it would be interesting to see if the Ser O^{γ}'s distribution around the His differs in catalytic and ordinary Ser–His interactions. To investigate this, the consensus Ser O^{γ}–His–Asp O^{δ} template, without the Asp O^{δ} atom, was parsed against the representative 95% by sequence non-identical PDB dataset. Figure 5.15 shows the distribution of Ser O^{γ} atoms for non-catalytic (red) and catalytic (blue) atoms relative to the His sidechain. The catalytic atoms cluster in a non-ideal hydrogen bonding position whereas there is a cluster

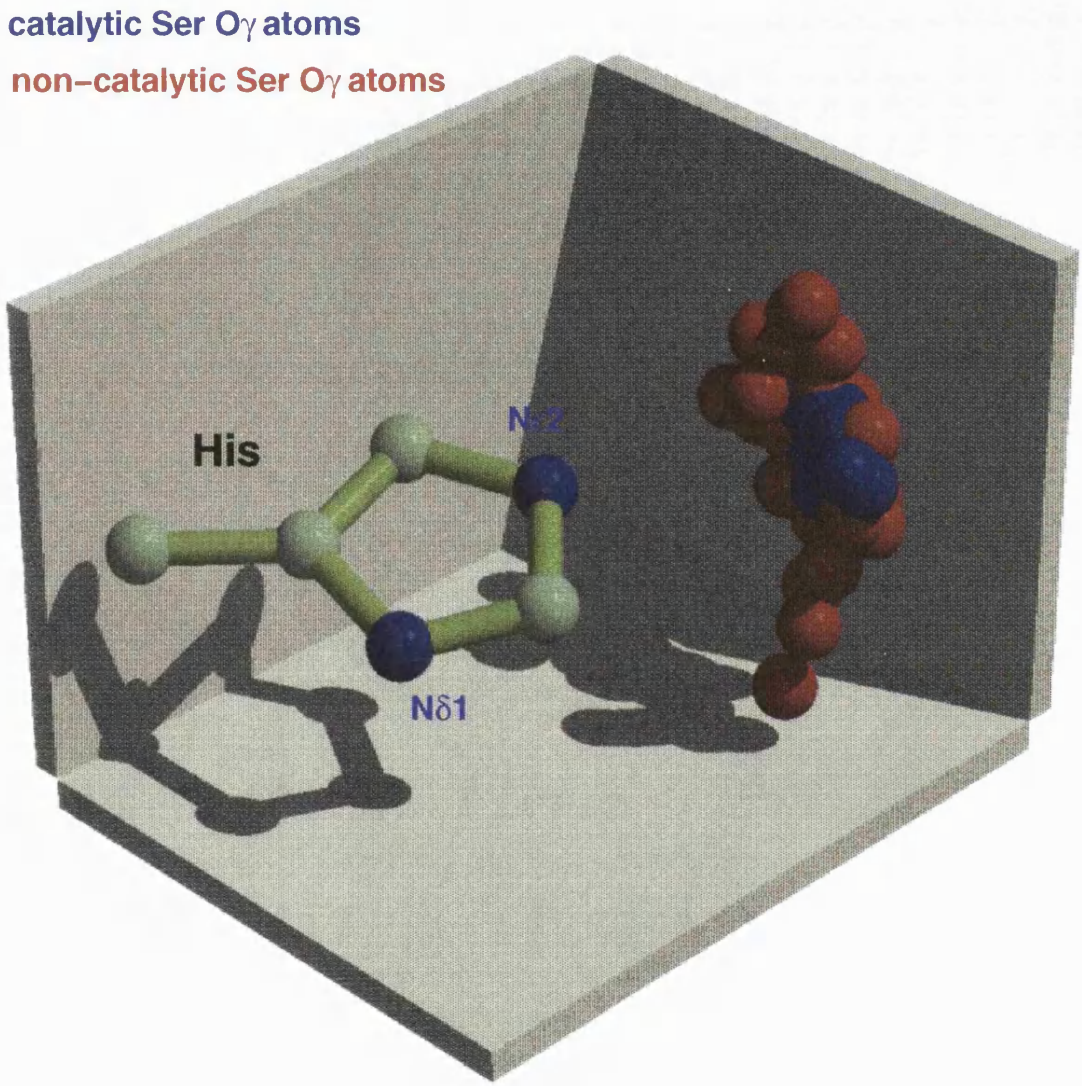


Figure 5.15: A 3D diagram showing the distribution of catalytic (blue) and non–catalytic (red) Ser O^{γ} atoms with respect to the His sidechain found in the 95% by sequence non–identical protein dataset.

of non–catalytic atoms in a position to form a strong hydrogen bond with the His $N^{\epsilon 2}$. There are however other non–catalytic Ser O^{γ} atoms in non–ideal hydrogen bonding positions. This shows that, though the catalytic Ser O^{γ} is in a distinct position with respect to the His sidechain, other factors determine whether it is catalytic. These include: the accessibility to ligands, the orientation of the Ser sidechain and the presence of an electrostatic group, such as the Asp in the His–Asp pair.

5.7 Conclusion

Having studied all the enzymes in the PDB with the catalytic residues of type Nu:–His–ELEC, it is striking how conserved in structure these triads are. Though the residue types of the Nu: and ELEC groups can vary according to enzyme type we find that the positions of the functional atoms in the triads are conserved. With the exception of the cysteine proteinases, one template is able to define the active site of all the serine proteinases, acetylcholinesterase and haloalkane dehalogenase. This suggests that convergent evolution has drawn the functional atoms into optimal catalytic positions. In addition, these triads are confined to the active sites of enzymes though we did find several examples of other unidentified Nu:–His–ELEC triads which may have biological relevance. This indicates that as the number of protein structures deposited in the PDB increases other common 3D templates, functional in more than one enzyme type, will occur.

5.8 References

Altschul, S.F., Gish W., Miller W., Eugene W.M & Lipman D.J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215** 403-410.

Baker E.N. (1980) Structure of Actinidin, after refinement at 1.7Å resolution *J. Mol. Biol.* **141** 441–466

Baker E.N. & Drenth J. (1987) In Jurnak F.A. and McPherson A. (eds.) *Biological Macromolecules and Assemblies* Wiley, New York **3** 341–348

Bairoch A. & Boeckmann B. (1994) The SWISS–PROT protein sequence databank: current status *Nucleic Acid Research* **22** 3578–3589

Blow D. M., Birktoft J. J. & Hartley B.S. (1969) Role of a buried acid group in the mechanism of action of chymotrypsin *Nature* **221** 337–340

Buehner M., Ford G.C., Moras D., Olsen K.W. & Rossmann M.G. (1974) Three-dimensional structure of D–glyceraldehyde–3–phosphate dehydrogenase *J. Mol. Biol.* **99** 25–49

Brady L., Brzozowski A. M., Derewenda Z. S., Dodson E., Dodson G., Tolley S., Turkenburg J. P., Christianson L., Huge Jensen B., Norskov L., Thim L. & Menge U. (1990) A serine protease triad forms the catalytic centre of triacylglycerol lipase *Nature* **343** 767–770

Christianson D.W. & Alexander R.S. (1989) Carboxylate–histidine–zinc interactions in protein structure and function *J. Am. Chem. Soc.* **111** 6412–6419

Christianson D.W. & Alexander R.S. (1990) Another catalytic triad? *Nature* **346** 225

Drenth J., Jansonius J.N., Koekoek R., Swen H.M. & Wolthers B.G. (1968) Structure of Papain *Nature* **218** 929–934

Findlay D., Herries D.G., Mathais A.P., Rabin B.R. & Ross C.A. (1962) The active site and mechanism of action of bovine pancreatic ribonuclease 7. The catalytic mechanism, *Biochem J.* **85** 152–153

Franken S.M., Rozeboom H.J., Kalk K.H. & Dijkstra B.W. (1991) Crystal structure of haloalkane dehalogenase: an enzyme to detoxify halogenated alkanes *Embo J.* **10** 1297–1302

Fujinaga M., Sielecki A.R., Read R.J., Ardelt W., Laskowski M. & James M.N.G. (1987) Crystal and molecular structures of the complex of α–chymotrypsin with its inhibitor turkey ovomucoid domain at 1.8Å resolution. *J. Mol.Biol.* **195** 397–418

Fujinaga M. & James M.N.G (1987) Rat submaxillary serine proteinase, tonin. Structure solution and refinement at 1.8Å resolution *J. Mol. Biol.* **195** 373–391

Grochulski P., Li Y., Schrag J. D., Bouthillier F., Smith P., Harrison D., Rubin B. & Cygler M. (1993) Insights into interfacial activation from an 'open' structure of *Candida rugosa* lipase *J. Biol. Chem.* **268** 12843–12847

Harel M., Schalk I., Ehret–Sabattier L., Bouet F., Goeldner M., Hirth C., Axelson P., Silman I. & Sussman J. (1993) Quaternary ligand binding to aromatic residues in the active–site gorge of acetylcholinesterase *Proc. Nat. Acad. Sci. (USA)* **90** 9031–9035

Hirs C.H., Halmann M. & Kycia J.H. (1965) Dinitrophenylation and inactivation of bovine pancreatic ribonuclease A *Arch. Biochem Biophys.* **111** 209–222

Jongsun K. & Rees D.C. (1992) Crystallographic structure and functional implications of the nitrogenase molybdenum–iron protein from *Azotobacter vinelandii Nature* **360** 553–560

Kim M.J., Yamamoto D., Matsumotot K., Inoue M., Ishida T., Mizuno H., Sumiya S & Kitamura K. (1992) Crystal structure of Papain-E64-C com-

plex. Binding diversity of E64-C to Papain S-2 and S-3 subsites *Biochem J.* **287** 797–803

Liao D-I., Breddam K., Sweet R., Bullock T. & Remington S. J. (1992) Refined atomic model of wheat serine carboxypeptidase II at 2.2Å resolution *Biochemistry* **31** 9796–9812

Muller Y.A., Schumacher G., Rudolph R. & Schulz G.E. (1994) The refined structures of a stabilized mutant and of wild–type pyruvate oxidase for *Lactobacillus plantarum J. Mol. Biol* **237** 315–334

Nordlaund P., & Eklund H. (1993) Structure and function of the *Escherichia coli* ribonucleotide reductase protein R2 *J. Mol. Biol.* **232** 123–164

Ollis D.L., Cheah E., Miroslaw C., Dijkstra B., Frolow F., Franken S.M., Harel M., Remington S.J., Silman I., Schrag J., Sussman J.L., Verschueren K.H.G. & Goldman A. (1992) The α/β hydrolase fold *Protein Engineering* **5** 197–211

Parge H.E., Hallewell R.A. & Tainer J.A. (1992) Atomic structures of wild–type and thermostable mutant recombinant Cu, Zn superoxide dismutase *Proc. Natl. Acad. Sci* **89** 6109–6113

Pickersgill P.W., Rizkallow P., Harris G.W. & Goodenough P.W. (1991) Determination of the structure of papaya ω. *Acta. Crystallogr., Sect. B* **47** 766–780

Shimada Y., Sugihara A., Izumi T. & Tominaga Y. (1990) cDNA cloning and characterisation of a novel thermostable lipase from *Bacillus sp. J. Biochem* **107** 703–707

Sussman J.L., Harel M., Frolow F., Oefner C., Goldman A., Toker L. & Silman I. (1991) Atomic–structure of acetylcholinesterase from *Torpedo californica* - A protoypic acetylcholine–binding protein *Science* **253** 872–879

Varughese K.I., Su Y., Cromwell D., Hasnain S. & Xuong N.H. (1992) Crystal structure of actinidin-E-64 complex *Biochemistry* **31** 5172–5176

Verscheuren H.G., Seljee F., Rozeboom HJ., Kalk K.H & Dijkstra B.W. (1993) Crystallographic analysis of the catalytic mechanism of haloalkane dehalogenase *Nature* **363** 693–698

Wallace A.C., Laskowski R.A. & Thornton J.M. (1996) Derivation of 3D coordinate templates for searching structural databases: application to the Ser–His–Asp catalytic triads in the serine proteinases and lipases *Protein Science* **5** 1001–1013

Wang J., Xiang Y-F. & Lim C. (1994) The double catalytic triad, Cys 25–His 159–Asp 158 and Cys 25–His 159–Asn 175, in papain catalysis: role of Asp 158 and Asn 175 *Protein Engineering* **7** 75–82

Wright C. S., Alden R. A. & Kraut J. (1969) Structure of subtilisin *bpn* at 2.5Å resolution. *Nature* **221** 235–242

Chapter 6

Catalytic residues and ligand binding sites

6.1 Introduction

In chapter 5 we saw that one 3D consensus template is able to describe the active site of the serine–proteinases, lipases and the α/β–hydrolase enzymes. This indicates that a common geometry can occur in the catalytic residues of enzymes of diverse function and fold. The template consisted of just the catalytic Nu: atom, ELEC atom and the His sidechain; it did not take into account the relative orientation of the Nu: and ELEC sidechains with respect to the His for the different enzymes in the group. In fact, as Figure 6.1 shows, these sidechain groups originate from different orientations.

We now wish to see if the orientation of the catalytic sidechain residues of these enzymes is determined by the relative position of the ligands. Actually, we would only expect the orientation of the Nu: sidechain to be influenced as it needs to be in the ideal geometry to interact directly with the substrate.

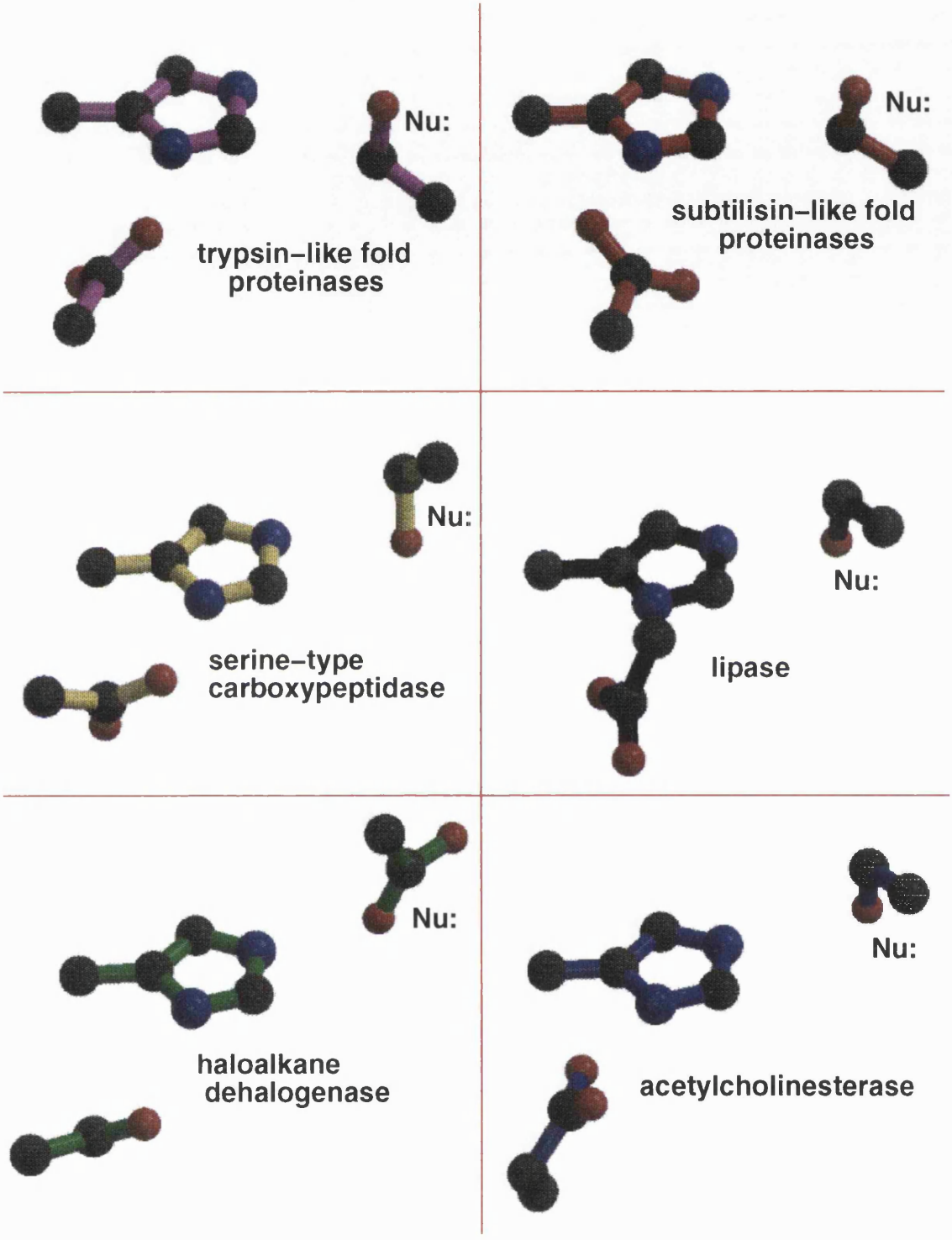


Figure 6.1: A 3D representation of the catalytic triads of the serine–proteinases, lipases and α/β–hydrolase fold enzymes.

6.1.1 Identification of ligands bound in active site

For a given protein structure, we identified the ligands bound to the active site as this will give us a good indication of the position of the ligand binding site. This is a far from straightforward process, but it is possible because we know the location of the active site in our protein structures from the consensus templates. For this study, we defined a ligand atom by two criteria: firstly, it was within 15Å of the catalytic residues and secondly, it was not part of the same polypeptide chain as any of the catalytic residues.

The ligands located for each of the enzymes are summarised in Table 6.1. Analysis of the Ser–His–Asp catalytic triad in chapter 3 indicated that those triad's with an *rms* deviation greater than 1.4Å from the Nu:–His–ELEC consensus template had non–native inhibitors bound to their active site that perturbed the catalytic triads geometry. Conversely, those triads with an *rms* less than 1.4Å have either peptide inhibitors or are the *apo*–form of the enzyme. For this reason we have divided the ligands according to whether the catalytic triad they are associated with in the protein structure has an *rms* greater to or less than 1.4Å from the mean Nu:–His–ELEC consensus template.

6.1.2 Method to compare ligand binding site conformation

In chapters 3 and 5 we saw that for each of the serine–proteinase, lipase and α/β hydrolase fold enzymes, a sidechain consensus template (*e.g.* for the serine proteinases, Ser C^{α}, Ser C^{β}, Ser O^{γ}, Asp C^{α}, Asp C^{β}, Asp O^{δ_1}, Asp O^{δ_2} and the His sidechain) was constructed that could identify every sidechain Nu:–His–ELEC catalytic triad of each enzyme group member. We now wish to compare the ligand orientations around the sidechain triads of each of these enzymes.

Group	Number PDB chains	Ligands < 1.4Å	Ligands > 1.4Å
trypsin–like fold proteinases	167	117	22
subtilisin–like fold proteinases	35	15	6
serine type carboxypeptidase	7	2	0
lipases	13	2	0
α/β–hydrolase enzyme: haloalkane dehalogenase	9	7	0
α/β–hydrolase enzyme: acetylcholinesterase	7	4	0

Table 6.1: A summary of the number of ligands found in the dataset of serine–proteinases, lipases and α/β–hydrolase fold enzymes. The ligands are divided according to those whose sidechain catalytic triads are greater or less than 1.4Å from the NU:–His–ELEC consensus template.

Taking each enzyme group in turn, every member had its catalytic Ser–His–Asp sidechains transformed onto that group's sidechain Ser–His–Asp consensus template using the TESS program described in chapter 4. In addition, the transformation matrix was applied to the ligand (if any) for that structure. This means that the catalytic triad for every member of that enzyme group will be superimposed and will allow us to compare the relative orientation of the ligands.

6.1.3 Comparing the ligand binding sites

The trypsin–like fold proteinases

This is the largest group with 167 peptide chains. Figure 6.2 is a 3D representation of the distribution of inhibitors around the Ser–His–Asp sidechain consensus template. The position of the inhibitors is represented by contours generated by the program SURFNET (Laskowski, 1995); these are the positions where the

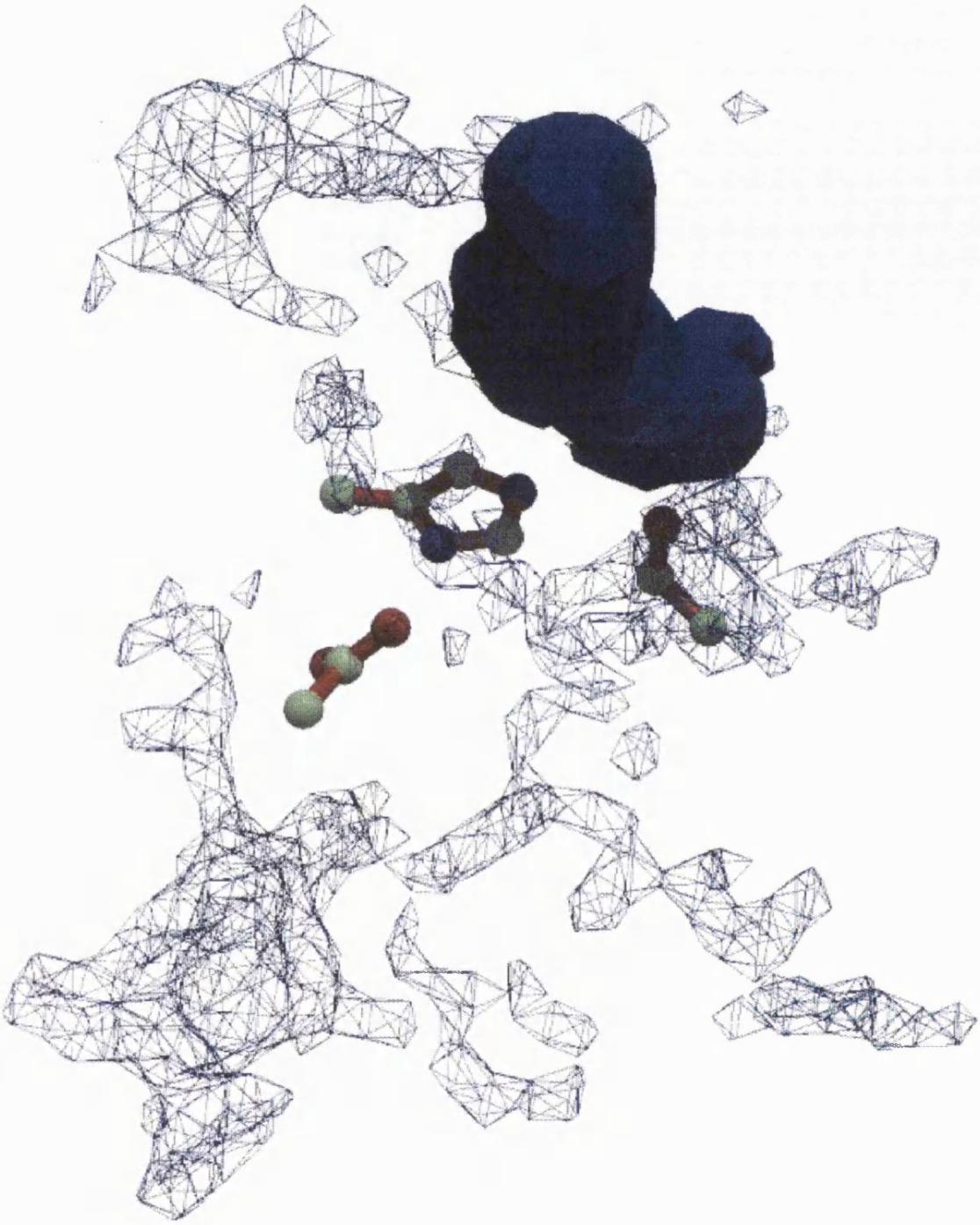


Figure 6.2: A 3D representation showing the distribution of ligands around the Ser–His–Asp consensus template for the trypsin–like proteinases. The grid–like contours in blue represent the ligands from those structures whose triads are more than 1.4Å from the consensus template whereas the solid contours are ligands from triads whose *rms* deviation is less than 1.4Å from the consensus template

inhibitor atoms cluster around the catalytic triad. The grid–like contours are the position of the inhibitors for structures whose Nu:–His–ELEC triads are above 1.4Å *rms* distance whereas the solid contours are less than this value. Clearly, the low *rms* contours cluster in a tight position directly above the nucleophilic Ser O^{γ}, whereas high *rms* inhibitors are dispersed in a large volume around the Ser–His–Asp triad.

The subtilisin–like fold proteinases

Table 6.1 shows that this group has 35 ligands with triads less than the 1.4Å deviation cut–off; this should give us a good indication of the position of the ligand binding site. Figure 6.3 shows the relative position of the inhibitors for this group of proteinases. The ligand binding sites lie directly above the Ser sidechain and suggests that the sidechain of the nucleophilic Ser is orientated so its O^{γ} atom is in close proximity to the scissile peptide bond of the substrate. Note also that these ligands are in approximately the same position as those for the trypsin–like proteinases.

The serine–type carboxypeptidase

The serine–type carboxypeptidase group has only 2 ligands both of which are non–protein benzylsuccinate from 1*whs* and 1*wht* (Liao *et al.*, 1992). This type of inhibitor does not appear to affect the conformation of the Ser–His–Asp triad as they are only 0.68Å and 0.80Å respectively from the Nu:–His–ELEC consensus template. This is reflected in the inhibitor contours in Figure 6.4; though not directly above the Ser O^{γ} electrostatic group, they are in its vicinity.

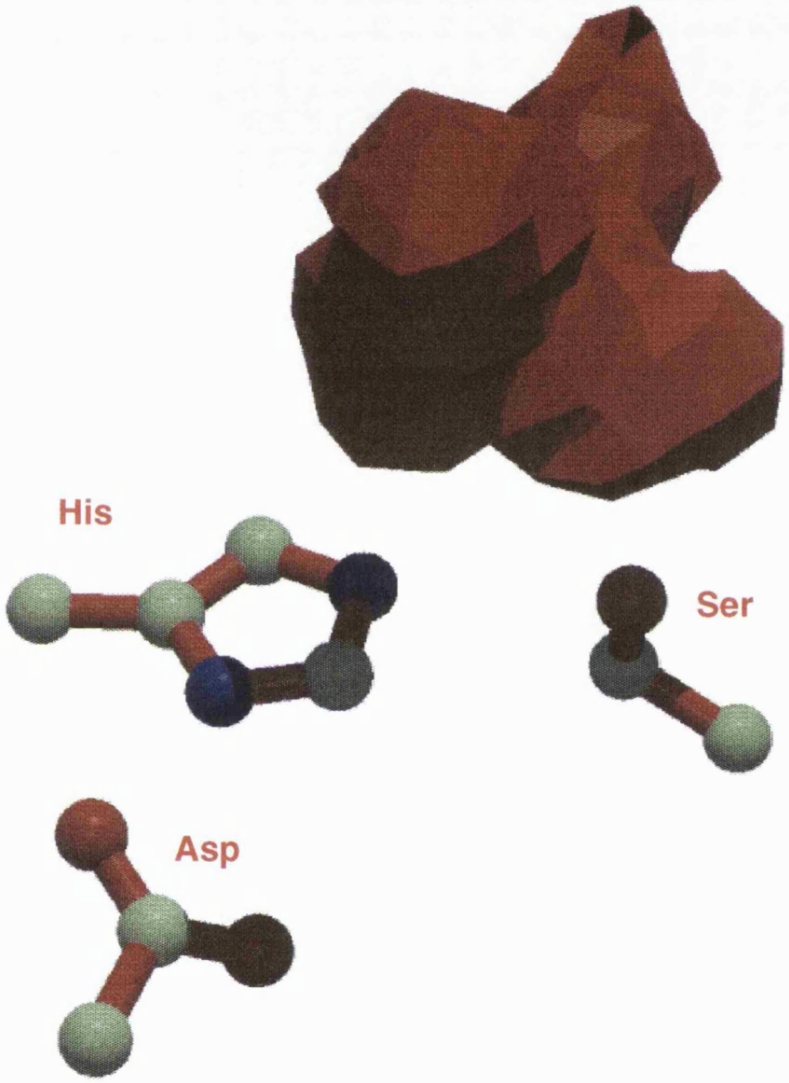


Figure 6.3: A 3D representation of the position of the inhibitors relative to the sidechain consensus templates for the subtilisin–like proteinases. The inhibitors were extracted from structures whose triads had an *rms* less than 1.4Å from the Nu:–His–ELEC consensus template.

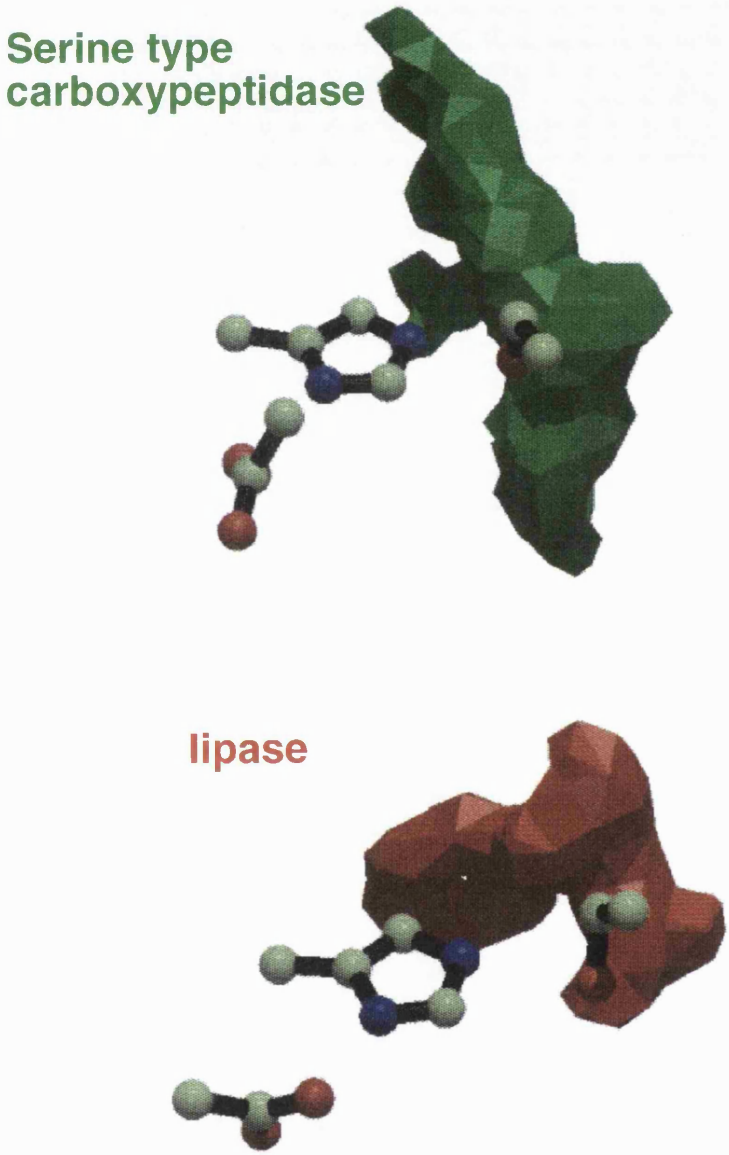


Figure 6.4: A 3D representation of the position of the inhibitors relative to the sidechain consensus templates of serine–type carboxypeptidase and lipase.

Lipase

There are only two lipase structures complexed with inhibitors: *1lpa* and *1lpb* both with an 11–carbon alkyl phosphonate (Van Tilbergh *et al.*, 1993). This inhibitor does not alter the conformation of the catalytic triad in the structure as it is only 0.53Å and 0.83Å respectively from the Nu:–His–ELEC consensus template. Figure 6.4 shows it binds in the vicinity of the nucleophilic Ser sidechain and is in a good position to interact with the Ser O^{γ} atom.

The α/β–hydrolase fold enzymes

There are two members of the α/β–hydrolase fold family: acetylcholinesterase with a Ser–His–Glu catalytic triad and haloalkane dehalogenase with an Asp–His–Asp triad (chapter 5). Figure 6.1 shows that the nucleophilic Asp and Ser sidechains are pointing in different directions when compared to the trypsin–like fold catalytic Ser. We would therefore expect the ligand binding sites to be in different positions also.

For haloalkane dehalogenase, 7 of the 9 structures have inhibitors bound to them; of these, *2dhd* and *2dhc* (Verschueren *et al.*, 1993) are structures with the substrate, dichloroethane, and the acyl–enzyme intermediate bound to the active site. The other 5 structures have either Cl or I atoms as ligands.

Figure 6.5 clearly shows the contours representing the ligands sitting directly below the nucleophilic Asp residue. In addition, they are placed below the plane of the His ring whereas the inhibitors of the serine–proteinases in Figure 6.2 and 6.3 are above it.

There are 4 inhibitors present in the acetylcholinesterase structures (Sussman *et al.*, 1991); Figure 6.5 shows them to be in close proximity to the Ser nucleophilic sidechain indicating the importance of the orientation of the nucleophilic sidechain with respect to the ligand binding site.

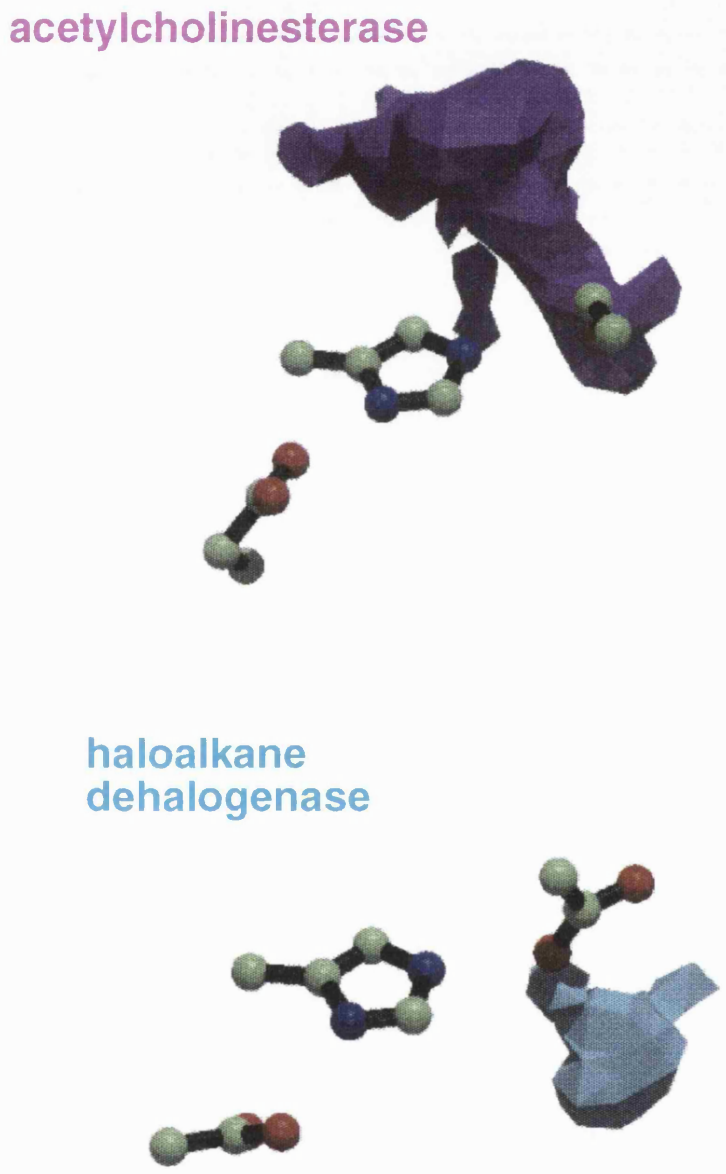


Figure 6.5: A 3D representation of the of the position of the inhibitors relative to the sidechain consensus templates for haloalkane dehalogenase and acetylcholinesterase (Sussman *et al.*, 1991).

6.1.4 Superposition of all ligand binding sites

Figure 6.6 shows all ligand contours described above superimposed and it summarises the heterogeneity in the conformation of these sites with respect to the catalytic triad. It is also worth noting that the binding site is orientated in such a way that neither the His nor ELEC group interacts directly with the ligand. Obviously, such an interaction would compromise the roles these residues have as the acid/base catalyst.

6.1.5 Conclusion

There is a clear relationship between the orientation of the Nu: sidechain of the catalytic triad and the ligand binding site; this group is always orientated so its nucleophilic atom is able to interact with the substrate. In addition, the ligand binding site is identifiable automatically with prior knowledge of the catalytic residues. Therefore, as the database of 3D enzyme active site templates increases, it will also be possible to produce a database of ligand binding sites automatically.

6.2 References

Bone R., Fujishige A., Kettner C.A. & Agard D.A. (1991) Structural basis for broad specificity in α–lytic protease mutants *Biochemistry* **30** 10388–10398

Laskowski R.A. (1995). SURFNET: A program for visualizing molecular surfaces, cavities and intermolecular interactions. *J. Mol. Graph.* **13** 323–330.

Liao D-I., Breddam K., Sweet R., Bullock T. & Remington S. J. (1992) Refined atomic model of wheat serine carboxypeptidase II at 2.2Å resolution *Biochemistry* **31** 9796–9812

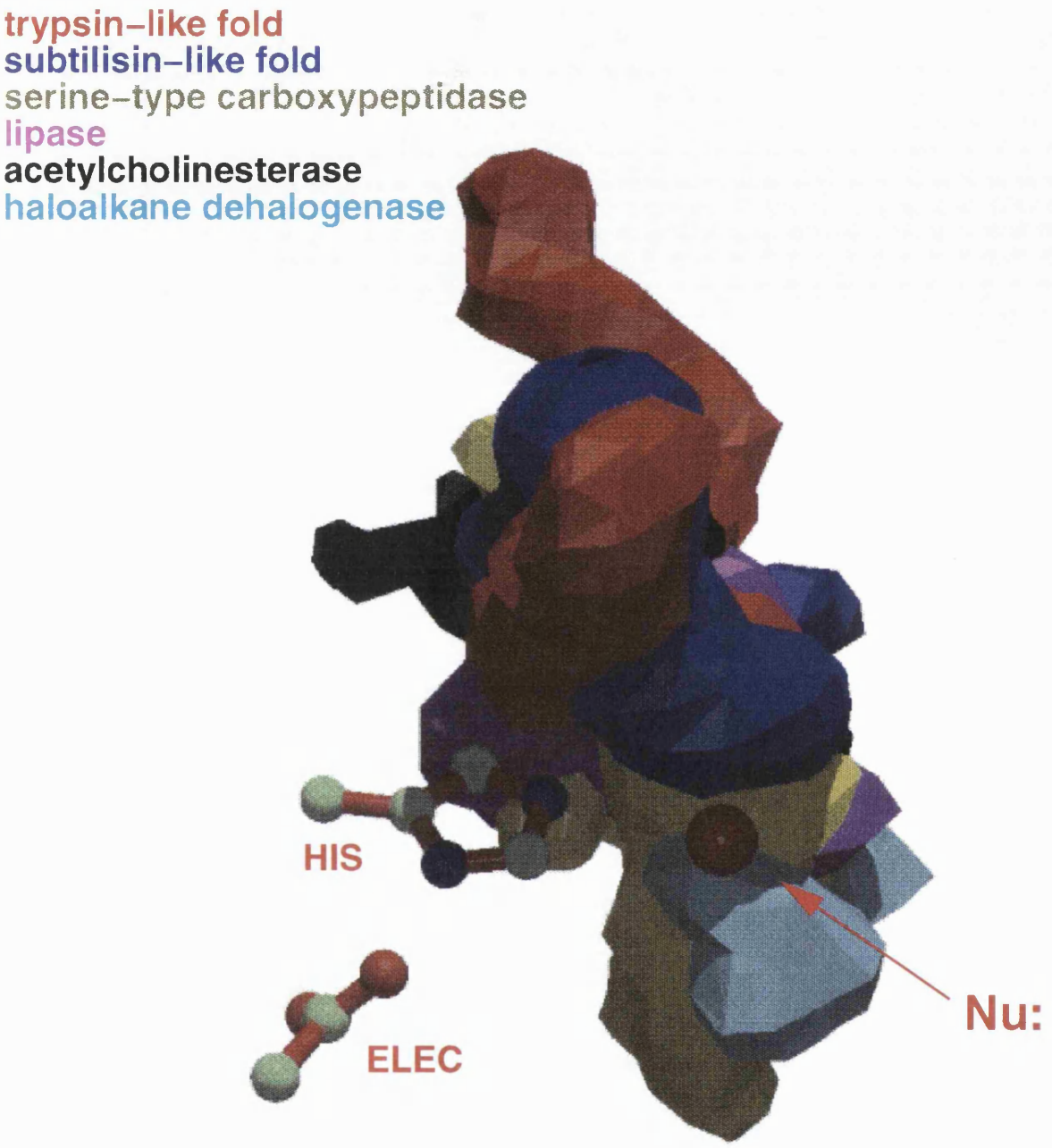


Figure 6.6: A 3D representation of the distribution of the ligands from the serine proteinases, lipases and α/β–hydrolase fold enzymes with respect to the Asp and His sidechain and the Nu: atom.

Sussman J.L., Harel M., Frolow F., Oefner C., Goldman A., Toker L. & Silman I. (1991) Atomic–structure of acetylcholinesterase from *Torpedo californica* - A protoypic acetylcholine–binding protein *Science* **253** 872–879

Van Tilbergh H., Egloff M.P., Martinez C., Rugani N., Verger R. & Cambillau C. (1993) Interfacila activation of the lipase-procolipase complex by mixed micelles revealed by X–ray crystallography *Nature* **362** 814–820

Timkovich R. & Dickerson R.E. (1976) The structure of *Paracoccus denitrificans* from cytochrome C551 *J. Biol. Chem.* **201** 4033–4062

Chapter 7

The role of His in metal binding sites

7.1 Introduction

Metals are ubiquitous in all organisms and have a diverse range of functions. These include metabolic regulation, nerve transmission, muscle contraction and cell motility, cell division, growth, secretion and membrane permeability.

One of the ways metals induce these physiological effects is to bind to proteins. They can stabilise protein structure, for example Ca^{2+} has binding sites in both elastase (Li De La *et al.*, 1990) and trypsin (Marquart *et al.*, 1983). Secondly, metals can modulate protein action; calmodulin (Babu *et al*, 1985) is Ca^{2+} modulated in its biological interactions with other proteins. Thirdly, metals can be directly involved in the enzymatic activity of proteins, for example the matrix metalloproteinases collagenase (Borkakoti *et al.*, 1994), stromelysin and gelatinase (Murphy *et al.*, 1991; Docherty *et al.*, 1992) all have a bound zinc in the enzyme active site which is directly involved in catalysis.

When a metal binds to a protein it ligates to several groups in the protein

Classification	Cation				Ligand			
Hard	H^+	Li^+	Na^+	K^+	H_2O	OH^-	ROH	NH_3
	Mg^{2+}	Ca^{2+}	Mn^{2+}	Cr^{3+}	CO_3^{2-}	$RCOO^-$		
	Co^{3+}							
Soft	Cu^+	Ag^+	Au^+	Tl^+	RSH	RS^-	R_2S	CN^-
	Pd^{2+}	Pt^{2+}	Cd^{2+}	Hg^+	H^-	I^-		
Borderline	Zn^{2+}	Cu^{2+}	Ni^{2+}	Fe^{2+}	Pyridine	RNH_2	Imidazole	
	Fe^{3+}	Co^{2+}	Sn^{2+}	Pb^{2+}				
	Pb^{2+}	Rh^{3+}	Ir^{3+}	Ru^{3+}				

Table 7.1: The 'hard' and 'soft' classification of the Lewis acids and bases.

structure and acts as a Lewis acid, accepting a lone pair of electrons from the Lewis base amino acid atoms. The major metal binding amino acids in proteins (Gurd & Wilcox, 1956; Voet & Voet, 1990) are carboxyl (aspartate and glutamate), imidiazole (histidine), indole (tryptophan), thiol (cysteine), thioester (methionine) and hydroxyl (serine, threonine, and tyrosine). The number of atoms that are packed around the metal (the coordination number) is dependent on the size, charge and polarisability of the metal and ligand.

Some metals are polarisable, that is, when they are placed in an electric field, there tends to be charge separation. The amount of polarisability leads to the concept of soft and hard metals, these are summarised in Table 7.1. In general 'hard' metals coordinate to 'hard' ligands and vice versa. Most metals and ligands of biochemical interest are hard or borderline, the exception being thiols and hydride ions. Zinc is a metal of borderline hardness and accommodates nitrogen, oxygen and sulfur in its coordination polyhedra whereas divalent ions of calcium or magnesium are hard and in general only ligate to oxygen atoms in protein structures. Table 7.2 summarises the preferred complexes made by metals. Vallee & Auld (1990) have shown that catalytic zinc sites have a binding

Property	Na^+,K^+	Mg^{2+},Ca^{2+}	Zn^{2+},Cd^{2+},Co^{2+} Cu^{2+},Fe^{2+},Mo^{2+}
Complex formation	Weak	Moderate	Strong
Preferred ligand atom	O	O	N and S

Table 7.2: Complex formation properties of metals in biochemistry.

frequency of His >> Glu > Asp = Cys. Cys is a soft ligand; its outer orbital electrons are polarised toward the Zn metal forming a partial π–bond, thereby reducing the polarisation and catalytic strength of the Zn metal. The His residue is the most suitable because it is a hard ligand and not easily polarisable; it may aid Zn in promoting the nucleophilicity of a bound solvent molecule.

The amino acid sidechains which ligate the zinc in metalloproteins are densely packed within the protein structure and often make hydrogen bond contacts with other residues. Argos *et al.*, 1978 pointed out that such interactions may both orientate and enhance the electrostatic force between the metal ions and its ligands. An example of this type of interaction is the metal–histidine–carboxylate triad. This is found in the active sites of several enzymes. Figure 7.1 is a diagrammatic representation of the two possible tautomeric forms of this triad. The conformations depend on which His nitrogen atom ($N^{\epsilon 2}$ or $N^{\delta 1}$) interacts with the metal; in this chapter tautomer ϵ is defined as His $N^{\epsilon 2}$ interacting with metal and tautomer δ is a His $N^{\delta 1}$–metal interaction. The sidechains of Asp, Asn, Glu, or Gln hydrogen bonded to a His increases its pK_a by 2 units (Carver & Bradbury, 1984) and therefore also its ligand strength; if a His makes a hydrogen bond with a carbonyl the ΔpK_a is about half as much (Perutz *et al.*, 1985). The histidine–carboxylate hydrogen bond also reduces the entropic barrier to the organisation of the metal binding site, since the hydrogen bond orientates the His into the

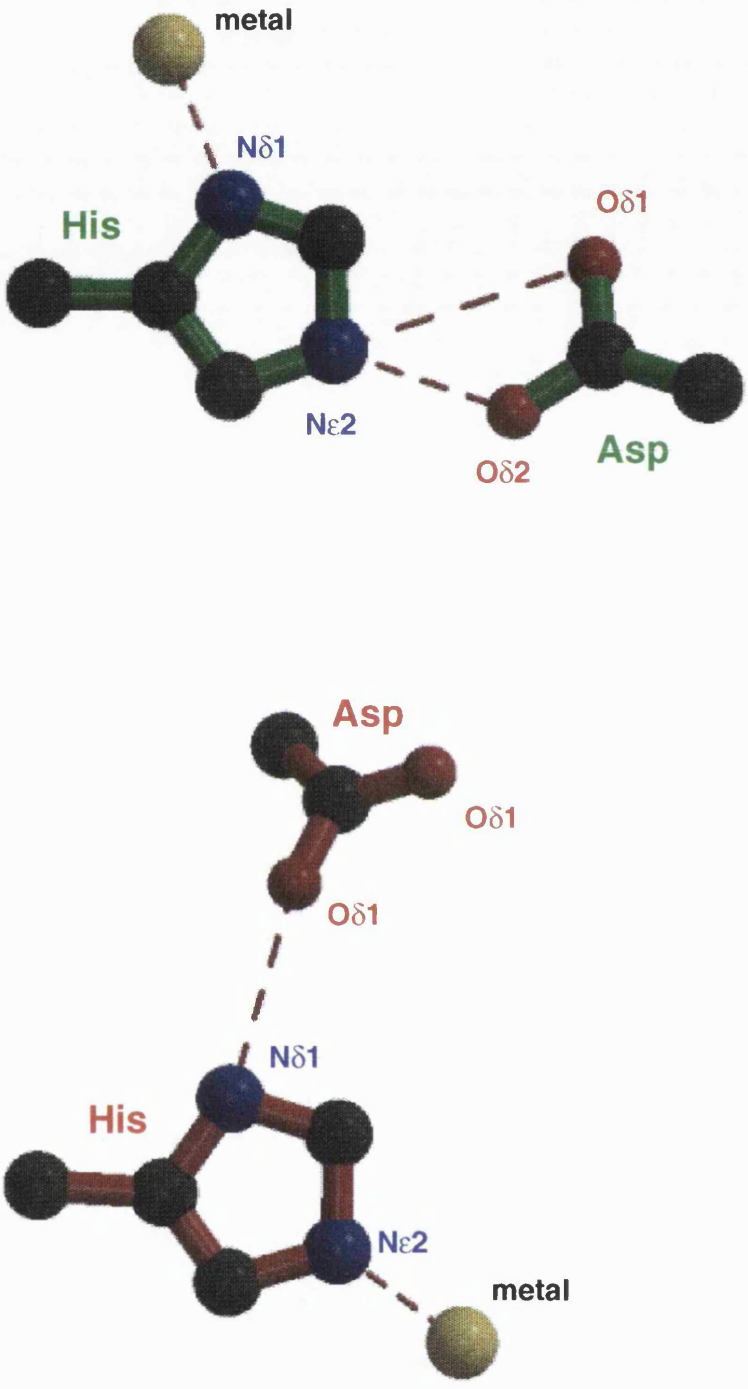


Figure 7.1: The two possible tautomeric forms of the metal–His–ELEC triad. The triad with green bonds is taken from the catalytic centre of thermolysin, 1*tmn* (Monzingo & Matthews, 1984); here the His N^{ϵ_2} interacts with the Zn metal (tautomer ϵ). The other triad in red bonds originates from Cu, Zn–superoxide dismutase 2*sod* (Tainer *et al.*, 1982) and here the Zn metal interacts with the N^{δ_1} (tautomer δ).

proper conformation to interact with the metal (Argos *et al.*, 1978).

In chapter 5 we saw that the His–ELEC diad acts as an acid/base catalyst as part of the catalytic Nu:–His–ELEC catalytic triad. Similarly, the metal–histidine–carboxylate triads discussed above can be thought of as metal–His–ELEC. Christianson & Alexander (1990) first noted the similarity in the His–Asp diad of the serine proteinases and the diad that ligates Zn in carboxypeptidase, thermolysin and carbonic anhydrase.

This chapter can be divided into two main sections. Firstly, all metals that interact with at least one His residue are extracted and classified. We find that all these interactions originate from functional regions of proteins.

Secondly, the types of electrostatic groups, ELEC, hydrogen bonded to the His residue are studied in terms of their structure and function. There are a number of factors that determine the geometry of these triads but we find that they adopt an 'ideal' conformation in catalytic metal binding sites. There are, however, examples where this is not the case: in the Fe:S catalytic centre of aconitase (E.C.4.2.1.1) a metal–His–ELEC triad has been located with non–ideal geometry.

Using the coordinates of the metal–His–ELEC consensus template, it finds functional metal binding sites automatically; however, it is not possible to separate non–catalytic functional from catalytic triads.

7.2 Metal binding sites in the PDB

Figure 7.2 is a flow diagram summarising how the metal–His–ELEC triads were extracted from the PDB. The dataset used was all proteins in the PDB with metal binding sites; these were extracted simply by searching for a metal in every PDB structure. Tables 7.3 shows that we located 79 unique enzymes by E.C. number

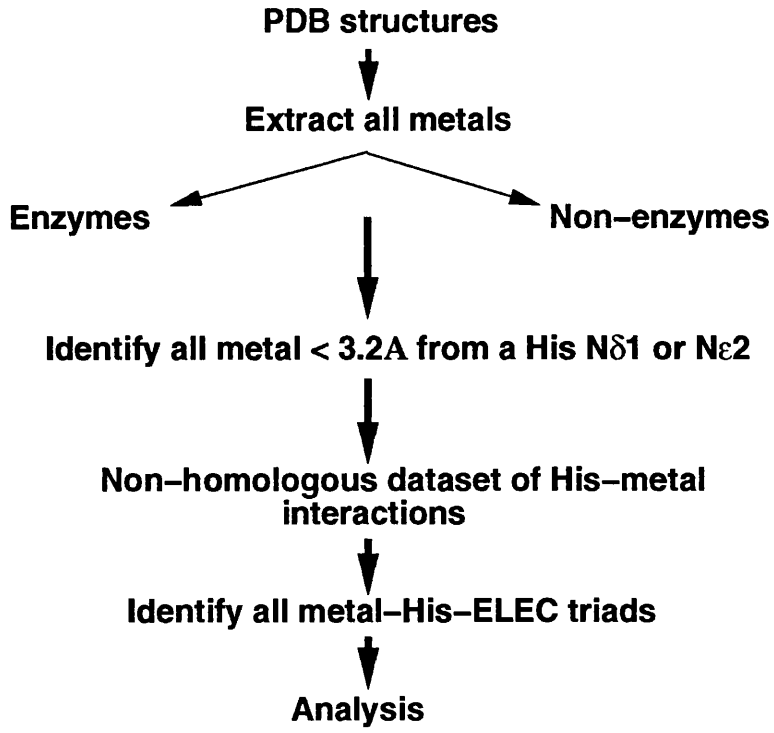


Figure 7.2: The method used to extract the metal–His–ELEC triads from the PDB.

(Bielka *et al.*, 1992) though some of these enzymes are homologues *e.g* tryspin and chymotrypsin.

There also were 40 unique non–enzyme proteins found (Table 7.4); these were classified according to their name in the PDB, however, due to the inherent inaccuracies in these files (Hooft *et al.*, 1996) some proteins of the same function may have a different name.

For each of the unique enzyme or proteins, the number of PDB structures and associated metals are listed. In addition, we calculated the number of these metals that are coordinated to a His; that is, if either of the His N^{δ_1} or N^{ϵ_2} atoms are within 3.3Å of the metal. We chose this cut–off based on the analysis of Einsphar & Bugg (1984); they found the metal–ligand bond length depended on more than one factor and could be anywhere between 1.2 and 3.3Å.

For example, Table 7.3 indicates that there are 13 alcohol dehydrogenase

E.C.1.1.1.1 structures in the PDB; in these structures there are a total of 2 coppers, both of which are liganded to a His, as well as 40 Zn metals, 19 of which have a His liganded to them.

The next procedure was to create a non-homologous dataset; this gives us only one unique example of each enzyme or non–enzyme protein with a metal–His interaction. In total, 39 of the 79 unique enzymes and 23 of the 40 unique proteins have at least one metal ligated to a His; Tables 7.5 and 7.6 summarise these matches. In fact, we extracted all metal–ligand interactions at a given metal–His binding site as this helps us to classify them in terms of function. For the non–enzyme proteins, a problem arose because these proteins were extracted from the PDB according to their name. After checking these maually it was found that 'electron transport' and 'oxygen transport proteins' could represent more than one different protein. This factor was taken into account when the His–metal interactions were classified; there are two representatives of 'oxygen transport proteins', the Fe hemoglobins/myoglobins as well as Cu containing hemocyanin. 'Electron transport proteins' was found to represent either the Fe containing cytochrome B and C or Cu containing amicyanin.

The 40 enzymes without a metal–His interaction are mostly non–catalytic structural metal sites. However there are catalytic sites such as the Ca^{2+} metal site in phospholipase A_2. In addition, some have arisen as a result of protein engineering experiments; for example, in the lysozyme structure *3lhm* (Inaka *et al.*, 1991) a Ca binding site has been introduced by mutation experiments.

There are a few unusual enzyme His–metal interactions. Tissue kallikrein (tonin) E.C.3.4.21.35 (Fujinaga *et al.*, 1987) is a serine–proteinase with a Ser 195–His 57–Asp 102 catalytic triad typical of this group of enzymes. In fact, Table 7.5 indicates that a Zn is ligated to the His 57 residue; it is thought that Zn inhibits this enzyme under physiological conditions and this is most prob-

name	no. PDB structures	AU	CA	CD	CO	CU	FE	HG	LI	MG	MN	NA	NI	ZN
Oxidoreductases E.C.1														
ALCOHOL DEHYDROGENASE E.C.1.1.1.1	13					2 **2**								19 **40**
L-IDITOL 2-DEHYDROGENASE E.C.1.1.1.14	1													1 **1**
ISOCITRATE DEHYDROGENASE (NADP+) E.C.1.1.1.42	4		0 **2**							0 **2**				
L-LACTATE DEHYDROGENASE (CYTOCHROME) E.C.1.1.2.3	2						2 **2**							
GALACTOSE OXIDASE E.C.1.1.3.9	3					2 **2**						0 **3**		
ALCOHOL DEHYDROGENASE (ACCEPTOR) E.C.1.1.99.8	1		0 **2**											
PYRUVATE OXIDASE E.C.1.2.3.3	2									0 **4**		0 **2**		
DEHYDROGENASE E.C.1.4.99.3	2					3 **3**	1 **1**							
DIHYDROFOLATE REDUCTASE E.C.1.5.1.3	16		0 **16**											
TRIMETHYLAMINE DEHYDROGENASE E.C.1.5.99.7	1						0 **8**							
NITRITE REDUCTASE E.C.1.7.99.3	1					6 **6**								
L-ASCORBATE OXIDASE E.C.1.10.3.3	4					36 **36**								
CYTOCHROME-C PEROXIDASE E.C.1.11.1.5	23						26 **28**							
CATALASE E.C.1.11.1.6	2						0 **3**							
PEROXIDASE E.C.1.11.1.7	2		0 **4**				3 **3**							
PROTOCATECHUATE 3,4-DIOXYGENASE E.C.1.13.11.3	1						6 **6**							
UNSPECIFIC MONOOXYGENASE E.C.1.14.14.1	2						0 **4**							
CAMPHOR 5-MONOOXYGENASE E.C.1.14.15.1	18						0 **18**							
SUPEROXIDE DISMUTASE E.C.1.15.1.1	15				2 **2**	31 **31**	10 **10**				6 **6**			29 **29**
RIBONUCLEOSIDE-DIPHOSPHATE REDUCTASE E.C.1.17.4.1	3						8 **8**	0 **27**			4 **4**			
NITROGENASE E.C.1.18.6.1	1						0 **4**			0 **2**				
Transferases E.C.2														
ASPARTATE CARBAMOYLTRANSFERASE E.C.2.1.3.2	20													0 **39**
TRANSKETOLASE E.C.2.2.1.1	4		0 **8**											
CHLORAMPHENICOL O-ACETYLTRANSFERASE E.C.2.3.1.28	5				5 **10**									
CYCLOMALTODEXTRIN GLUCANOTRANSFERASE E.C.2.4.1.19	2		2 **4**											
LACTOSE SYNTHASE E.C.2.4.1.22	1		0 **1**											
AMIDOPHOSPHORIBOSYLTRANSFERASE E.C.2.4.2.14	1						0 **16**							
6-PHOSPHOFRUCTOKINASE E.C.2.7.1.11	2									0 **6**				
PROTEIN KINASE E.C.2.7.1.37	1										0 **2**			
PROTEIN-N(PI)-PHOSPHOHISTIDINE-SUGAR PHO E.C.2.7.1.69	2									0 **1**	0 **1**			1 **1**
PROTEIN-TYROSINE KINASE E.C.2.7.1.112	2						0 **9**							2 **2**
PHOSPHOGLYCERATE KINASE E.C.2.7.2.3	2									0 **2**				
NUCLEOSIDE-DIPHOSPHATE KINASE E.C.2.7.4.6	3									0 **4**				
DNA-DIRECTED DNA POLYMERASE E.C.2.7.7.7	5									0 **3**	0 **1**			0 **2**
Hydrolases E.C.3														
TRIACYLGLYCEROL LIPASE E.C.3.1.1.3	4		0 **8**											
PHOSPHOLIPASE A2 E.C.3.1.1.4	10		0 **18**											
ALKALINE PHOSPHATASE E.C.3.1.3.1	1									0 **2**				4 **4**
FRUCTOSE-BISPHOSPHATASE E.C.3.1.3.11	7									0 **6**	0 **6**			0 **6**
DEOXYRIBONUCLEASE I E.C.3.1.21.1	2		0 **6**											

name	no. PDB structures	AU	CA	CD	CO	CU	FE	HG	LI	MG	MN	NA	NI	ZN
RIBONUCLEASE H E.C.3.1.26.4	1									0 1				
RIBONUCLEASE T1 E.C.3.1.27.3	16		0 18									0 1		0 1
PANCREATIC RIBONUCLEASE E.C.3.1.27.5	1				0 4									
MICROCOCCAL NUCLEASE E.C.3.1.31.1	11		0 11											
ALPHA-AMYLASE E.C.3.2.1.1	5		3 9											
OLIGO-1,6-GLUCOSIDASE E.C.3.2.1.10	6					24 24								
LYSOZYME E.C.3.2.1.17	2		0 1									0 1		
EXO-ALPHA-SIALIDASE E.C.3.2.1.18	16		0 22											
LICHENINASE E.C.3.2.1.73	4		0 4											
CELLULOSE 1,4-BETA-CELLOBIOSIDASE E.C.3.2.1.91	1		0 1											
LEUCYL AMINOPEPTIDASE E.C.3.4.11.1	4									0 1				0 7
BACTERIAL LEUCYL AMINOPEPTIDASE E.C.3.4.11.10	1													2 2
METHIONYL AMINOPEPTIDASE E.C.3.4.11.18	1				1 2									
DIPEPTIDYL-PEPTIDASE I E.C.3.4.14.1	1						0 7							
CARBOXYPEPTIDASE A E.C.3.4.17.1	11			0 10										11 11
TRYPSIN E.C.3.4.21.4	36		0 37											
THROMBIN E.C.3.4.21.5	1		0 7											
TISSUE KALLIKREIN E.C.3.4.21.35	1													1 1
PANCREATIC ELASTASE E.C.3.4.21.36	14		0 13											
SUBTILISIN E.C.3.4.21.62	21		3 41									0 2		
ENDOPEPTIDASE K E.C.3.4.21.64	4		0 4					1 1				0 1		
THERMITASE E.C.3.4.21.66	4		0 9									0 3		
PROTEIN C (ACTIVATED) E.C.3.4.21.69	4		0 2											
STREPTOGRISIN B E.C.3.4.21.81	1		0 1											
CARICAIN E.C.3.4.22.30	1							1 1						
RHIZOPUSPEPSIN E.C.3.4.23.21	2		0 2											
MEPRIN A E.C.3.4.24.18	1													1 1
ASTACIN E.C.3.4.24.21	5				1 1	1 1		1 1					1 1	1 1
PSEUDOLYSIN E.C.3.4.24.26	1		0 1											1 1
THERMOLYSIN E.C.3.4.24.27	14		0 56											13 13
ADAMALYSIN E.C.3.4.24.46	1		0 1											1 1
N-ACETYLMURAMOYL-L-ALANINE AMIDASE E.C.3.5.1.28	1													1 1
ADENOSINE DEAMINASE E.C.3.5.4.4	1													1 1
ADENOSINETRIPHOSPHATASE E.C.3.6.1.3	7		0 1							0 6				
Lyases E.C.4														
RIBULOSE-BISPHOSPHATE CARBOXYLASE E.C.4.1.1.39	5									2 10				
2,2-DIALKYLGLYCINE DECARBOXYLASE (PYRUVA E.C.4.1.1.64	3								0 1			0 4		
CARBONATE DEHYDRATASE E.C.4.2.1.1	62	0 4			3 3	1 2		3 30			1 1	0 1	1 1	54 56
ACONITATE HYDRATASE E.C.4.2.1.3	7						0 27							
PHOSPHOPYRUVATE HYDRATASE E.C.4.2.1.11	6		0 1							0 2	0 2			0 2
DNA-(APURINIC OR APYRIMIDINIC SITE) LYAS E.C.4.2.99.18	1						0 4							

name	no. PDB structures	AU	CA	CD	CO	CU	FE	HG	LI	MG	MN	NA	NI	ZN
Isomerases E.C.5														
MANDELATE RACEMASE E.C.5.1.2.2	3									0 **2**	0 **1**			
XYLOSE ISOMERASE E.C.5.3.1.5	**46**				16 **32**					24 **70**	25 **55**			2 **4**
MUCONATE CYCLOISOMERASE E.C.5.5.1.1	1										0 **1**			
CHLOROMUCONATE CYCLOISOMERASE E.C.5.5.1.7	1										0 **2**			
Ligases E.C.6														
METHIONINE-TRNA LIGASE E.C.6.1.1.10	2													0 **2**
GLUTAMATE-AMMONIA LIGASE E.C.6.3.1.2	2										2 **4**			

Table 7.3: List of all the enzymes in the January 1995 release of the PDB which have bound metals. For each enzyme, the number of PDB structures are given. For each metal type there are two numbers in each box, the bottom number is the total number of metals in the PDB structures for that enzyme. The top number is the number of these metals liganded to a His.

ably its mechanism of inhibition. The serine–proteinase endopeptidase K and the cysteine–proteinase caricain have the heavy metal Hg ligated to their active sites. These are the result of heavy–metal derivative experiments during the detemination of the X–ray structure of these proteins. Methylamine dehydrogenase 1*mda* (Chen *et al*, 1992), E.C.1.4.99.3 has a Cu, however, this enzyme is complexed with amicyanin which is a Cu–containg electron transport protein. Two enzymes, α–amylase (Matsura *et al.*, 1984) and cyclomaltodextrin glucanotransferase E.C.2.4.1.19 (Klein & Schulz, 1991) have Ca bound to the surface of their structures; both these metals are thought to stabilise their bound protein's structure. These are unusual because Ca is a hard metal and prefers to ligate to oxygen atoms; in fact, the Ca–His distances in these enzymes are 2.34Å and 2.37Å respectively which are quite long for these types of ligand–metal bonds.

name	no. PDB structures	AU	CA	CD	CO	CU	FE	HG	LI	MG	MN	NA	NI	ZN
ADHESION PROTEIN	1		0 **3**											
AIDS-RELATED VIRUS GAG POLYPROTEIN	1													1 **1**
AMYLOID PROTEIN	1		0 **10**											
APOPROTEIN	1		0 **1**											
BINDING PROTEIN	8		0 **14**	1 **7**						0 **3**		0 **1**		
CALCIUM/PHOSPHOLIPID BINDING	36		2 **140**	0 **2**						0 **3**	0 **3**			
CARBOXYLIC ESTER HYDROLASE	2		0 **4**											
CELLULAR LIPOPHILIC TRANSPORT PROTEIN	1			0 **2**										
CHAPERONE PROTEIN	2									0 **2**				
CONTRACTILE SYSTEM PROTEINS	4		0 **10**											
DNA-BINDING REGULATORY PROTEIN	3									2 **2**				2 **4**
ELECTRON TRANSPORT	105		0 **1**	4 **4**		24 **28**	67 **237**	1 **1**						
ELONGATION FACTOR	2									0 **2**				
EXCITATION ENERGY TRANSFER	1									5 **7**				
FINGER DNA BINDING DOMAIN	3													3 **3**
GALACTOSE-BINDING PROTEIN	3		0 **3**											
GENE-REGULATING PROTEIN	2											0 **2**		
GLUCOCORTICOID RECEPTOR	3													0 **8**
HORMONE	13											0 **4**		19 **19**
HYDROLASE INHIBITOR(SERINE PROTEINASE)	5		2 **12**											
IMMUNOGLOBULIN	2						0 **1**							0 **1**
INTEGRAL MEMBRANE PROTEIN PORIN	1		0 **3**											
LECTIN(AGGLUTININ)	25		0 **53**	1 **2**							43 **44**		1 **1**	
METALLOTHIONEIN	7			0 **26**								0 **1**		0 **2**
MUSCLE PROTEIN	2		0 **4**											
NUCLEOCAPSID PROTEIN	4													6 **6**
ONCOGENE PROTEIN	18									0 **21**				
OXYGEN TRANSPORT	138				2 **2**	16 **16**	243 **245**					0 **1**	2 **2**	
PANCREATIC HORMONE	14													0 **1**
PERIPLASMIC BINDING PROTEIN	2		0 **2**											
PHEROMONE-BINDING	1			1 **4**										
PHOTOSYNTHETIC REACTION CENTER	5						9 **9**			18 **19**				
PLATELET FACTOR	1												0 **1**	
PROTEIN (VIRAL)	4		0 **13**				4 **4**							1 **2**
SIGNAL TRANSDUCTION PROTEIN	3									0 **3**				
STORAGE	3		0 **2**	0 **2**			1 **2**				4 **4**			
TOXIN	1													0 **3**
TRANSCRIPTION REGULATION	8			0 **2**										4 **7**
TRANSFERRIN	1						2 **2**							
TRANSPORT	7					4 **4**	3 **4**			0 **1**				

Table 7.4: List of all the non–enzyme proteins in the January 1995 release of the PDB which have bound metals. For each enzyme, the number of PDB structures are given. For each metal type there are two numbers in each box, the bottom number is the total number of metals in the PDB structures for that unique protein. The top number is the number of these metals liganded to a His.

PDB code	metal	interacting residues
Oxidoreductases E.C.1		
		ALCOHOL DEHYDROGENASE E.C.1.1.1.1
2oxi	ZN ZN A 1	$N^{\epsilon 2}$ HIS A 67 S^{γ} CYS A 174
		L-IDITOL 2-DEHYDROGENASE E.C.1.1.1.14
1sdg	ZN ZN 375	S^{γ} CYS 46 $N^{\epsilon 2}$ HIS 67 $O^{\epsilon 2}$ GLU 174 C1 SOR 376
		L-LACTATE DEHYDROGENASE (CYTOCHROME) E.C.1.1.2.3
1fcb	FE HEM A 560	$N^{\epsilon 2}$ HIS A 43 $N^{\epsilon 2}$ HIS A 66
		GALACTOSE OXIDASE E.C.1.1.3.9
1gof	CU CU 700	OH TYR 272 OH TYR 495 $N^{\epsilon 2}$ HIS 496 $N^{\epsilon 2}$ HIS 581 C ACY 703
		DEHYDROGENASE E.C.1.4.99.3
1mda	CU CU A 0	$N^{\delta 1}$ HIS A 53 S^{γ} CYS A 92 $N^{\delta 1}$ HIS A 95 S^{δ} MET A 98
		NITRITE REDUCTASE E.C.1.7.99.3
1afn	CU CU A 501	$N^{\delta 1}$ HIS A 95 S^{γ} CYS A 136 $N^{\delta 1}$ HIS A 145 S^{δ} MET A 150
1afn	CU CU A 502	$N^{\epsilon 2}$ HIS A 100 $N^{\epsilon 2}$ HIS A 135 O HOH A 503 $N^{\epsilon 2}$ HIS B 306
		L-ASCORBATE OXIDASE E.C.1.10.3.3
1aoz	CU CU A 701	$N^{\delta 1}$ HIS A 445 S^{γ} CYS A 507 $N^{\delta 1}$ HIS A 512 S^{δ} MET A 517
1aoz	CU2 C2O A 702	$N^{\epsilon 2}$ HIS A 106 $N^{\epsilon 2}$ HIS A 450 $N^{\epsilon 2}$ HIS A 506
1aoz	CU3 C2O A 702	$N^{\delta 1}$ HIS A 62 $N^{\epsilon 2}$ HIS A 104 $N^{\epsilon 2}$ HIS A 508
1aoz	CU4 C1O A 703	$N^{\epsilon 2}$ HIS A 60 $N^{\epsilon 2}$ HIS A 448
1aoz	CU CU 812	$N^{\epsilon 2}$ HIS A 286 $N^{\epsilon 2}$ HIS B 286 O HOH 605
		CYTOCHROME-C PEROXIDASE E.C.1.11.1.5
1cca	FE HEM 1	$N^{\epsilon 2}$ HIS 175 O HOH 313
		PEROXIDASE E.C.1.11.1.7
1arp	FE HEM 345	$N^{\epsilon 2}$ HIS 184 O HOH 415
		PROTOCATECHUATE 3,4-DIOXYGENASE E.C.1.13.11.3
2pcd	FE FE M 600	OH TYR M 408 OH TYR M 447 $N^{\epsilon 2}$ HIS M 460 $N^{\epsilon 2}$ HIS M 462 O HOH M 801 O HOH M 827
		Cu,Zn-SUPEROXIDE DISMUTASE E.C.1.15.1.1
1sos	CU CU A 152	$N^{\delta 1}$ HIS O 44 $N^{\epsilon 2}$ HIS O 46 HIS O 61 $N^{\epsilon 2}$ HIS O 118 O HOH 191
1sos	ZN ZN 262	$N^{\epsilon 2}$ HIS 94 $N^{\epsilon 2}$ HIS 96 $N^{\delta 1}$ HIS 119 O HOH 263
		FE-SUPEROXIDE DISMUTASE E.C.1.15.1.1
1ids	FE FE A 200	$N^{\epsilon 2}$ HIS A 28 $N^{\epsilon 2}$ HIS A 76 $O^{\delta 2}$ ASP A 160 $N^{\epsilon 2}$ HIS A 164 O HOH 1001
		MN-SUPEROXIDE DISMUTASE E.C.1.15.1.1
1abm	MN A 199	$N^{\epsilon 2}$ HIS A 26 $N^{\epsilon 2}$ HIS A 74 $O^{\delta 2}$ ASP A 159 $N^{\epsilon 2}$ HIS A 163 O HOH A 200
		RIBONUCLEOSIDE-DIPHOSPHATE REDUCTASE E.C.1.17.4.1
1rub	FE1 FE0 401	$O\delta_1$ ASP A 84 $O^{\delta 2}$ ASP A 84 $O^{\epsilon 1}$ GLU A 115 $N^{\delta 1}$ HIS A 118 O HOH 522 O HOH 749
1rib	FE2 FE0 401	$O^{\epsilon 1}$ GLU A 115 $O^{\epsilon 2}$ GLU A 115 $O^{\epsilon 2}$ GLU A 204 $O^{\epsilon 2}$ GLU A 238 $N^{\delta 1}$ HIS A 241 O HOH 522
Lyases E.C.2		
		CHLORAMPHENICOL O-ACETYLTRANSFERASE E.C.2.3.1.28
1cia	CO CO 222	$O^{\epsilon 1}$ GLU 23 $O^{\epsilon 2}$ GLU 23 $N^{\delta 1}$ HIS 27
		CYCLOMALTODEXTRIN GLUCANOTRANSFERASE E.C.2.4.1.19
1cgt	CA CA 685	$O\delta_1$ ASN 139 O ILE 190 $O\delta_1$ ASP 199 $O^{\delta 2}$ ASP 199 O HIS 233 O HOH 738
		PROTEIN-N(PI)-PHOSPHOHISTIDINE-SUGAR PHOSPHATASE E.C.2.7.1.69
1glc	ZN ZN 4	$N^{\epsilon 2}$ HIS F 75 $N^{\epsilon 2}$ HIS F 90 $O^{\epsilon 1}$ GLU G 478 $O^{\epsilon 2}$ GLU G 478 O HOH 1
		PROTEIN-TYROSINE KINASE E.C.2.7.1.112
6ins	ZN ZN 1	$N^{\epsilon 2}$ HIS F 10D O HOH 119
6ins	ZN ZN 2	$N^{\epsilon 2}$ HIS E 10B O HOH 1
Hydrolases E.C.3		
		ALKALINE PHOSPHATASE E.C.3.1.3.1
1alk	ZN ZN A 450	$O\delta_1$ ASP A 327 $O^{\delta 2}$ ASP A 327 $N^{\epsilon 2}$ HIS A 331 $N^{\epsilon 2}$ HIS A 412 P PO4 A 453
1alk	ZN ZN A 451	$O\delta_1$ ASP A 51 $O^{\delta 2}$ ASP A 51 OG SER A 102 $O\delta_1$ ASP A 369 $O^{\delta 2}$ ASP A 369 $N^{\epsilon 2}$ HIS A 370
		ALPHA-AMYLASE E.C.3.2.1.1
1cdg	CA CA 692	$O\delta_1$ ASN 139 O ILE 190 $O\delta_1$ ASP 199 $O^{\delta 2}$ ASP 199 O HIS 233 O HOH 24
		OLIGO-1,6-GLUCOSIDASE E.C.3.2.1.10
1azn	CU CU A 200	O GLY A 45 $N^{\delta 1}$ HIS A 46 S^{γ} CYS A 112 $N^{\delta 1}$ HIS A 117 S^{δ} MET A 121
		BACTERIAL LEUCYL AMINOPEPTIDASE E.C.3.4.11.10
1amp	ZN ZN 501	$O\delta_1$ ASP 117 $O^{\delta 2}$ ASP 117 $O^{\epsilon 1}$ GLU 152 $O^{\epsilon 2}$ GLU 152 $N^{\epsilon 2}$ HIS 256 O HOH 935
1amp	ZN ZN 502	$N^{\epsilon 2}$ HIS 97 $O\delta_1$ ASP 117 $O\delta_1$ ASP 179 $O^{\delta 2}$ ASP 179 O HOH 934 O HOH 935
		METHIONYL AMINOPEPTIDASE E.C.3.4.11.18
1mat	CO CO 401	$O^{\delta 2}$ ASP 108 $N^{\epsilon 2}$ HIS 171 $O^{\epsilon 1}$ GLU 204 $O^{\epsilon 2}$ GLU 204 $O^{\epsilon 2}$ GLU 235
		CARBOXYPEPTIDASE A E.C.3.4.17.1
1cbx	ZN ZN 309	$N^{\delta 1}$ HIS 69 $O^{\epsilon 1}$ GLU 72 $O^{\epsilon 2}$ GLU 72 $N^{\delta 1}$ HIS 196 C1 BZS 500
		TISSUE KALLIKREIN E.C.3.4.21.35
1ton	ZN ZN 200	$N^{\epsilon 2}$ HIS 57 $N^{\epsilon 2}$ HIS 97 $N^{\epsilon 2}$ HIS 99
		SUBTILISIN E.C.3.4.21.62
1sca	CA CA 403	O ALA 37 O HIS 39 O LEU 42
		ENDOPEPTIDASE K E.C.3.4.21.64
1ptk	HG HG 73	$O\delta_1$ ASP 39 $O^{\delta 2}$ ASP 39 O HIS 69 $N^{\delta 1}$ HIS 69 S^{γ} CYS 73
		CARICAIN E.C.3.4.22.30
1ppo	HG HG 217	S^{γ} CYS 25 $N^{\delta 1}$ HIS 159 O HOH 229
		MEPRIN A E.C.3.4.24.18
1iaf	ZN ZN 999	$N^{\epsilon 2}$ HIS 92 $N^{\epsilon 2}$ HIS 96 $N^{\epsilon 2}$ HIS 102 OH TYR 149 O HOH 300
		ASTACIN E.C.3.4.24.21
1iab	CO CO 999	$N^{\epsilon 2}$ HIS 92 $N^{\epsilon 2}$ HIS 96 $N^{\epsilon 2}$ HIS 102 OH TYR 149 O HOH 300
		PSEUDOLYSIN E.C.3.4.24.26
1ezm	ZN ZN 300	$N^{\epsilon 2}$ HIS 140 $N^{\epsilon 2}$ HIS 144 $O^{\epsilon 1}$ GLU 164 $O^{\epsilon 2}$ GLU 164 O HOH 14
		THERMOLYSIN E.C.3.4.24.27
1npc	ZN ZN 323	$N^{\epsilon 2}$ HIS 143 $N^{\epsilon 2}$ HIS 147 $O^{\epsilon 1}$ GLU 167 $O^{\epsilon 2}$ GLU 167 O HOH 326
		ADAMALYSIN E.C.3.4.24.46
1iag	ZN ZN 999	$N^{\epsilon 2}$ HIS 142 $N^{\epsilon 2}$ HIS 146 $N^{\epsilon 2}$ HIS 152 O HOH 300
		N-ACETYLMURAMOYL-L-ALANINE AMIDASE E.C.3.5.1.28
1lba	ZN ZN 151	$N^{\delta 1}$ HIS 17 $N^{\delta 1}$ HIS 122 S^{γ} CYS 130 O HOH 199
		ADENOSINE DEAMINASE E.C.3.5.4.4
1add	ZN ZN 400	$N^{\epsilon 2}$ HIS 15 $N^{\epsilon 2}$ HIS 17 $N^{\epsilon 2}$ HIS 214 $O\delta_1$ ASP 295 O HOH 461

PDB code	metal	interacting residues
		Lyases E.C.4
		RIBULOSE-BISPHOSPHATE CARBOXYLASE E.C.4.1.1.39
4rub	MG MG A 491	$O\delta_1$ ASP A 203 $O^{\epsilon 1}$ GLU A 204 $N^{\epsilon 2}$ HIS A 294 CZ CBX A 201 C2 CAP A 490
		CARBONATE DEHYDRATASE E.C.4.2.1.1
1cah	CO CO 262	$N^{\epsilon 2}$ HIS 94 $N^{\epsilon 2}$ HIS 96 $N^{\delta 1}$ HIS 119 C BCT 500 O HOH 263
		Isomerases E.C.5
		XYLOSE ISOMERASE E.C.5.3.1.5
1xim	CO CO A 396	$O^{\epsilon 2}$ GLU A 217 $N^{\epsilon 2}$ HIS A 220 $O\delta_1$ ASP A 255 $O^{\delta 2}$ ASP A 255 $O\delta_1$ ASP A 257
		Ligases E.C.6
		GLUTAMATE-AMMONIA LIGASE E.C.6.3.1.2
1lgr	MN MN 470	$O^{\epsilon 1}$ GLU 129 $O^{\epsilon 2}$ GLU 129 $N^{\delta 1}$ HIS 269 $O^{\epsilon 1}$ GLU 357 $O^{\epsilon 2}$ GLU 357

Table 7.5: A list of all the enzyme structures in the January 1995 release of the PDB which have a metal ligated by one or more His residues.

7.3 The structure of metal–His interactions in the PDB

In this section, the structure of the metal–His interactions found in the non–homologous dataset (Tables 7.5 and 7.6) are analysed. In fact, it was mentioned in the introduction that a third element, the ELEC group, is often found hydrogen bonded to His, forming the metal–His–ELEC triad as shown in Figure 7.1.

We now investigate the types of metal–His–ELEC triads found in these metal binding sites. In order to extract all the metal–His–ELEC triads, the TESS program was run, firstly using the tautomer ϵ seed triad from tonin *1ton* (Asp 102 $O^{\delta 2}$, His 57 sidechain, Zn 200) (Fujinaga & James, 1987) and then the tautomer δ triad from superoxide dismutase *1sos* (Asp 122 $O^{\delta 1}$, Zn 155, His 69 sidechain) (Parge *et al.*, 1992), both with a distance cut–off of 3.0Å. In both cases we searched for any metal atom at the coordinate position of the seed metals and any non–carbon amino acid atom at the position of the seed templates' ELEC atom; the

PDB code	metal	interacting residues					
		AIDS-RELATED VIRUS GAG POLYPROTEIN					
2znf	ZN ZN 19	S^{γ} CYS 3	S^{γ} CYS 6	$N^{\epsilon 2}$ HIS 11	S^{γ} CYS 16		
		BINDING PROTEIN					
1hsl	CD CD 756	$N^{\delta 1}$ HIS A 129	O HOH 261	O HOH 263			
		CALCIUM-BINDING PROTEIN					
1clm	CA CA 152	$O\delta_1$ ASP 129	$O\delta_1$ ASP 131	$O\delta_1$ ASP 133	O HIS 135	$O^{\epsilon 1}$ GLU 140	$O^{\epsilon 2}$ GLU 140
		COMPLEX(TRANSCRIPTION REGULATION/DNA)					
2drp	ZN ZN A 171	S^{γ} CYS A 113	S^{γ} CYS A 116	$N^{\epsilon 2}$ HIS A 129	$N^{\epsilon 2}$ HIS A 134		
2drp	ZN ZN A 172	S^{γ} CYS A 143	S^{γ} CYS A 146	$N^{\epsilon 2}$ HIS A 159	$N^{\epsilon 2}$ HIS A 164		
		DNA-BINDING PROTEIN					
1cmc	MG MG A 106	$N^{\epsilon 2}$ HIS A 14	$O^{\epsilon 2}$ GLU A 19	O TYR A 104	OXT TYR A 104		
1bbo	ZN ZN 60	S^{γ} CYS 4	S^{γ} CYS 7	$N^{\epsilon 2}$ HIS 20	$N^{\epsilon 2}$ HIS 24		
1bbo	ZN ZN 61	S^{γ} CYS 32	S^{γ} CYS 35	$N^{\epsilon 2}$ HIS 48	$N^{\epsilon 2}$ HIS 54		
		ELECTRON TRANSPORT PROTEIN (AMICYANIN)					
1aaz	CD CD 188	ND1 HIS B 75	O HOH 1	O HOH 15			
1aaz	CD CD 189	NE2 HIS B 12	O HOH 16	O HOH 17	O HOH 18		
1aan	CU CU 200	$N^{\delta 1}$ HIS 53	S^{γ} CYS 92	$N^{\delta 1}$ HIS 95	S^{δ} MET 98		
		ELECTRON TRANSPORT PROTEIN (CYTOCHROME C)					
1bbh	FE HEM A 132	$N^{\epsilon 2}$ HIS A 125					
		EXCITATION ENERGY TRANSFER					
3bcl	MG BCL 1	NE2 HIS 105					
3bcl	MG BCL 3	NE2 HIS 290					
3bcl	MG BCL 4	NE2 HIS 282					
3bcl	MG BCL 6	NE2 HIS 140					
3bcl	MG BCL 7	ND1 HIS 289					
		FINGER DNA BINDING DOMAIN					
3znf	ZN ZN 31	S^{γ} CYS 5	S^{γ} CYS 8	$N^{\epsilon 2}$ HIS 21	$N^{\epsilon 2}$ HIS 27		
		HORMONE					
1izb	ZN ZN 101	$N^{\epsilon 2}$ HIS B 10	O HOH 64				
1izb	ZN ZN 102	$N^{\epsilon 2}$ HIS D 10					
		LECTIN					
1con	CD CD 1	$O^{\epsilon 2}$ GLU A 8	$O^{\delta 2}$ ASP A 10	$O\delta_1$ ASP A 19	$N^{\epsilon 2}$ HIS A 24	O HOH 2	O HOH 3
1scs	CO CO 1	$O^{\epsilon 2}$ GLU 8	$O^{\delta 2}$ ASP 10	$O\delta_1$ ASP 19	$N^{\epsilon 2}$ HIS 24	O HOH 11	O HOH 12
5cna	MN MN A 239	$O^{\epsilon 2}$ GLU A 8	$O^{\delta 2}$ ASP A 10	$O\delta_1$ ASP A 19	$N^{\epsilon 2}$ HIS A 24	O HOH 5	O HOH 6
1lec	MN MN 250	$O^{\epsilon 2}$ GLU 129	$O^{\delta 2}$ ASP 131	$O\delta_1$ ASP 140	$N^{\epsilon 2}$ HIS 145	O HOH 305	O HOH 332
1scr	NI NI 1	$O^{\epsilon 2}$ GLU 8	$O^{\delta 2}$ ASP 10	$O\delta_1$ ASP 19	$N^{\epsilon 2}$ HIS 24	O HOH 15	O HOH 16
		NUCLEOCAPSID PROTEIN					
1aaf	ZN ZN 56	S^{γ} CYS 15	S^{γ} CYS 18	$N^{\epsilon 2}$ HIS 23	S^{γ} CYS 28		
1aaf	ZN ZN 57	S^{γ} CYS 36	S^{γ} CYS 39	$N^{\epsilon 2}$ HIS 44	S^{γ} CYS 49		
		OXYGEN TRANSPORT (COPPER)					
1coh	CO COH B 1	1coh	NE2 HIS B 92				
1coh	CO COH D 1	1coh	NE2 HIS D 92				
1hc1	CU CU 665	$N^{\epsilon 2}$ HIS 194	$N^{\epsilon 2}$ HIS 198	$N^{\epsilon 2}$ HIS 224			
1hc1	CU CU 666	$N^{\epsilon 2}$ HIS 344	$N^{\epsilon 2}$ HIS 348	$N^{\epsilon 2}$ HIS 384			
		OXYGEN TRANSPORT (FE)					
1bab	FE HEM A 143	$N^{\epsilon 2}$ HIS A 88					
1nih	NI HNI A 1	1nih	NE2 HIS A 87				
1nih	NI HNI C 1	1nih	NE2 HIS C 87				
		PHEROMONE-BINDING					
1mup	CD CD 201	$N^{\epsilon 2}$ HIS 108	$O^{\epsilon 1}$ GLN 119	$N^{\epsilon 2}$GLN 119	$N^{\delta 1}$ HIS 145		
		PHOTOSYNTHETIC REACTION CENTER					
1prc	FE FE 607	NE2 HIS L 190	NE2 HIS L 230	NE2 HIS M 217	OE1 GLU M 232	OE2 GLU M 232	NE2 HIS M 264
1prc	FE HEM 609	SD MET C 74	NE2 HIS C 91				
1prc	FE HEM 610	SD MET C 110	NE2 HIS C 136				
1prc	FE HEM 611	SD MET C 233	NE2 HIS C 248				
1prc	FE HEM 612	NE2 HIS C 124	NE2 HIS C 309				
1prc	MG BCL 601	$N^{\epsilon 2}$ HIS M 180					
1prc	MG BCL 602	$N^{\epsilon 2}$ HIS L 173	CBB BCL 603				
1prc	MG BCL 603	$N^{\epsilon 2}$ HIS M 200					
1prc	MG BCL 604	$N^{\epsilon 2}$ HIS L 153					
		PROTEIN OF ELECTRON TRANSPORT					
2cdv	FE HEM 1	$N^{\epsilon 2}$ HIS 70	$N^{\epsilon 2}$ HIS 106				
2cdv	FE HEM 2	$N^{\epsilon 2}$ HIS 35	$N^{\epsilon 2}$ HIS 52				
2cdv	FE HEM 3	$N^{\epsilon 2}$ HIS 22	$N^{\epsilon 2}$ HIS 34				
2cdv	FE HEM 4	$N^{\epsilon 2}$ HIS 25	$N^{\epsilon 2}$ HIS 83				
		STORAGE AND ELECTRON TRANSPORT					
1fha	FE FE 200	$O^{\epsilon 1}$ GLU 27	$O^{\epsilon 1}$ GLU 62	$O^{\epsilon 2}$ GLU 62	$N^{\delta 1}$ HIS 65	O HOH 12	O HOH 13
1bcf	MN MN A 600	$O^{\epsilon 2}$ GLU A 51	$O^{\epsilon 1}$ GLU A 94	$O^{\epsilon 2}$ GLU A 94	$O^{\epsilon 1}$ GLU A 127	$N^{\delta 1}$ HIS A 130	
1bcf	MN MN A 601	$O^{\epsilon 1}$ GLU A 18	$O^{\epsilon 2}$ GLU A 18	$O^{\epsilon 1}$ GLU A 51	$N^{\delta 1}$ HIS A 54	$O^{\epsilon 2}$ GLU A 127	
		TRANSCRIPTION REGULATION					
1ard	ZN ZN 1	S^{γ} CYS 106	N CYS 109	S^{γ} CYS 109	$N^{\epsilon 2}$ HIS 122	$N^{\epsilon 2}$ HIS 126	
		TRANSFERRIN					
1ifg	FE FE 693	$O\delta_1$ ASP 60	OH TYR 92	OH TYR 192	$N^{\epsilon 2}$ HIS 253	C CO3 695	
1ifg	FE FE 694	$O\delta_1$ ASP 395	OH TYR 435	OH TYR 528	$N^{\epsilon 2}$ HIS 597	C CO3 696	
		VIRUS					
1lcf	ZN ZN 70	S^{γ} CYS 24	$N^{\delta 1}$ HIS 26	S^{γ} CYS 43	S^{γ} CYS 46		

Table 7.6: A non-homologous list of all the non-enzyme protein structures in the January 1995 release of the PDB which have a metal ligated by one or more His. In several instances there are more than one PDB code for each protein group. This occurs because the proteins were classified according to name and sometimes this refers to more than one protein by function.

Atom	x	y	z
metal–His N^{δ_1}	0.4	-3.1	0.2
ELEC	0.8	-3.9	0.3
metal–His N^{ϵ_2}	3.8	1.8	0.1
ELEC	4.1	2.4	0.2
His C^{β}	-1.5	-0.1	-0.0
His C^{γ}	0.0	0.0	0.0
His N^{δ_1}	0.8	-1.1	0.0
His C^{δ_2}	0.8	1.1	0.0
His C^{ϵ_1}	2.1	-0.7	-0.0
His N^{ϵ_2}	2.1	0.6	-0.0

Table 7.7: Coordinates of the two metal–His–ELEC conformations with respect to the His sidechain residue.

coordinates of the resultant consensus templates are given in Table 7.7. Table 7.8 shows the number and type of ELEC groups located for each of the metals in the dataset. In general, there is a predominance of carbonyl groups; this is not surprising considering that there is one per amino acid in the protein. Asp, Asn, Gln and Glu are also common ELEC groups; in the introduction we saw that these groups are thought to faciliate the activation of active site metals.

Figure 7.3 is a diagrammatic representation of the distribution of the metal and ELEC groups around the His for both tautomers δ and ϵ. The atoms are widely distributed indicating that there is stuctural distortion in the metal–His–ELEC triads. This distortion may be more common in a particular metal type. To check this we measured the distance of all metals for both triad tautomer δ and ϵ from the relevant sidechain N atom; Figure 7.4 shows the results. There is a clear peak around 2.2Å and it is comprised of more than one metal type, indicating that the structural heterogeneity of the metal–His–ELEC triad does not depend on the metal type.

Similarly, Figure 7.5 is a histogram of the number of hits against distance of

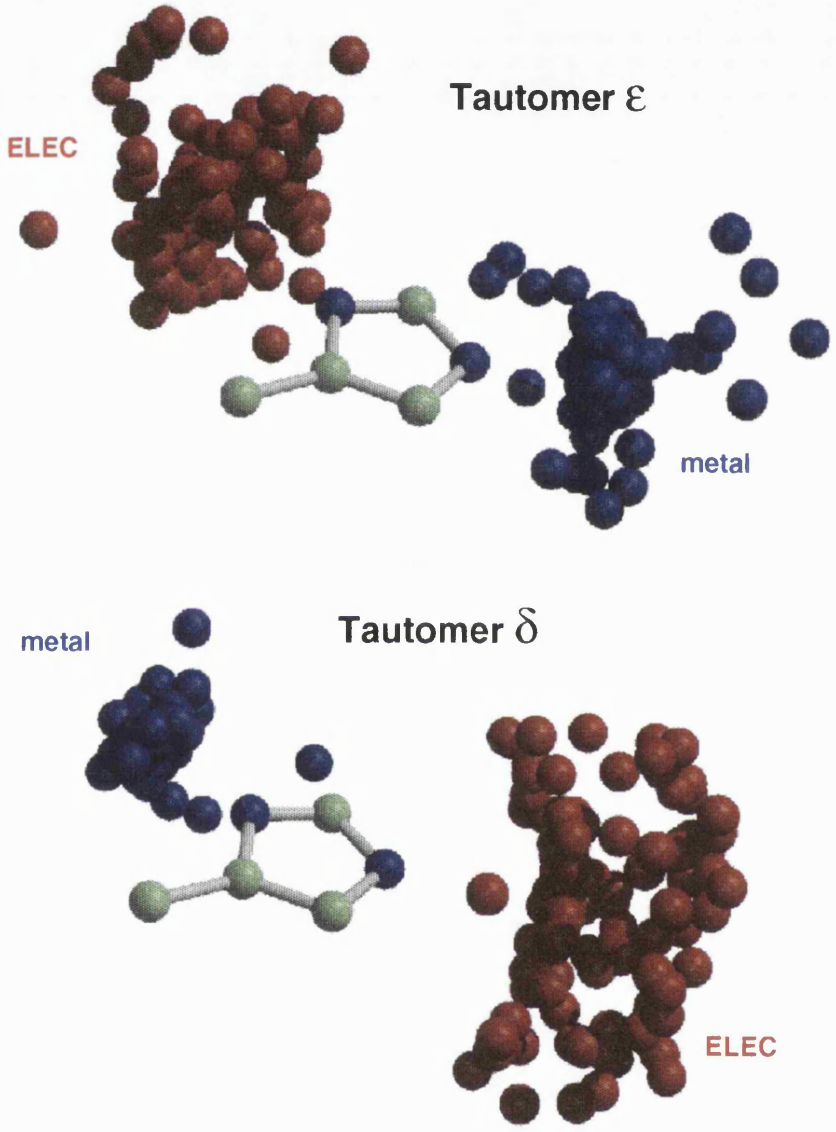


Figure 7.3: A diagram showing the distribution of the metal and ELEC atoms around the His sidechain for both tautomer δ and ϵ.

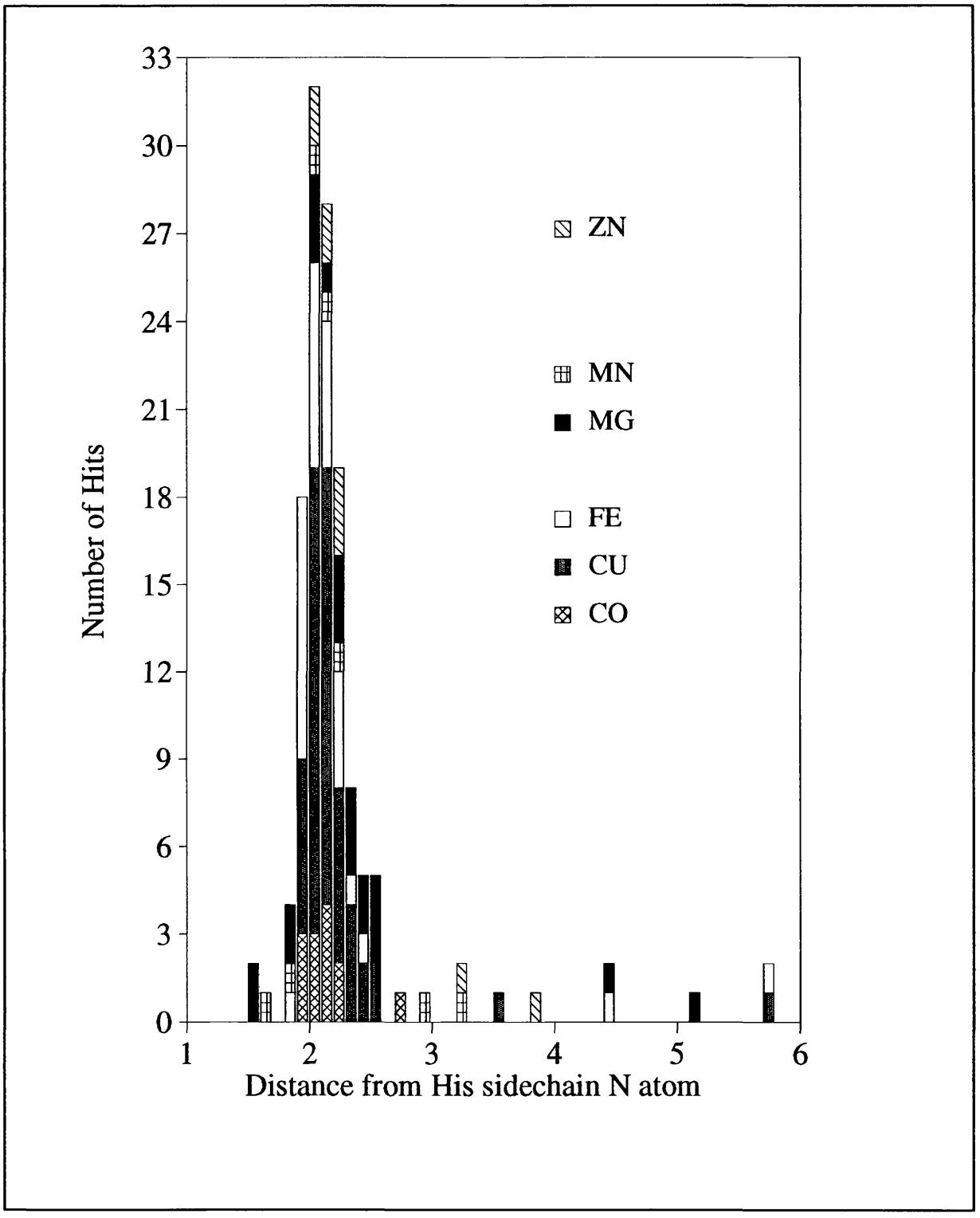


Figure 7.4: A histogram of the number of hits against distance of the metal from the sidechain N atom for both tautomers δ and ϵ.

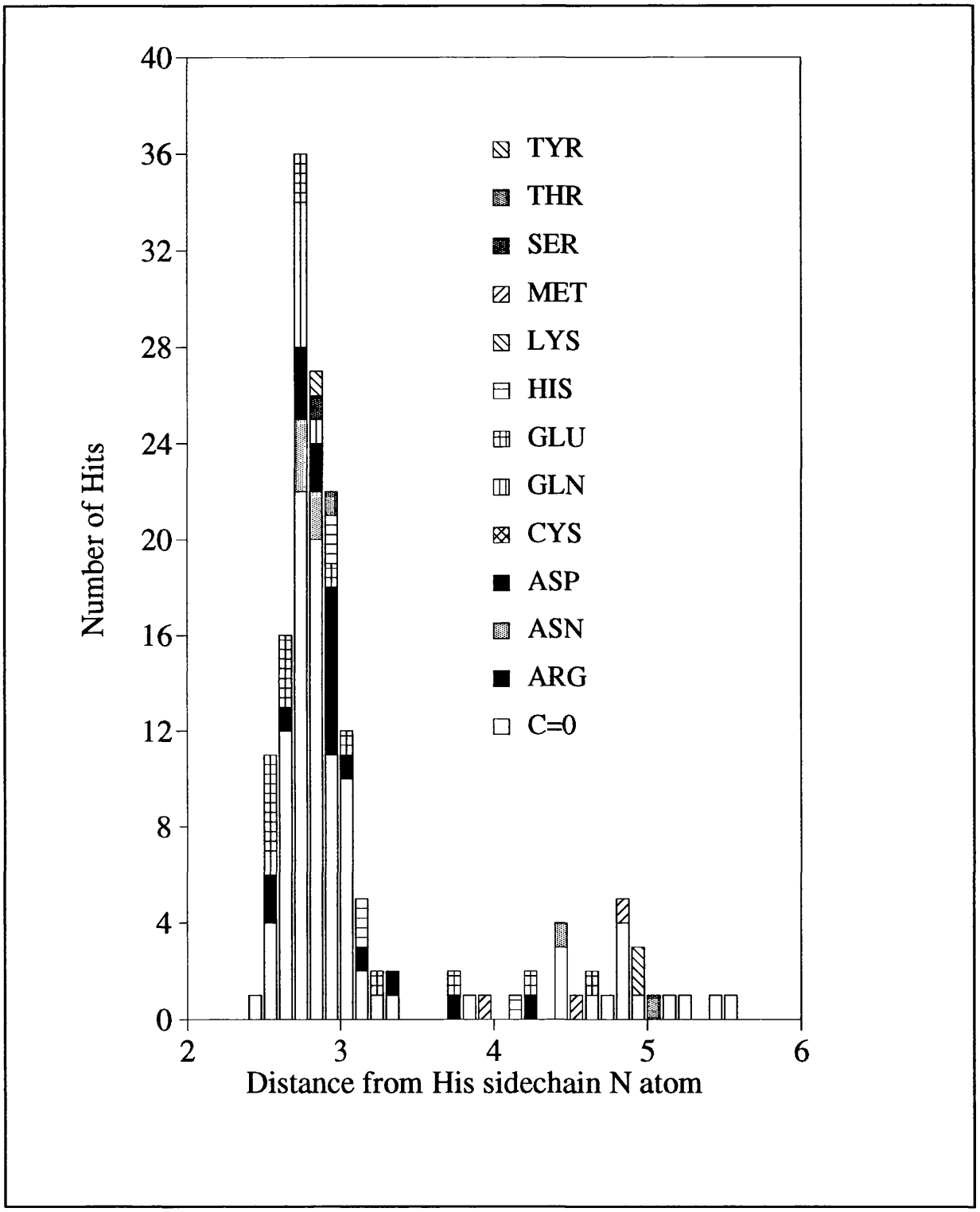


Figure 7.5: A histogram of the number of hits against distance of the metal from the sidechain N atom for both tautomers δ and ϵ.

Metal	ELEC group													
	ARG	ASN	ASP	CYS	GLN	GLU	HIS	LYS	MET	SER	THR	TYR	mainchain carbonyl	no ELEC group
CA														2
CD														3
CO	1		1		1	1	1						8	
CU		3	5		1	8	4		2	1	2		26	
FE			3		1	1		1					20	5
HG		1			1	1							4	
MG													6	
MN	1	3		1	1	1	1						5	
NI				1	1		1						6	
ZN		2	6		1	1						1	18	15

Table 7.8: The type of ELEC group associated with each metal as part of the metal–His–ELEC triad.

the ELEC group from the sidechain N atom for both tautomers δ and ϵ. Again, there is a peak of atoms around 1.8Å distance and all ELEC types occupy this position.

7.4 The structural heterogeneity of the metal–His–ELEC triad

The previous sections have shown that the structure of the metal–His–ELEC triad is not uniform. In this section the reasons for this heterogeneity will be discussed with reference to specific examples taken from both tautomers δ and ϵ of the metal–His–ELEC triad.

7.4.1 Zn–His interactions

Carbonic anhydrase E.C.4.2.1.1

- a metal center with optimal geometry

Most of the carbon dioxide produced during respiration requires transport out of the cell. Carbonic anhydrase hydrates CO_2 to carbonate, HCO_3^{2-}. It is one of the most efficient biological catalysts known and its rate reaches the limit of diffusion control. The crystal structure of this enzyme has been solved to 2.0Å by Eriksson *et al.* (1986). Important active site residues include Thr 199, Thr 20, Glu 106, His 64, Trp 209, Val 143, the Zn ion (liganded to His 94, His 96 and His 119) and the zinc bound hydroxide ion. The hydration of CO_2 occurs through chemically independent steps. The first step involves association of the substrate with enzyme and the chemical conversion of substrate into product. This involves Zn polarising its bound OH^- which then nucleophilically attacks CO_2, producing Zn bound HCO_3^-. Prior to the reaction, CO_2 is thought to lie close to Trp 209 and Val 143. Thr 109 and Glu 106 are involved in a hydrogen bonding network that facilitates proton transfer away from the active site; His 64 is thought to act as a proton buffer to the surface of the protein. The second step involves product dissociation and regeneration of the catalytically active nucleophile zinc hydroxide.

There are two metal–His–ELEC residues that have a sidechain ELEC group, a tautomer ϵ Zn 261–$N^{\epsilon 2}$ His 94 $N^{\delta 1}$–Gln 92 $O^{\epsilon 1}$ and tautomer δ Zn 261–$N^{\delta 1}$ His 119 $N^{\epsilon 2}$–Glu 117 $O^{\epsilon 2}$. A schematic view of the active site is shown in Figure 7.6. These two triads are different tautomeric forms but both have conformations near to the ideal and are 0.53Å and 0.57Å respectively from the mean consensus template. The third His 94 ligand has a mainchain ELEC group and this also adopts the ideal triad conformation.

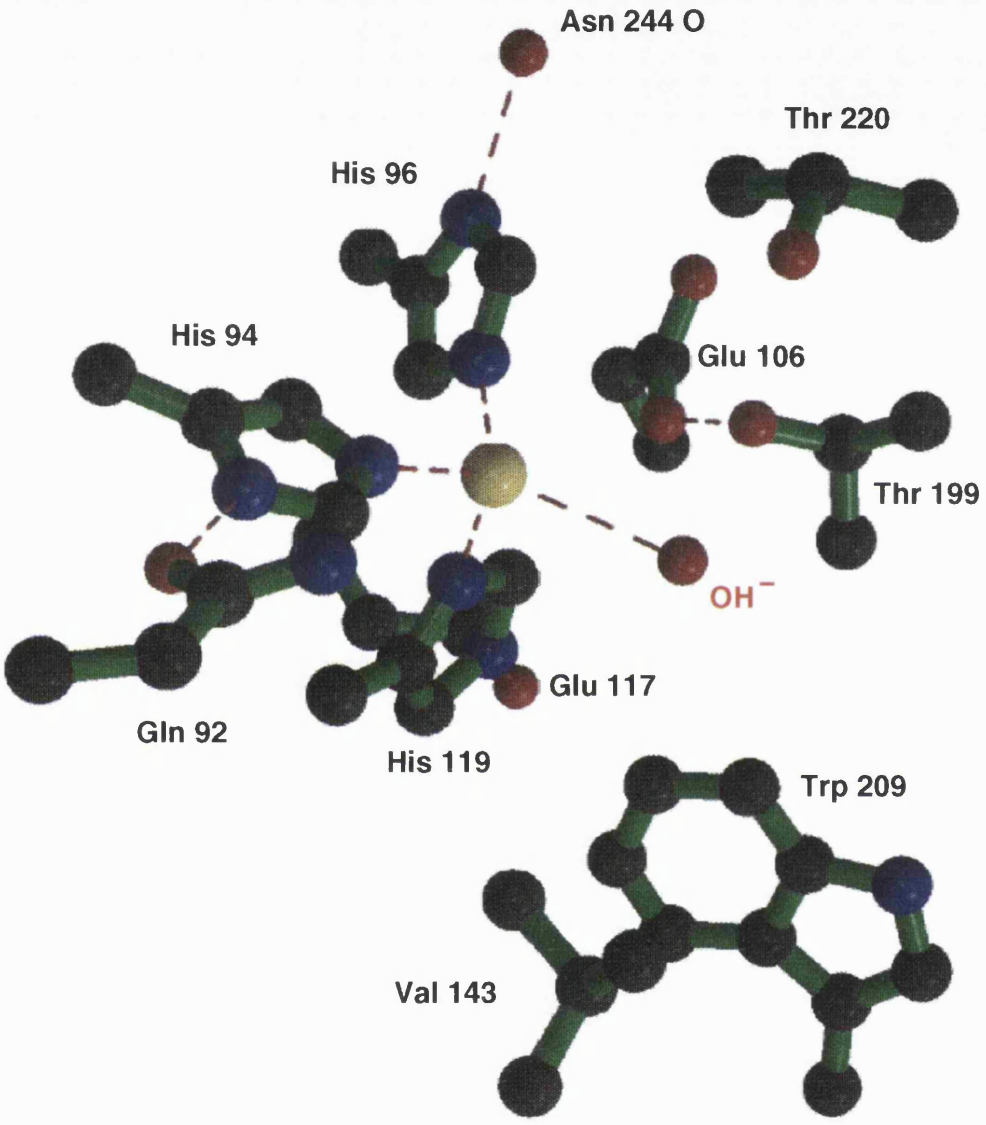


Figure 7.6: A schematic view of the active site of carbonic anhydrase (Eriksson *et al.*, 1986), showing the main catalytic residues and the zinc coordinated to 3 His residues.

The metalloproteinases E.C.3.4.21.x

Table 7.5 indicates that there are several metalloproteinases in the PDB that have a His coordinated active site Zn. These enzymes are astacin E.C.3.4.21.21, pseudolysin E.C.3.4.24.26, thermolysin E.C.3.4.24.27, adamalysin E.C.3.4.24.46 and carboxypeptidase A E.C.3.4.17.1. In general these enzymes all use Zn to activate water that then acts as a nucleophile, attacking the carbonyl bond of the peptide substrate.

Thermolysin has been studied extensively in solution and in the crystal (Cappalonga *et al*, 1992; Hausrath & Matthews, 1984). It displays specificity toward P_1 sidechains of the substrate such as phenylalanine or leucine. Figure 7.7 is a diagrammatic representation of its active sites taken from the thermolysin structure 1*tmn* (Monzingo & Matthews, 1984). There are three triads located, all of tautomer ϵ; Zn 805–$N^{\epsilon 2}$ His 142 $N^{\delta 1}$–Asp 170 $O^{\delta 1}$, Zn 805–$N^{\epsilon 2}$ His 146 $N^{\delta 1}$–Asn 226 $O^{\delta 1}$ and Zn 805–$N^{\epsilon 2}$ His 231 $N^{\delta 1}$–Asp 226 $O^{\delta 1}$. The first two of these are seen coordinating to the active site zinc and are around 1Å from the mean consensus template. The third triad has an *rms* deviation around 2Å from the mean consensus template. In fact, the Asp 226 $O^{\delta 1}$–$N^{\delta 1}$ His 231 diad acts an acid/base catalyst and is not involved in binding the Zn at all. This suggests that our cut–off of 3.0Å was too large, it is, however an interesting example showing the His–Asp pair has more than one function depending on its immediate chemical environment.

7.4.2 Fe–His interactions

- multi centered complexes with distorted geometry

The majority of the Fe–His metal interactions originate from Fe–bound heme or cytochromes and the ELEC groups are generally mainchain carbonyls. Non–heme

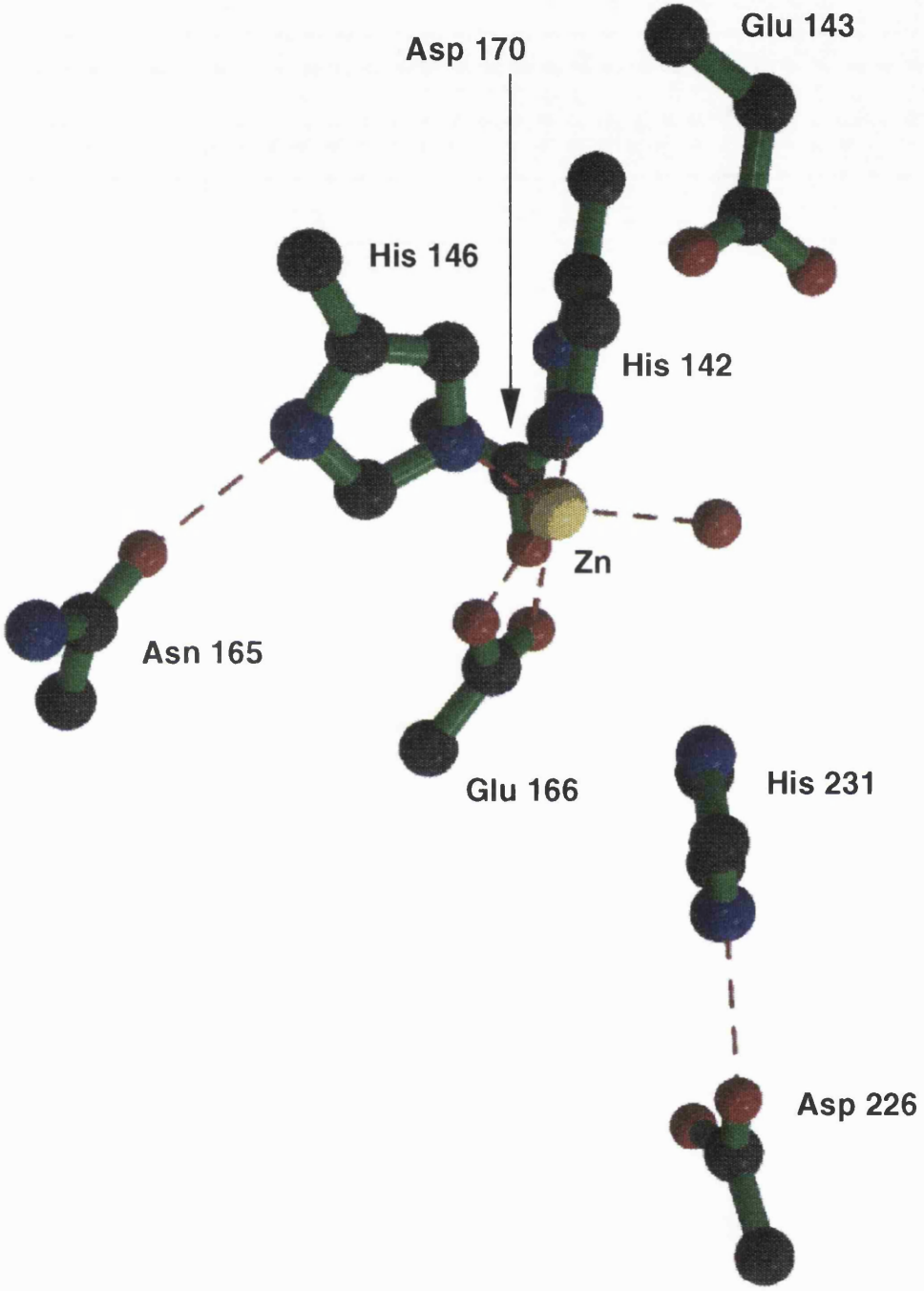


Figure 7.7: A representation of the active sites of the metalloproteinase thermolysin 1*tmn* (Monzingo & Matthews, 1984)

proteins with mononuclear iron have a diversity of roles including several types of oxygen reactions, iron transport, and water insertion. There are two non–heme proteins in the PDB whose structures have been solved and have a metal–His–ELEC triad, R2 subunit of ribonucleotide reductase (RNR) (Rosenzweig *et al.*, 1993) and hemerythrin (Stenkamp *et al.*, 1985).

Hemerythrin (Hr) is an oxygen transporting protein found in invertebrates. Diferrous Hr (deoxy-Hr) binds one O_2 and is simultaneously oxidised to the diferric state (oxy–Hr). Figure 7.8 is a diagram of the active site Fe; there are three metal–His–ELEC triads, all of tautomer ϵ. Two of these tautomers, Fe 1–$N^{\epsilon 2}$ His 54 $N^{\delta 1}$–Gln 24 $N^{\epsilon 2}$ and Fe 1–$N^{\epsilon 2}$ His 25 $N^{\delta 1}$–Asp 22 $O^{\delta 1}$ have high *rms* deviations of around 2.0Å. The reason for these high *rms* distances is that the Asp ELEC groups are distorted away from their ideal hydrogen bonding position.

The R2 subunit of ribonucleotide reductase (RNR) (Rosenzweig *et al.*, 1993) catalyses the conversion of ribonucleotides to deoxyribonucleotides. Figure 7.9 is a 3D representation of the active site Fe atoms and interacting residues of RNR. RNR has two water bound irons in the active B2 subunit of the protein and despite the irons having octahedral coordination, the bidentate Asp ligand distorts the octahedron towards trigonal bipyramidal. There are two metal–His–ELEC triads both with tautomer δ and these are both near the ideal geometry.

Therefore, unlike Hr, the coordinating residues of RNR are not dominated by His. In addition, the Fe atoms of RNR are only bridging twice whereas Hr has 3 bridging species.

These factors contribute to the differing chemical behaviour of the RNR and Hr centers.

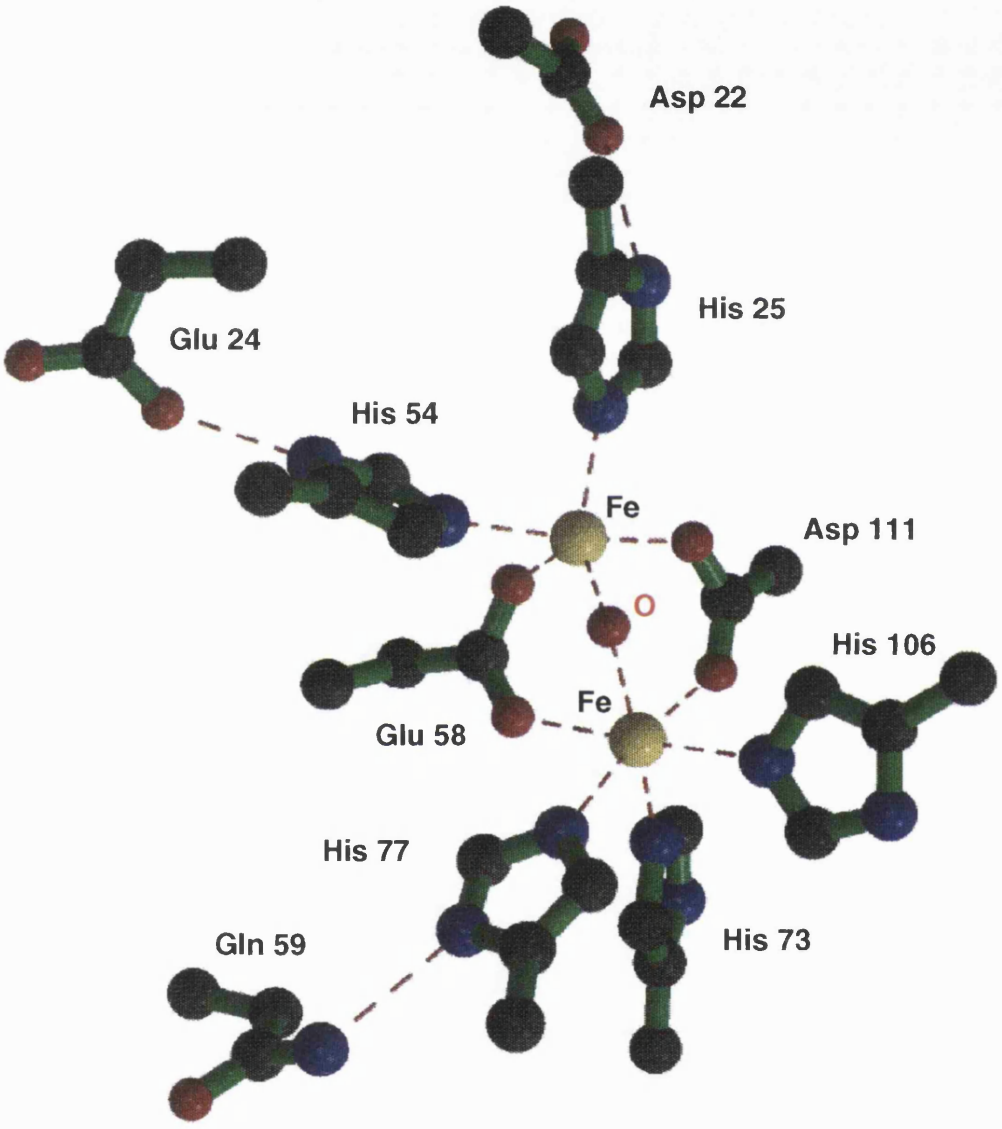


Figure 7.8: A diagram of the active site of hemerythrin (Stenkamp *et al.*, 1985). The Gln 59 N^{ϵ_2} (which is hydrogen bonded to His 77 N^{δ_2} in the diagram) is wrongly assigned and should be swapped with Gln 59 O^{ϵ_1}.

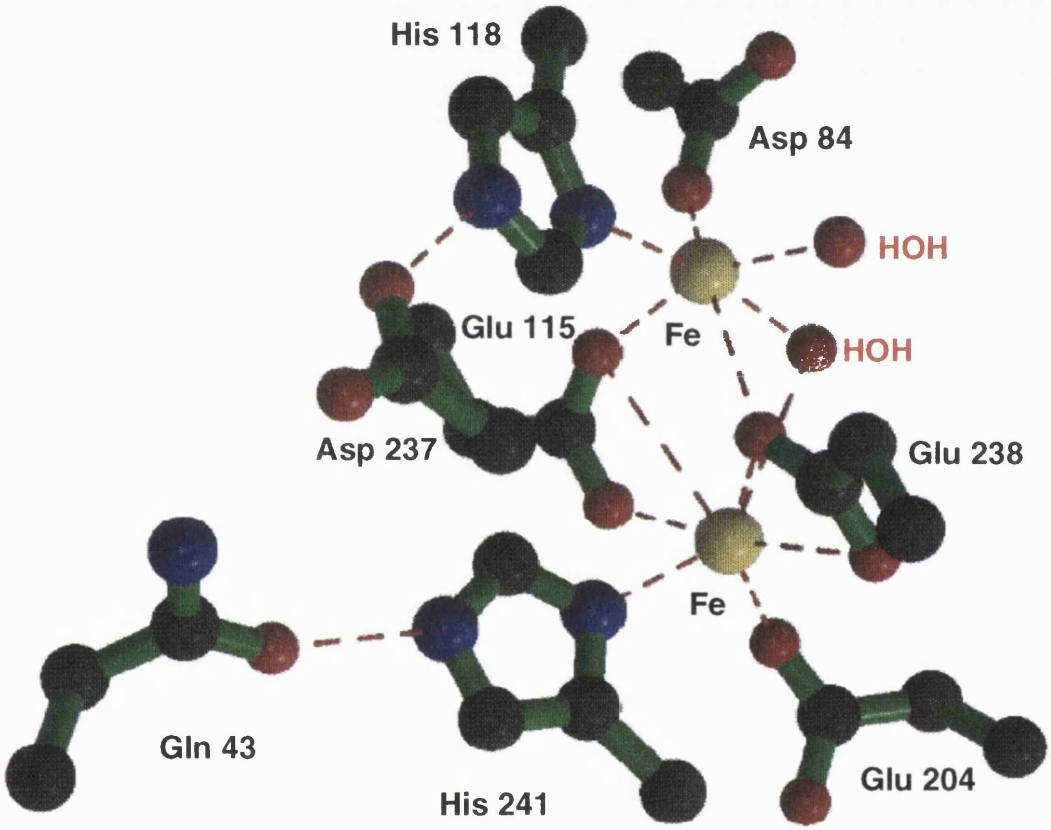


Figure 7.9: A diagrammatic representation of the diferric centre of the oxygen binding protein ribonuclease reductase subunit R2 (Rosenzweig *et al.*, 1993)

7.4.3 Cu–His interactions

-forming three distinct functional groups

Most of the Cu containing proteins in the PDB originate from blue oxidases. These form a sub–group of copper proteins whose metal ions are classified according to their distinct spectroscopic properties: Type I absorbs in the visible region, type II has an undetectable absorption and type III absorbs in the U.V. region.

Type I Cu

The type I copper proteins, azurin from *Pseudomonas aeroginosa* (Adman, 1979), *Alcaligenes denitrificans* (Norris *et al.*, 1983) and poplar plastocyanin (Guss & Freeman, 1983) both have a β–fold and are electron transfer proteins. Both also have tautomer δ metal binding sites with the ELEC groups being non–carbon mainchain atoms.

Type II Cu

Cu, Zn–superoxide dismutase (Tainer *et al.*, 1982) protects the cell against damage by converting the toxic superoxide radical (O_2^-) to hydrogen peroxide and molecular oxygen. The Zn in the active site is buried and is 6.3Å from the solvent accessible type II Cu. The Cu has four His residues ligated in a distorted quadratic arrangement. Figure 7.10 is a diagram of the type II Cu center taken from the superoxide dismutase structure 1*cob* (Djinovic *et al.*, 1992); there is also a water molecule ligated to the Cu centre. The Cu and Zn of Cu,Zn–superoxide dismutase are both ligated to His 61 in the active site. His 61 and the Zn increase the redox potential and thus catalytic activity of the superoxide bound Cu. The triads identified for this enzyme are His 69 N^{δ_1}–N^{δ_1} His 61 N^{ϵ_2}–Cu 810 and Arg 141 N^{η_2}–N^{δ_1} His 46 N^{ϵ_2}–Cu 810. The first triad is unusual because the His 61 is contacting the His 69 via the Zn, these two residues are not directly hydrogen

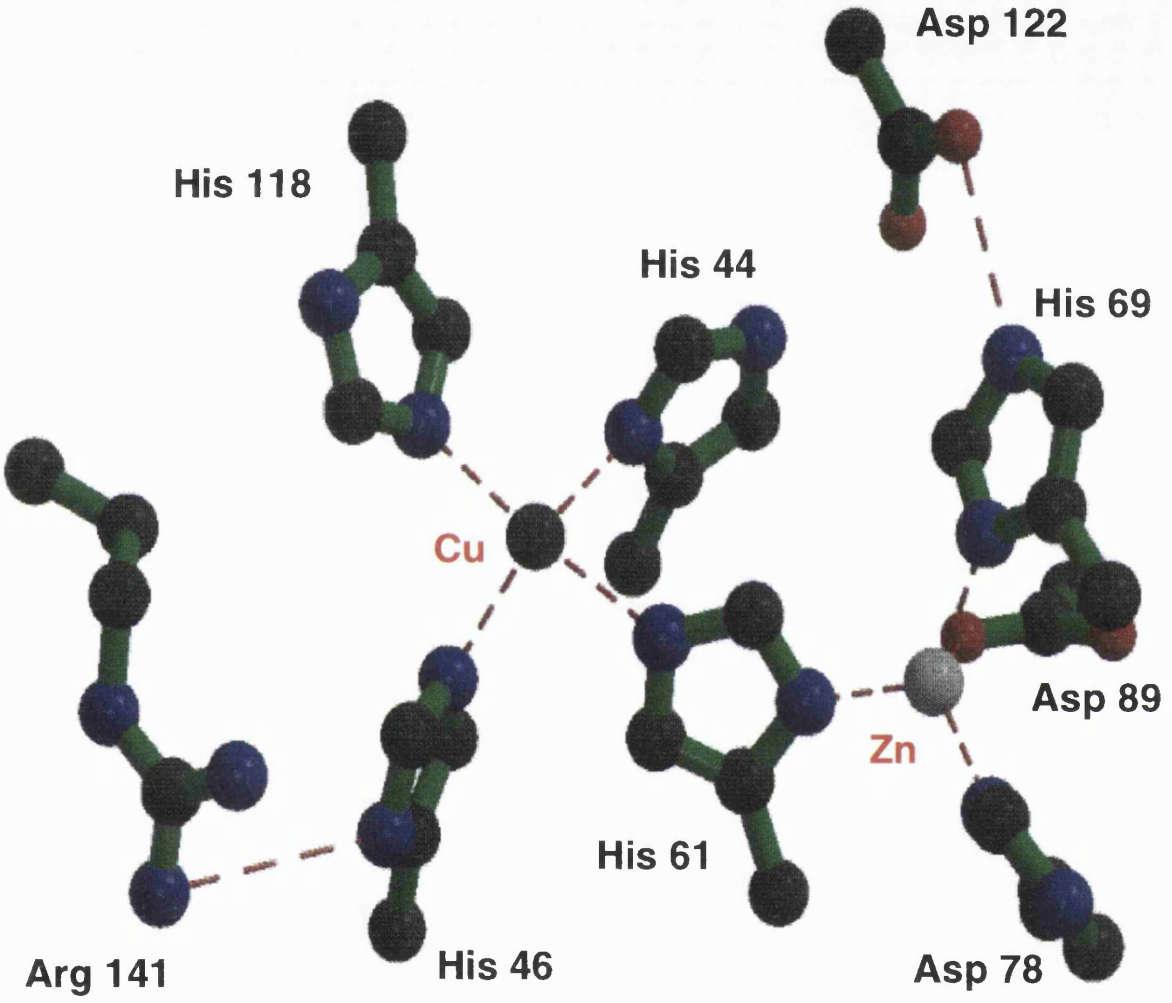


Figure 7.10: A representation of the active site Cu and Zn metals from superoxide dismutase, 1*cob* (Djinovic *et al.*, 1992)

bonded to each other. The second triad has a high *rms* distance because the Arg $N^{\eta 2}$ is out of the plane of the His 46 residue (Figure 7.10); though Arg 144 is sequentially conserved and important electrostatically and mechanistically it does not have the role of an ELEC in the ELEC–His diad. The third triad of Asp 122 $O^{\delta 1}$–$N^{\epsilon 2}$ His 69 $N^{\delta 1}$–Zn is an authentic metal–His–ELEC triad and is 1.02Å deviation from the mean consensus template.

Type III Cu

Hemocyanin from *Panulirus interruptus* (Volbeda & Hol, 1989) has type III copper; this consists of a pair of copper atoms ligated by 6 His residues at their $N^{\epsilon 2}$ atoms. It has an α fold (Gaykema *et al.*, 1984) and the same function as the Type I Cu proteins. Figure 7.11 is a diagram of the metal centre of this protein; there is one tautomer ϵ triad formed but this is distorted having a typical *rms* value around 1.8Å from the mean template and the sidechain of the Glu is pointing away from the His 348 $N^{\delta 1}$. The other His residues surrounding the Cu metals have either mainchain oxygen atoms making up their triad or their hydrogen bonding potential is unsatisfied.

7.5 Comparison of the Nu:–His–ELEC and metal–His–ELEC triads

We saw in chapter 5 that a Nu:–His–ELEC template could be defined that is able to identify all serine proteinase, lipase and α/β hydrolase fold enzymes within the PDB with the exclusion of all other non–catalytic interactions. This triad has the nucleophilic Nu: group interacting with the His $N^{\epsilon 2}$ atom and is therefore equivalent to the tautomer ϵ conformation in the metal–His–ELEC triad. In addition, we found the cysteine proteinases have the other, tautomer δ conformation and

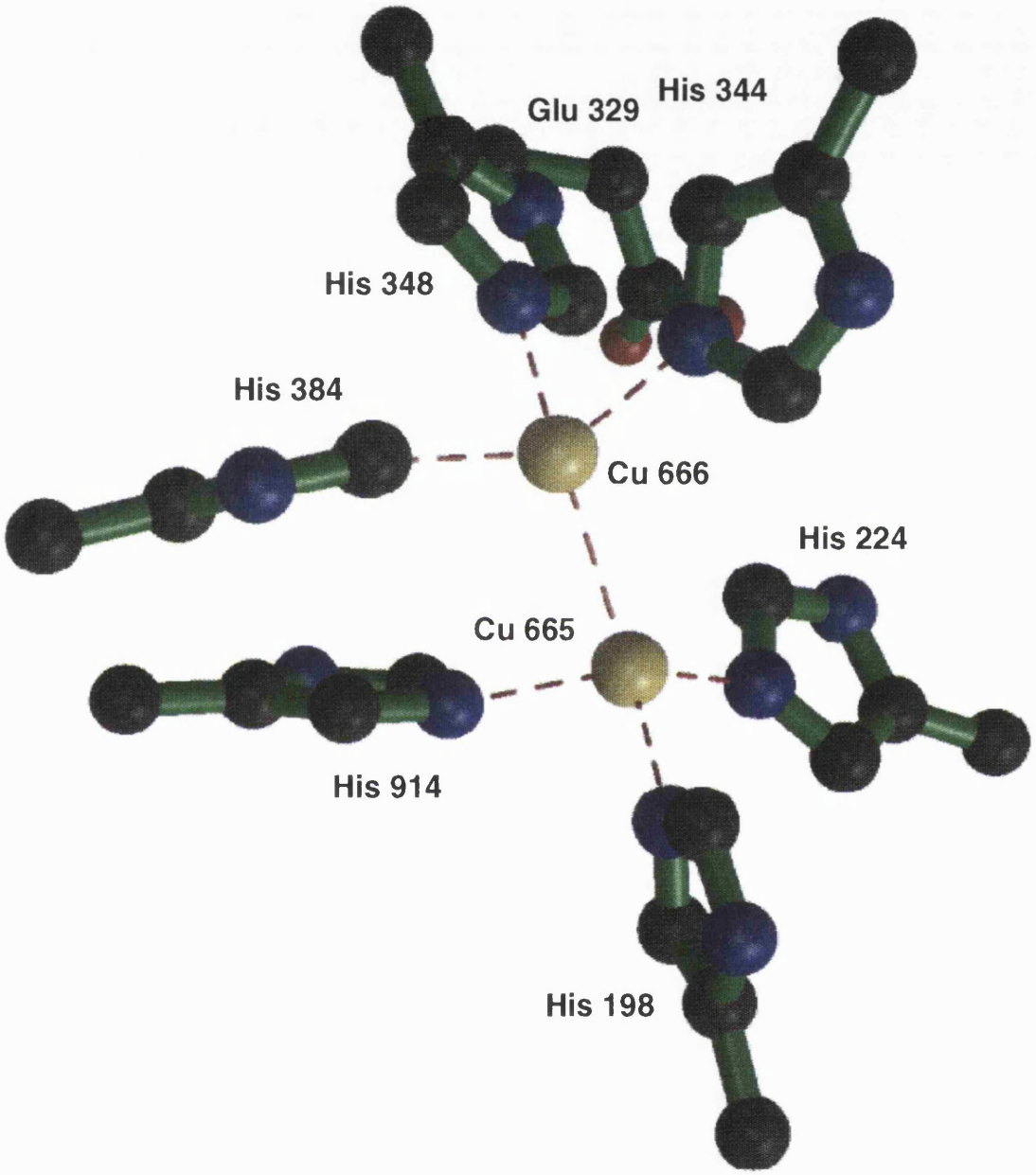


Figure 7.11: A diagrammatic representation of the type III Cu centre taken from the hemocyanin structure 1*hc*1 (Volbeda & Hol, 1989)

a consenus template has been constructed for this enzyme.

The His–ELEC diad, in concert with either a nucleophilic or metal group, forms triads of different function. In addition, these triads are always in functional regions of the protein and are clearly a result of convergent evolution as they occur in enzymes and proteins of diverse tertiary fold and functional type.

We have now defined two templates for each of the Nu:–His–ELEC and metal–His–ELEC triads (tautomer δ and ϵ). It would be interesting to compare the conformation of these triads. In this comparison the coordinates of the triads for the metal–His–ELEC group are used when the ELEC group is a sidechain atom. Figure 7.12, is a 3D representation of the consensus templates. The top triad is tautomer ϵ which is equivalent to the Group 1–2–3 template in Chapter 5. The *rms* distance between these two templates is 1.08Å. The metal atom (black) is in an ideal position to interact with the His $N^{\epsilon 2}$ and the nucleophilic Nu: atom is 1.25Å from the metal atom. This illustrates the difference between the mechanism of action of the two triads; the nucleophilic atom is positioned to attack the substrate, not to form a stable hydrogen bond with the His $N^{\epsilon 2}$. Indeed, if this hydrogen bond were stable then the role of the His–ELEC diad as an acid/base catalyst would be hindered.

The electrostatic atoms, ELEC, of the two triads are 0.76Å from each other despite performing the same functional role. This is above the atomic resolution of an atom in a well resolved X–ray structure and illustrates the distortion that occurs in the metal binding site His–ELEC diad.

The other consensus template conformations, tautomer δ, are shown in the bottom triad in Figure 7.12. The only Nu:–His–ELEC triad of this conformation is from the cysteine proteinases. The metal atom is in an ideal position to interact with the His $N^{\delta 1}$ atom and the nucleophilic S^{γ} atom from the enzyme lies 2.1Å below the metal atom. The positions of these atoms are, like the tautomer ϵ

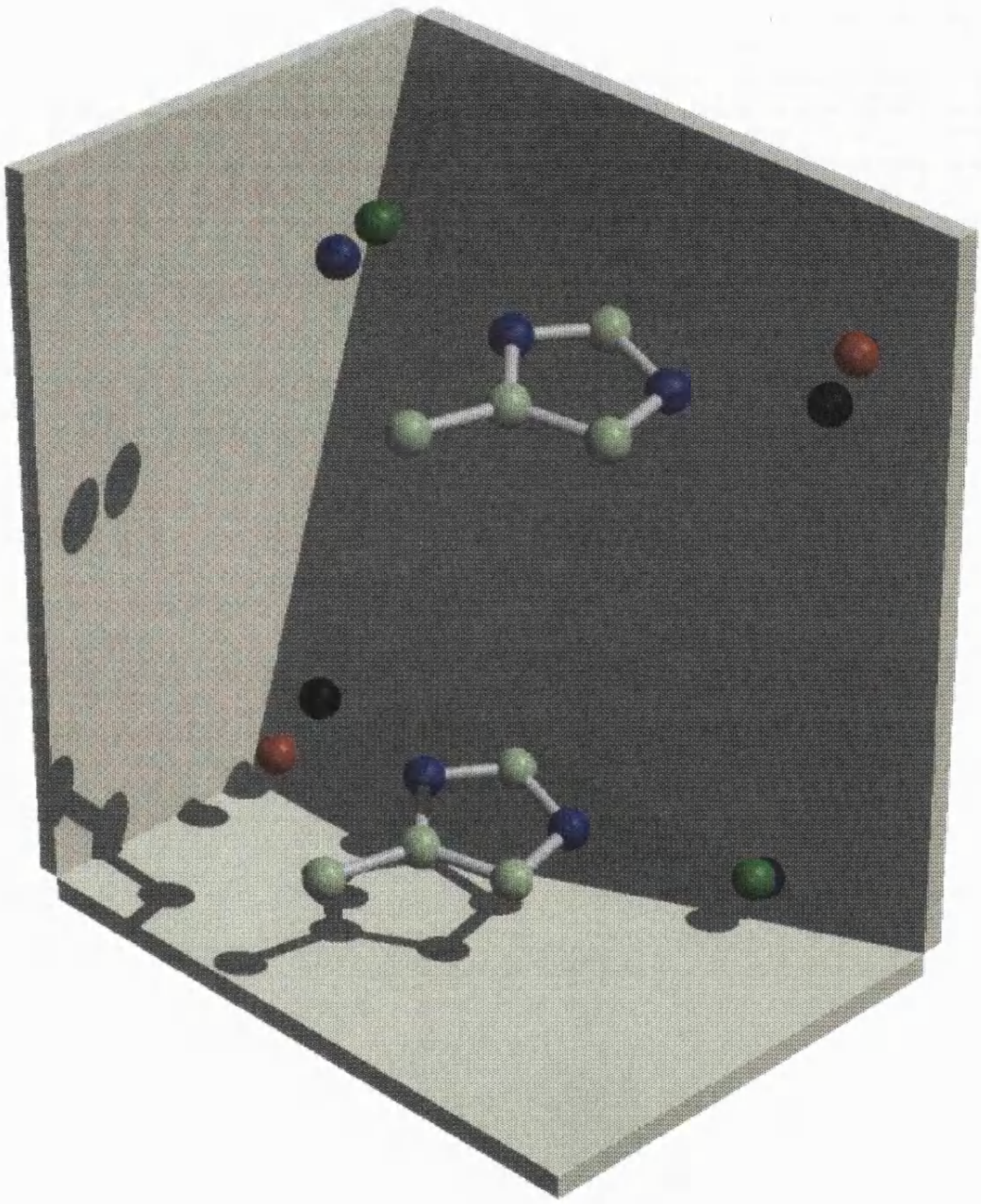


Figure 7.12: A 3D representation of the consensus templates of the catalytic triads of type metal–His–ELEC and Nu:–His–ELEC for the tautomers δ (bottom) and ϵ (top). Metal atoms are black; ELEC group for the metal–His–ELEC triad are green; ELEC group for the Nu:–His–ELEC triad are blue; red atoms are Nu: groups.

conformation mentioned above, a reflection of their functional differences. The electrostatic atoms are only 0.44Å from each other indicating their similar functional role.

7.6 Conclusion

The analysis of metal–His–ELEC triads in the PDB reveals that they occur in a wide range of catalytic and functional metal centres yet there are no cases where this triad is found as part of a structural metal center. This occurs because His is a hard ligand and is most suitable for activating bound metals for their functional role.

In general, metal–His–ELEC triads found in enzyme active sites are in an ideal conformation and this triad functions to activate the bound metal. For example, in carbonic anhydrase which has a single metal ligated to several triads of type metal–His–ELEC, the triads are found to be conserved in a geometry that allows ideal interactions to occur.

When the ELEC group is a sidechain atom it usually originates from Asp, Asn, Glu and Gln. In addition, these triads are usually found in enzyme active site metal centres. We have already noted that the ΔpK_a for the ligated His sidechain atom is greater when the ELEC group is one of these species. This enables the pK_a of the His ligand to be 'fine–tuned' therefore optimising the electrostatic properties of the catalytic metal.

Conversely, non–enzymatic metal centres have many mainchain ELEC groups. This reflects the different function of these sites such as ligand binding and electron transport. Due to the heterogeneity of the metal binding sites in proteins it is not possible to identify catalytic metal–His–ELEC triads with the exclusion of all other interactions. In reality, these sites should be first compared with more

accurate crystal structures of metal–protein interactions such as those deposited in the Cambridge Structural Database (Allen *et al.*, 1979).

Furthermore, the TESS program could be modified to make the metal part of the reference frame (chapter 4), this would allow a more direct comparison of metal centers and we could include all metal interactions rather than just metal-His.

7.7 References

Adman E.T. (1979) A comparison of the structures of electron transfer proteins *Biochim. Biophys. Acta.* **549** 107–144

Allen F.H., Bellard S., Brice M.D., Cartwright B.A., Doubleday A., Higgs H., Hummelink T., Hummelink-Peters B.G., Kennard O., Motherwell W.D.S., Rodgers J.R. & Watson D.G. (1979). The Cambridge Crystallographic Data Centre: computer-based search, retrieval, analysis and display of information. *Acta Cryst.*, **B35**, 2331–2339.

Argos P., Garavito R.M., Eventoff W., Rossmann M.G. & Branden C.I. (1978) Structural stability of metal binding sites *J. Mol. Biol.* **126** 141–158

Babu Y.S., Sack J.S., Greenhough T.J., Bugg C.E. Means A.R. & Cook W.J. (1985) Three–dimensional structure of calmodulin *Nature* **315** 37–40

Bielka H., Dixon H. B. F., Karlson P., Liebecq C., Sharon N., Van Lenten E. J., Velick S. F., Vliegenthart J. F. G. & Webb E.C. (1992) *E. C. Enzyme nomenclature 1992: recommendations of the Nomenclature Committee of the International Union Of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes. Nomenclature Committee*

of the International Union of Biochemistry Academic Press, Inc., (London) Ltd.

Bode E., Gomis–Rueth X., Huber R., Zwilling R. & Stoecker W. (1992) Structure of astacin and implications for activation of astacins and Zn–ligation of collagenases *Nature* **358** 164–168

Borkakoti N., Winkler F.K., Williams D.H., D'Arcy A., Broadhurst M.J., Brown P.A., Johnson W.H. & Murray E.J. (1994) Structure of the catalytic domain of human fibroblast collagenase complexed with inhibitor *Nature Struct. Biol.* **1** 106–110

Cappalonga A.M., Alexander R.S. & Christianson D.W. (1992) Structural comparison of sulfodiimine and sulfonamide inhibitors in their complexes with Zn enzymes *J. Biol. Chem.* **267** 19192–19197

Carver J.A. & Bradbury J.H. (1984) Assignment of H–1–NMR resonances of histidine and other aromatic residues in met–myoglobin, cyano–myoglobin, oxy–myoglobin, and (carbon monoxy) myoglobin *Biochemistry* **23** 4890–4905

Chakrabarti P. (1990) Geometry of interaction of metal ions with histidine residues in protein structures *Protein Engineering* **4** 57–63

Chen L., Durley R., Poliks B.J., Hamada K., Chen Z., Matthews F.S., Davidson V.L., Satow Y., Huizanga E., Vellieux F.M.D. & Hol W.G.J. (1992) Crystal structure of an electron–transfer complex between methylamine dehydrogenase and amicyanin *Biochemistry* **31** 4959–4964

Christianson D.W. & Alexander R.S. (1990) Another catalytic triad? *Nature* **346** 225

Collyer C.A. & Blow D.M. (1990) Observations of reaction intermediates and the mechanism of aldose–ketose interconversion by D–xylose isomerase *Proc. Natl. Acad. Sci.* **87** 1362–1366

Cooper J.B., Driessen H.P.C., Wood S.P., Zhang Y. & Young D. (1994) Crystallisation and preliminary X–ray analysis of the iron–dependent superoxide dismutase from *Mycobacterium tuberculosis J. Mol. Biol.* **235** 1156–1158

Djinovic K., Coda A., Antolini L., Pelosi G., Desideri A., Falconi M., Rotilio G. & Bolognesi M. (1992) Crystal–structure and refinement of the semisynthetic cobalt–substituted bovine erythrocyte superoxide dismutase at 2.0Å resolution *J. Mol. Biol.* **226** 227–238

Docherty A.J.P., O'Connell J., Crabbe T., Angal S., & Murphy G. (1992) The matrix metalloproteinases and their natural inhibitors. Prospects for treating degenerative tissue diseases *Trends Biotech.* **10** 200–207

Eriksson A.E., Jones T.A. & Liljas A. (1986) *Zinc Enzymes* Bertini I., Luchinat C., Maret W., Zeppezauer M. eds. 317–328 *Burkhauser, Cambridge, Massachussetts*

Fujinaga M. & James M.N.G (1987) Rat submaxillary serine proteinase, tonin. Structure solution and refinement at 1.8Å resolution *J. Mol. Biol.* **195** 373–391

Gaykema W.P.J., Hol W.G.J., Vereijken J.M., Soeter N.M., Bak H.J. & Beintema J.J. (1984) 3.2Å structure of the copper containing oxygen carrying protein *Panulirus interruptus* haemocyanin. *Nature* **309** 23–29

Gurd F.R.N. & Wilcox P.E. (1956) Complex formation between metallic cations and proteins, peptides and amino acids *Adv. Protein Chem.* **11** 311–418

Guss & Freeman (1983) Structure of oxidised poplar plastocyanin at 1.6Å resolution *J. Mol. Biol.* **169** 521–563

Hooft R.W.W., Vriend G., Snader C. & Abola E.E. (1996) Errors in protein structures *Nature* **381** 272

Inaka K., Kuroki R., Kikuchi M. & Matsushima M. (1991) Crystal of the apo– and holo–mutant human lysozymes with an introduced Ca binding site *J. Biol. Chem.* **266** 20666–20671

Kendrew J.C., Dickerson R.E., Strandberg R.E. Hart R.G. & Davies D.R. (1960) Structure of myoglobin. A 3D Fourier synthesis at 2Å resolution *Nature* **185** 422–427

Klein C. & Schultz G.E. (1991) Structure of cyclodextrin glycosyltransferase refined at 2.0Å resolution *J. Mol. Biol.* **217** 737–758

Kloek A.P., Yang J., Mathews F.S. & Goldberg D.E. (1993) Expression, characterisation and crystallisation of oxygen–avid *Ascaris* hemoglobin domains *J. Biol. Chem.* **268** 17669–17685

Kukimoto M., Nishiyana M., Murphy M.E.P., Turley S., Adman E.T., Horinouchi S. & Beppu T. (1993) X–ray crystal structure and site–directed mutagenesis of a nitrite reductase from *Alcaligenes faecalis* S–6; role of two Cu atoms in nitrite reduction *Biochemistry* **33** 5246–5252

Lah M.E., Dixon M.M., Partridge K.A., Stallings W.C., Fee J.A. & Ludwig M.L. (1995) Structure–function in *E. coli* iron superoxide dismutase: comparisons with the manganese enzyme from *Thermus thermophilus Biochemistry* **34** 1646–1660

Lauble H., Kennedy M.C., Beinart H. & Stout D.C. (1994) Crystal structures of aconitase with *trans*–aconitase and nitrocitrate bound *J. Mol. Biol.* **237** 437–451

Li De La I., Papamichael E., Sakarelos C., Dimicoli J. L. & Prange T. (1990) Interaction of the peptide CF3–LEU–ALA–NH–C6H4–CF3 with porcine pancreatic elastase. X–ray studies at 1.8Å. *J. Mol. Recog.* **3** 36–44

Ludwig M.L., Metzger A.L., Partridge K.A. & Stallings W.C. (1991) Manganese superoxide dismutase from *Thermus thermophilus.* A structural model refined at 1.8Å resolution *J. Mol. Biol.* **219** 335–338

Matsura Y., Kusunoki M., Harada W. & Kakudo M. (1984) Structure and possible catalytic residues of taka–amylase A *J. Biochem.* **95** 697–702

Marquart M., Walter J., Deisenhofer J., Bode W. & Huber R. (1983) The geometry of the reactive site and of the peptide groups in trypsin, trypsinogen and its complexes with inhibitors *Acta Crystallogr., sect. B* **39** 480–490

Meader W.E., Means A.R. & Quiocho F.A. (1992) Target enzyme recognition by calmodulin: 2.4Å structure of a calmodulin–peptide complex *Science* **257** 1251–1255

Messerschmidt A., Ladenstein R., Huber R., Bolognesi M., Avigliano L., Petruzzelli R., Rossi A. & Finazzi–Agro A. (1991) Refined crystal structure of ascorbate oxidase at 1.9Å resolution *J. Mol. Biol.* **224** 179–205

Monzingo A.F. & Matthews B.W. (1984) Binding of N-carboxymethyl dipeptide inhibitors to thermolysin determined by X–ray crystallography. A novel class of transition state analogues for zinc peptides *Biochemistry* **23** 5724–5729

Murphy G.J.P., Murphy G. & Reynolds J.J (1991) The origin of matrix metalloproteinases and their familial relationships *FEBS lett.* **289** 4–7

Nar H., Messerschmidt A., Huber R., Van De Kamp M. & Canters G.W. (1991) Crystal structure analysis of oxidised *Pseudomonas aeroginosa* azurin at pH 5.5 amd pH 9.0. A pH–induced conformational transition involves a peptide bond flip *J. Mol. Biol.* **221** 765–777

Nordlund P., Sjuberg B.M. & Eklund H. (1990) 3D structure of the free radical protein of ribonucleotide reductase *Nature* **345** 593–596

Norris G.E., Anderson B.F. & Baker E.N. (1983) Structure of azurin from *Alcaligenes denitrificans* at 2.5Å resolution *J. Am. Chem. Soc.* **108** 2784–2785

Norris G.E., Anderson B.F. & Baker E.N. (1991) Molecular replacement solution of the structure of apolactoferrin, a protein displaying large scale conformational change *Acta. Crystallogr., Sect. B.* **47** 998–1002

Ohlendorf D.H., Lipscomb J.D. & Weber P.C. (1988) Structure and assembly of protocatechuate 3,4–dioxygenase *Nature* **336** 403–408

Parge H.E., Hallewell R.A. & Tainer J.A. (1992) Atomic structures of wild–type and thermostable mutant recombinant Cu, Zn superoxide dismutase *Proc. Natl. Acad. Sci, USA* **89** 6109–6113

Perutz M.F. (1970) Stereochemistry of cooperative effects of haemoglobin *Nature* **228** 726–734

Perutz M.F., Muirhead H., Cox J.M., Goamann L.C.G., Matthews F.S., McGandy E.L. & Webb L.E. (1968) 3D Fourier synthesis of horse oxyhaemoglobin at 2.8Å resolution: (1) X–ray analysis *Nature* **219** 29–32

Poulos T.L., Edwards S.L., Wariishi H. & Gold M.H. (1992) Crystallographic refinement of lignin peroxidase at 2.0Å. *J. Biol. Chem.* **268** 4429–4440

Rosenzweig A.C., Frederick C.A., Lippard S.J. & Nordlund P. (1993) Crystal structure of a bacterial non–heme iron hydroxylase that catalyses the biological oxidation of methane *Nature* **366** 537–543

Stenkamp R.E., Sieker L.C., Jensen L.H., Mc Callum J.D. & Sanders–Loehr J. (1985) Active site structures of deoxyhemerythrin and oxyhemerythrin *Proc. Nat. Acad. Sci. USA* **82** 713–716

Tainer J.A., Getzoff E.D., Beem K.M., Richardson J.S. & Richardson D.C. (1982) Determination and analysis of the 2Å structure of Cu, Zn superoxide dismutase *J. Mol. Biol.* **160** 181–195

Tanokura M. (1983) H-1-NMR nuclear magnetic–resonance titration cuves and micorenvironments of aromatic residues in bovine pancreatic ribonuclease A *Biochim. Biophys. Acta.* **742** 576–585

Vallee B.L. & Auld D.S. (1990) Zinc coordination, function, and structure of zinc enzymes and other proteins *Biochemistry* **29** 5647–5659

Voet D. & Voet J.G. (1990) Biochemistry *Wiley, New York*

Volbeda A. & Hol W.G.J. (1989) Crystal structure of hexameric haemocyanin from *Panulirus interruptus* refined at 3.2Å resolution *J. Mol. Biol.* **249** 249–262

Wang J., Mauro J.M., Edwards S.L., Oatley S.J., Fishel L.A., Ashford V.A., Xuong N–H. & Kraut J. (1990) X–ray structures of recombinant yeast cytochrome *c* peroxidase and 3 heme–cleft mutants prepared by site–directed mutagenesis *Biochemistry* **29** 7160–7173

Wery J.P., Schevxitz R.W., Clawson D.K., Bobbitt J.L., Dow E.R., Gamboa G., Goodson T. Junior, Hermann R.B., Kramer R.M., McClure D.B., Mihelich E.D., Putnam J.E., Sharp J.D., Stark D.H., Teater C., Warrick M.W. & Jones N.D. (1991) Structure of recombinant human rheumatoid arthritic synovial fluid phospholipase A_2 at 2.2Å resolution. *Nature* **352** 79–82

Chapter 8

Creating a database of 3D enzyme active site templates

8.1 Introduction

This chapter describes the work performed so far to create a database of 3D templates for enzyme active sites. It is clear that creating over 200 templates to describe all the enzymes in the PDB would be very time consuming so, ideally, we would like an automatic method. There are various options available, such as extracting the functional sequence motifs from databases such as PROSITE (Bairoch & Bucher, 1994), however in practice, this provides only a few potential templates (see below).

In addition, there are other problems encountered when a new consensus template is defined. To illustrate this, the derivation of 3D consensus templates from two different enzymes, ribonuclease and lysozyme, are described. It becomes clear that more than one template may been needed to describe the active site of more than one enzyme. Convergent evolution can result in functionally similar enzymes from different species that have catalytic residues in different conforma-

tions. In addition, there may be more than one conformation of the catalytic residues depending on the types of inhibitors bound to the active site. Furthermore, the lysozyme active site is known to undergo considerable conformational change during catalysis; since the structures of these enzymes are generally for the ground state, so is the consensus template.

An option that could overcome these problems would be to use the world wide web (WWW). An interface could be set up that enables experts on a given enzyme to deposit the 3D templates in the database.

8.2 Defining 3D templates automatically

Ideally, a method is needed that automatically defines the 3D enzyme active site templates. To do this, we need to identify the catalytic residues and atoms that are directly involved in the chemical catalysis reactions. For example, the serine proteinase has Ser, His and Asp as catalytic residues, yet the consensus template consists of only Asp O^{δ}, Ser O^{γ} and the His sidechain.

8.2.1 Automatically identifying catalytic residues

There are several possibilities available that may automatically identify catalytic residues. Firstly, the PROSITE (Bairoch & Bucher, 1994) sequence database has some active site sequence motifs. Table 8.1 summarises this information. There are only 8 enzymes in this database that have 2 or more active site residue motifs. Three of these, α–lytic proteinase, subtilisin and carboxypeptidase C are serine proteinases and only subtilisin has all 3 His, Asp and Ser residues documented. Therefore this is not a viable option.

The second strategy could be to extract the site record information directly from the PDB files. Table 8.2 is a summary of the site records for all the enzymes

OXIDOREDUCTASES	HYDROLASES	ISOMERASES
malate dehydrogenase E.C.1.1.1.37	**phospholipase A2 E.C.3.1.1.4**	**xylose isomerase E.C.5.3.1.5**
Leu 157 Asn 160	His 48 Asp 99	Lys 182 His 53
	β–amylase E.C.3.2.1.2 Asp 101 Glu 186	
	carboxypeptidase C E.C.3.4.16.5 His 397 Ser 146	
	α–lytic proteinase E.C.3.4.21.12 His 57 Ser 195	
	actinidin E.C.3.4.21.14 His 162 Cys 25 Asn 182	
	subtilisin E.C.3.4.21.62 His 62 Ser 215 Asp 32	

Table 8.1: Enzymes in the PROSITE database (Bairoch & Bucher, 1994) which have catalytic residues listed in the site records. Only those enzymes with more than 1 catalytic residue in the records are listed.

in the PDB. It is not possible to tell from these site records whether the residues are catalytic or constitute the ligand binding site. However, this appears to be the most extensive list available.

Another approach is by Lichtarge *et al.*, 1996. They have used an evolutionary trace method that predicts active site and functionally important residues from sequence conservation patterns in homologous proteins. These are mapped onto protein surfaces to generate clusters identifying functional interfaces. They have successfully identified the binding sites for the SH2 and SH3 modular signalling domains and DNA binding domains of the nuclear hormone receptors.

Zvelebil & Sternberg, 1988 tried to predict the location of key catalytic residues in proteins by performing an analysis of the structural environment of 17 enzymes whose active site residues were already identified. They found, in general, that the environment of catalytic residues is similar to that of polar sidechains that have low accessibility to solvent. They developed two algorithms based on this data which, with limited success, was able to identify catalytic residues in other enzyme active sites.

Peters *et al.*, 1996 have developed a program based purely on geometric criteria that searches for clefts on the protein surface, they locate more than 95% of ligand binding sites in the PDB; this could be used as a starting point for the identification of the active site residues.

Laskowski *et al.*, 1996 have found that the largest cleft in a representative dataset of protein structures in the PDB is in fact its ligand binding site in more than 80% of the proteins.

OXIDOREDUCTASES					
aldehyde reductase E.C.1.1.1.21	His 110	Tyr 48	Lys 77	Cys 298	
aldehyde reductase E.C.1.1.1.37	Arg 81	Arg 87	Asp 150	Arg 153	His 177
TRANSFERASES					
aspartate aminotransferase E.C.2.6.1.1	Lys 258				
HYDROLASES					
triacylglycerol lipase E.C.3.1.1.3	Ser 209	Glu 341	His 449		
deoxyribonuclease I E.C.3.1.21.1	Glu 39	Gly 78	His 134	Asp 212	His 252
ribonuclease T1 E.C.3.1.27.3	Tyr 38	Lys 40	Glu 58	Arg 77	
pancreatic ribonuclease E.C.3.1.27.5	His 12	Lys 41	Val 43	Asn 44	Thr 45
	His 119	Phe 120	Asp 121	Ser 123	
micrococcal nuclease E.C.3.1.31.1	Arg 35	Glu 43	Arg 87		
α–amylase E.C.3.2.1.1	Asp 229	Glu 257	Asp 328		
glucan 1,4–α–glucosidase E.C.3.2.1.3	Asp 55	Arg 305	Leu 177	Arg 54	
cellulose E.C.3.2.1.4	Asp 338	His 397	Ser 146	Glu 145	
lysozyme E.C.3.2.1.17	Glu 35	Asp 52			
exo-α-sialidase E.C.3.2.1.18	Arg 37	Arg 56	Asp 62	Met 99	Asp 100
	Trp 121	Trp 128	Leu 175	Glu 231	Arg 246
	Arg 309	Tyr 342	Glu 361		
methionyl aminopeptidase E.C.3.4.11.18	Asp 97	Asp 108	His 171	Glu 204	Glu 235
	Ala 2				
chymotrypsin E.C.3.4.21.1	His 57	Asp 102	Ser 195	Leu 45	
trypsin E.C.3.4.21.4	His 57	Asp 102	Ser 195		
thrombin E.C.3.4.21.5	His 57	Asp 102	Ser 195		
pancreatic elastase E.C.3.4.21.36	Asp 108	His60	Ser 203		
subtilisin E.C.3.4.21.62	Asp 32	His 64	Ser 221		
endopeptidase K E.C.3.4.21.64	Asp 39	His 69	Ser 224		
thermitase E.C.3.4.21.66	Asp 38	His 71	Ser 225		
protein C E.C.3.4.21.69	His 211	Asp 257	Ser 360		
papain E.C.3.4.22.2	Cys 25	His 159	Asn 175		
actinidain E.C.3.4.21.14	Cys 25	His 162	Asn 182	Gln 19	Trp 184
caricain E.C.3.4.22.30	Cys 25	His 159	Asn 179		
pepsin A E.C.3.4.23.1	Asp 32	Asp 215			
renin E.C.3.4.23.15	Asp 32	Asp 215			
rhizopepsin E.C.3.4.23.21	Asp 35	Asp 218			
endothiapepsin E.C.3.4.23.22	Asp 32	Asp 215			
mucoropepsin E.C.3.4.23.23	Asp 32	Asp 215			
astacin E.C.3.4.24.21	His 92	His 96	His 102	Tyr 149	
LYASES					
ribulose bisphosphate carboxylase E.C.4.1.1.39	Lys 201	Asp 203	Glu 204		
2,2-dialkylglycine decarboxylase(pyruvate) E.C.4.1.1.64	Gln 52	Met 53	Phe 79	Thr 110	Gly 111
	Asn 115	Ser 137	Trp 138	Met 141	Glu 210
	Ser 214	Ser 215	Asp 243	Ala 245	Gln 246
	Lys 272	Tyr 301	Thr 303	Asn 394	Arg 406
carbonate dehydrogenase E.C.4.2.1.1	Tyr 7	Val 63	His 64	Ser 65	His 67
	Asn 69	Gln 92	Glu 106	Glu 117	Thr 199
	His 200	Phe 91	Ala 121	Leu 131	Ala 135
	Leu 141	Val 143	Leu 198	Pro 201	Pro 202
	Tyr 204	Ser 206	Val 207	Trp 209	
aconitate hydratase E.C.4.2.1.3	Gln 72	Asp 100	His 101	His 147	Asp 165
	Ser 166	His 167	Asn 170	Asn 258	Gln 262
phosphopyruvate hydratase E.C.4.2.1.11	Glu 168	Glu 211	lys 345	His 373	Lys 396
mandelate racemase E.C.5.1.1.2	Lys 166	His 297			
triosephosphate isomerase E.C.5.3.1.1	Glu 165	His 95	Ser 96	Lys 13	
LIGASES					
glutamate–ammonia ligase E.C.6.3.1.2	Glu 129	Glu 131	His 269	Glu 212	Glu 220
	Glu 357	Arg 321	Gly 265	Arg 339	Arg 359
	Asp 50				
biotin-(acetyl-CoA-carboxylase) ligase E.C.6.3.4.15	Ser 89	Thr 90	Asn 91	Gln 112	Gly 115
	Tyr 132	Lys 183	Ile 187	Leu 188	Gly 204
	Ala 205				

Table 8.2: Potential catalytic residues extracted from the site records of the PDB files.

8.2.2 Identifying the atoms of the catalytic residues involved in catalysis

A method is also needed to identify the atoms of the catalytic residues that are involved in the chemical catalysis.

One approach would be to search for ligands in the PDB files to identify the atoms which are contacting the ligand; this is not a trivial problem due to the heterogeneity of the ligand entries in these files. This has been made easier because there is now a database of all ligands present in the PDB which is freely available from Brookhaven. This could be used in concert with the LIGPLOT (Wallace *et al.*, 1995) program to identify the atoms that contact the inhibitors in the enzyme active sites.

In general, however, it is necessary to define the template by manual techniques, usually by scanning the available literature describing the structure of the enzyme under investigation. Once the relevant catalytic atoms and residues have been identified, further complications can arise. To illustrate this, the derivation of consensus templates for two enzymes, ribonuclease and lysozyme is described.

8.3 Ribonuclease

Ribonucleases are found in both prokaryotes and eukaryotes (Beintema, 1990); they catalyse the hydrolysis of phosphodiester bonds in RNA chains. Structurally, the best understood ribonucleases are bovine pancreatic ribonuclease (RNase A E.C.3.4.27.5) and ribonuclease T_1 (RNase T_1 E.C.3.4.27.3) from the fungus *Aspergillus oryzae*. The crystal structures of ribonuclease H and barnase have also been solved.

8.3.1 Ribonuclease A

This is a pyrimidine–specific ribonuclease of 124 amino acids and is a member of a large superfamily of homologous bovine RNases (Beintema *et al.*, 1988). The superfamily has been divided into two classes: secretory ribonucleases which are found in the pancreas, and non–secretory ribonucleases which are found in the liver, lung, spleen and leucocytes. Its function is the degradation of microbial RNA. In other vertebrates, including man, RNase is found at very low levels and its function is unclear but may involve the breakdown of dietary RNA.

RNase A is a monomer with an $\alpha+\beta$ fold and four intra–chain disulphides. Ribonuclease S (RNase S) is a product of the cleavage of RNase A by subtilisin between residues 20 and 21. Although RNase S is less stable than the parent enzyme, its enzymatic activity and general tertiary fold (e.g. 1*rbc*, Varadarajan & Richards, 1992) is the same. There is also another isoform, RNase B, which has covalently attached carbohydrate at Asn-X–Thr/Ser attachment sites on the surface of the molecule.

Specificity and catalytic mechanism

RNase A is specific for the pyrimidines uridine and cytidine; this specificity is achieved by hydrogen bonding from the backbone NH and –OH atoms of the sequentially conserved residue Thr 45.

Chemical modification studies (Crestfield *et al.*, 1963; Hirs *et al*, 1965) and analysis of the pH dependence of the enzymatic activity (Findlay *et al.*, 1962) have shown that the residues His 12, His 119 and Lys 41 are involved in the catalytic mechanism (Blackburn & Moore, 1982). The mechanism is a two step process, firstly the trans–esterification occurs whereby the P–O5' bond at the 3' end of a pyrimidine is cleaved and a 2',3' cyclic nucleotide is formed which is then hydrolysed. Figure 8.1 is a 3D representation of the inhibitor deoxycytidyl–3',5'–

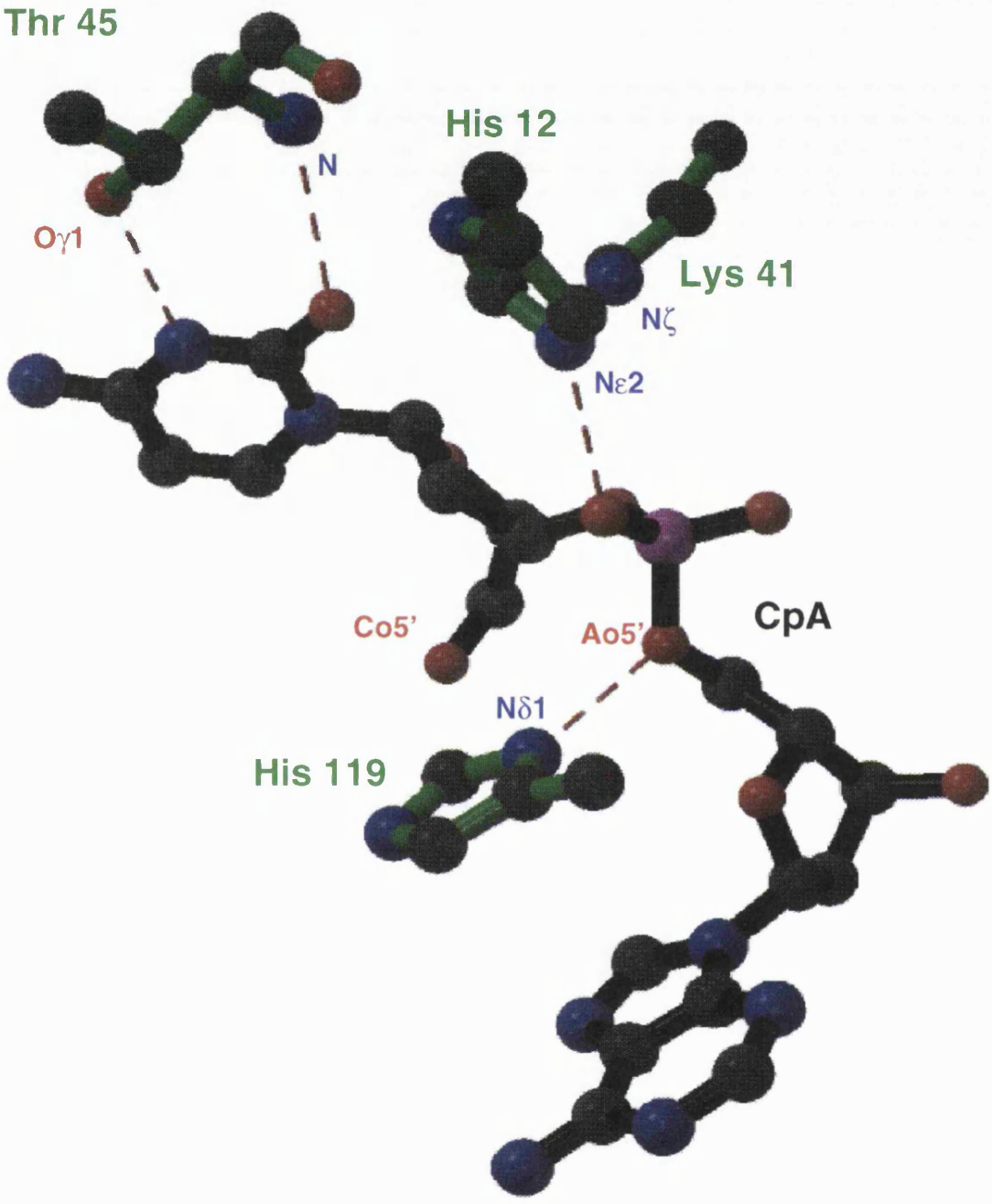


Figure 8.1: A 3D representation of the active site of ribonuclease A complexed with D(CPA) (Zegers *et al.*, 1994). The catalytic residues are His 119, His 12 and Lys 41.

deoxyadenosine, d($c_p a$) bound to ribonuclease A *1rpg* (Zegers *et al.*, 1994). The His 12 interacts with the 2'–oxygen (equivalent to CO5' in the inhibitor in Figure 8.1) and Lys 41 is in the vicinity. On the basis of NMR and crystallographic data, Roberts *et al.* (1969) proposed a model in which His 12 abstracts a proton from the 2'–oxygen which becomes a nucleophile and attacks the phosphorus (P), forming a penta–coordinated transition state. The 5' (AO5') leaving group is protonated by His 119, which acts as a general acid. Lys 41 forms a salt bridge with the charged oxygens on the phosphorus and thereby stabilises the transition state.

Borkakoti *et al.* (1982) noticed that in a phosphate complex of RNase A there are two distinct conformations of His 119 (A and B) . These two conformations are related by a 180^o rotation about the His 119 C^β–C^γ ($\chi 2$) bond. Santoro *et al.* (1993) described the same two RNase A conformations of His 119 in NMR solution studies of RNase A. Depending on the inhibitor complexes of RNase A, His 119 will adopt the A (e.g. O8–2'O–CMP, Borkakoti, 1983) or the B conformation (e.g. O3–2'–CMP, Borkakoti, 1983). It appears that in the absence of substrate His 119 is relatively mobile and can adopt either the A or B conformation. Borkakoti *et al.* (1982) suggested that the two conformations reflected the two different reactions i.e. the transesterification and hydrolysis reactions. deMel *et al.* (1992) suggest that His 119 is active in position B and inactive in position A, whereas Zegers *et al.*, 1994 suggest the A conformation is active and B inactive.

The consensus templates

Since there is ambiguity as to the catalytically active conformation of the His 119 residue, two consensus templates have been created describing both the A and B conformations. The atoms chosen for the consensus templates were the sidechain atoms of His 12, His 119 $N^{\delta 1}$ and Lys 41 N^ζ. The 'seed' template was taken

Template conformer A coordinates					
Residue	Number	Atom	x	y	z
Lys	41	N^{ζ}	5.3	-1.1	-2.8
His	119	N^{δ_1}	6.2	3.2	3.8
Template conformer B coordinates					
Residue	Number	Atom	x	y	z
Lys	41	N^{ζ}	5.1	-1.4	-2.9
His	119	N^{δ_1}	5.8	5.8	2.0
Reference frame atoms					
Residue	Number	Atom	x	y	z
His	12	C^{γ}	0.0	0.0	0.0
His	12	N^{δ_1}	0.8	-1.1	0.0
His	12	C^{δ_2}	0.8	1.1	0.0
His	12	C^{ϵ_1}	2.1	-0.7	0.0
His	12	N^{ϵ_2}	2.1	0.6	0.0

Table 8.3: Coordinates of the consensus templates that describe the two conformers, A and B, for the active site of ribonuclease A.

from 3*rn*3 (Borkakoti *et al.*, 1982) which has coordinates for both the A and B His 119 conformations and the dataset used was all RNase A structures in the January 1995 PDB. The 3*rn*3 PDB structure has coordinates for both the A and B conformations of the His 119 residue; therefore two consensus templates have been constructed, one for each conformation and Table 8.3 gives their resultant coordinates. Figure 8.2 is a 3D representation of the distribution of the His 119 N^{δ_1} atoms relative to His 12 for all structures in the ribonuclease A and ribonuclease S dataset. There are two distinct clusters of atoms (in blue) representing the A and B conformations. Also shown is the distribution the Lys N^{ζ} atoms (in red). Table 8.4 lists the PDB structures responsible for these clusters; there are no structures, other than PDB structures with alternate conformations, that have a hit in both groups A and B. To illustrate the distinct clustering of the A and B forms, the mean consensus template of the B conformation was used to

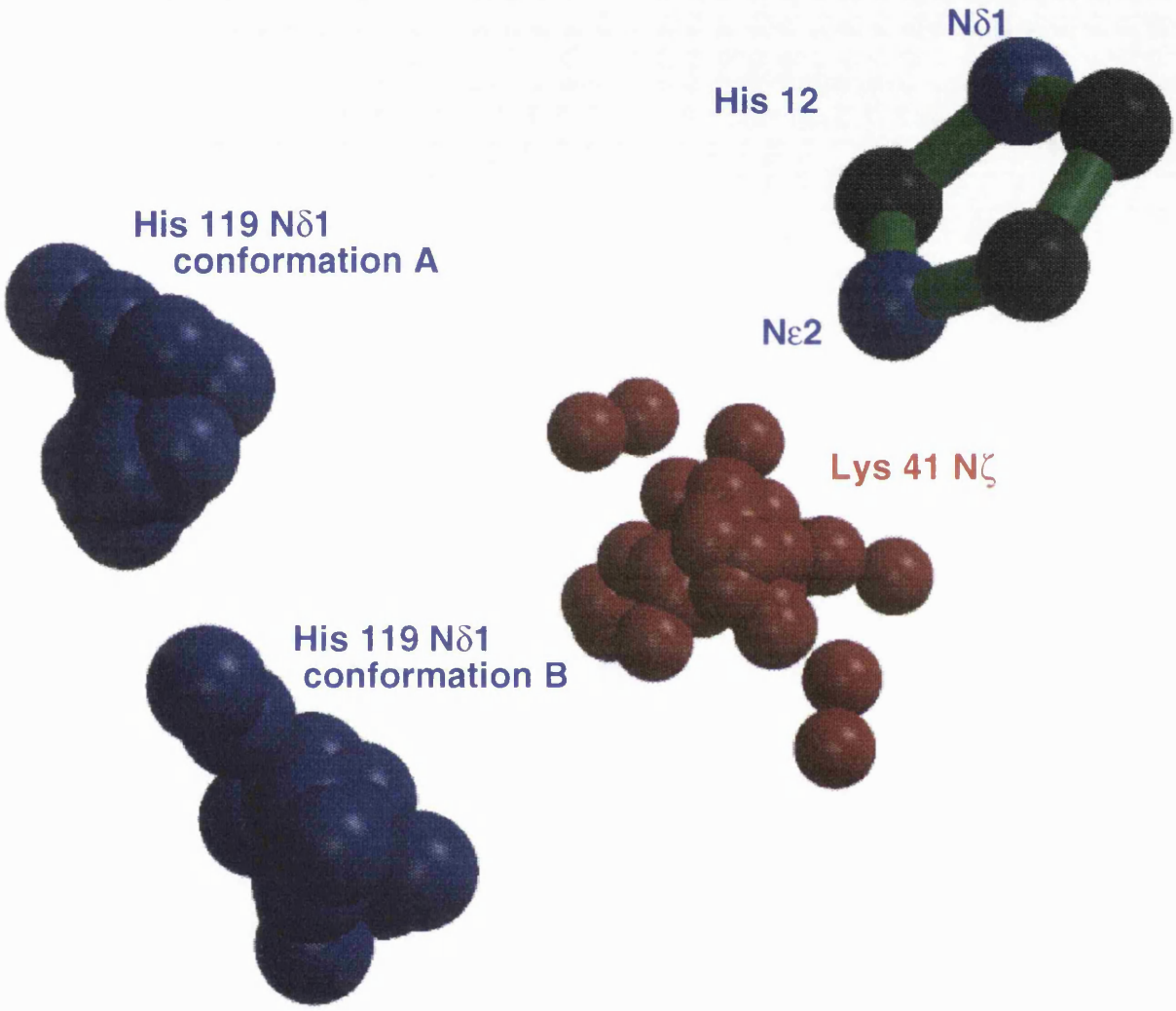


Figure 8.2: A 3D representation of the distribution of the His 119 $N^{\delta 1}$ active site atom conformations A and B for all the RNase A and RNase S structures in the PDB. Also shown is the sidechain of His 12 and the distribution of the Lys 41 N^{ζ} atoms.

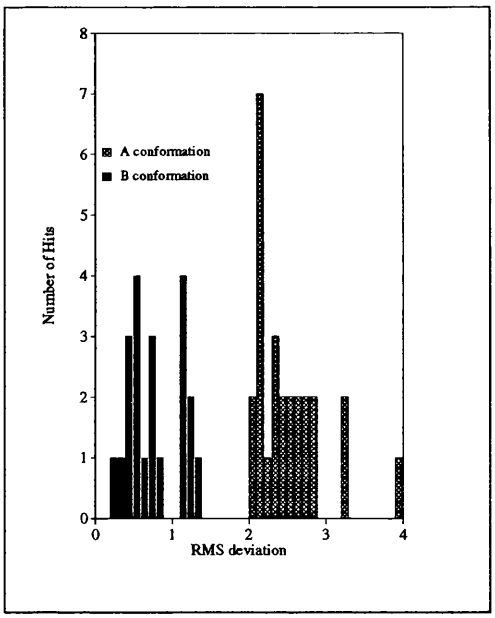


Figure 8.3: A histogram of the number of hits against *rms* distance from the B conformation of the RNase A consensus template. It shows that the A and B conformations of the His 119 residues are in distinct positions.

search through the RNase A dataset using a 5.0Å distance cut–off. The histogram in Figure 8.3 shows that the A and B conformations cluster at distinct distances from the B conformation consensus template.

Template search through the PDB

The 95% by sequence non–identical protein dataset was used to search for other proteins with residues and atoms in a similar conformation to the RNase A consensus template. This was in fact done twice, once for each of the template conformations A and B; in both cases an *rms* distance cut–off of 5Å was used.

Figure 8.4 is a histogram of the number of hits against *rms* deviation from the A conformation of the RNase A consensus template. All the hits located, other than the RNase structure in this dataset, have *rms* deviations greater than 2.5Å; they can therefore be discounted as having RNase A catalytic activity.

Figure 8.5 is a histogram of the number of hits against *rms* deviation when

TEMPLATE CONFORMER A: ribonuclease E.C.3.1.27.5						
ribonuclease A						
1bsrB 0.45	**3rn3** 1.16	1ras 1.50	1rat 0.90	2rat 0.74	3rat 0.92	4rat 0.79
5rat 0.31	6rat 0.43	7rat 0.48	8rat 0.44	9rat 0.41	1rbn 0.62	1rcn E 0.71
1rnc 0.35	1rnd 0.46	1rar 1.03	1rob 0.49	1rpg 0.33	1rph 0.53	1rtb 1.56
5rsa 0.34	6rsa 0.53	7rsa 0.48	1rtb 1.56			

TEMPLATE CONFORMER B: ribonuclease E.C.3.1.27.5						
ribonuclease A						
1bsr B 0.64	1rbn 0.97	**3rn3** 1.05	1rpf 1.29	1rph 0.78	9rsa B 0.71	1srn A 0.65
3srn A 0.49	4srn A 0.32	1ssa A 0.78	1ssb A 0.50			
ribonuclease S						
1rbc S 1.00	1rbd S 0.30	1rbe S 0.74	1rbf S 1.50	1rbg S 0.49	1rbh S 0.44	1rbi S 0.40
2rln S 0.29	1rnu 1.36	1rnv 0.54	2rns 1.38			

Table 8.4: A summary of the ribonuclease PDB structures and their *rms* deviations from their respective consensus templates, that adopt either the A or B conformation of their active site His 119 residue. Those PDB codes in bold have coordinates describing both conformations.

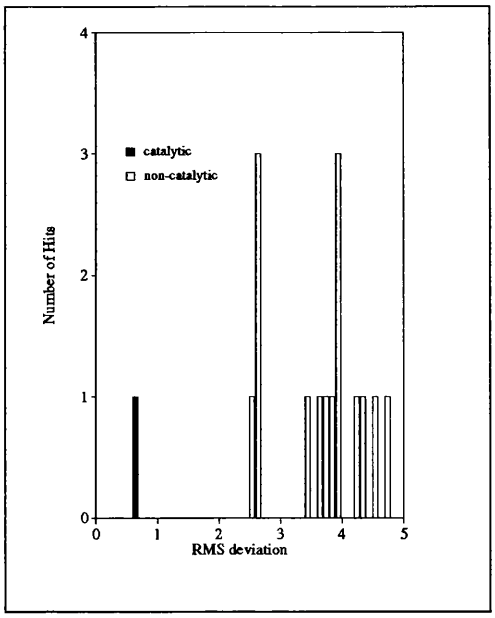


Figure 8.4: A histogram of the number of hits against *rms* distance when the A conformation of the RNase A consensus template was searched through the 95% by sequence non–identical protein dataset.

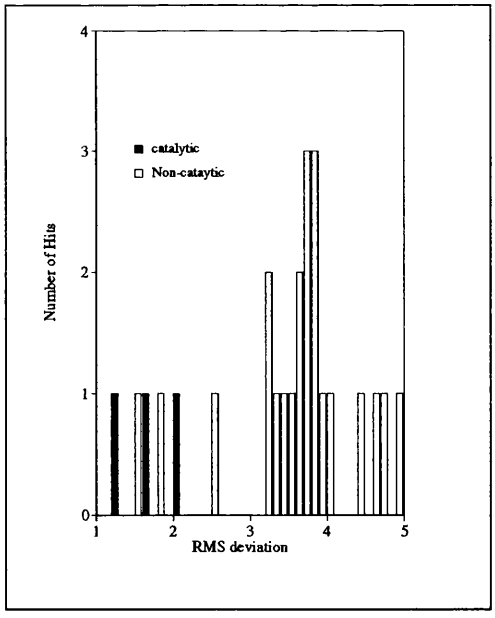


Figure 8.5: A histogram of the number of hits against *rms* distance when the B conformation of the RNase A consensus template was searched through the 95% by sequence non–identical protein dataset.

the B conformation consensus template was used to search through the 95% non–identical PDB dataset. There are several hits from proteins other than RNase A, however most of these are of relatively high *rms* distance, and can be discounted as potentially catalytic. There are however, two hits with *rms* distance comparable to the RNase A catalytic atoms (black bars). These are in fact from the residues His 285, His 321 and Lys 191 from the A and B chains of the same enzyme RUBISCO, ribulose–1,5–bisphosphate carboxylase E.C.4.1.1.39 (5*rub*, Schneider *et al.*, 1990) with an *rms* distance of 1.51Å and 1.86Å respectively.

RUBISCO, which is the most abundant naturally occurring enzyme catalyses the initial steps of two opposing metabolic pathways in plants: firstly, the initial step in photosynthetic carbon dioxide fixation, the carboxylation of ribulose–1,5–bisphosphate and secondly, oxygenation of ribulose–1,5–bisphosphate, the first step in photorespiration. RUBISCO contains eight large chains (L) and eight small chains (S) forming an L_8S_8 complex. The large L subunit consists of two domains; the N–terminal domain is folded into a central β–sheet with helices on each side and the C–terminal domain consists of a β/α barrel. The active site is located at the carboxy side of the strands in the barrel, with residues from the N–terminal domain coming into close proximity. Figure 8.6 is a 3D representation of the active site of RUBISCO from 9*rub* (Lundqvist & Schneider, 1991) with substrate ribulose–1,5–bisphosphate. Also shown is the active site Mg that is coordinated to Asp 193 (yellow bonds) and the substrate. The residues located, His 285, His 321 and Lys 191 (red bonds) are found in the large L chain and are also part of the active site. In addition, they are all conserved in the RUBISCO sequence (Schneider *et al.*, 1990). The Lys 191 is the site of carbamylation during activation and His 321 is involved in binding the phosphate group of the substrate (Lundqvist *et al.*, 1989). There is no clear functional role assigned to His 285.

When compared to ribonuclease A (Figure 8.1), the position of the active site

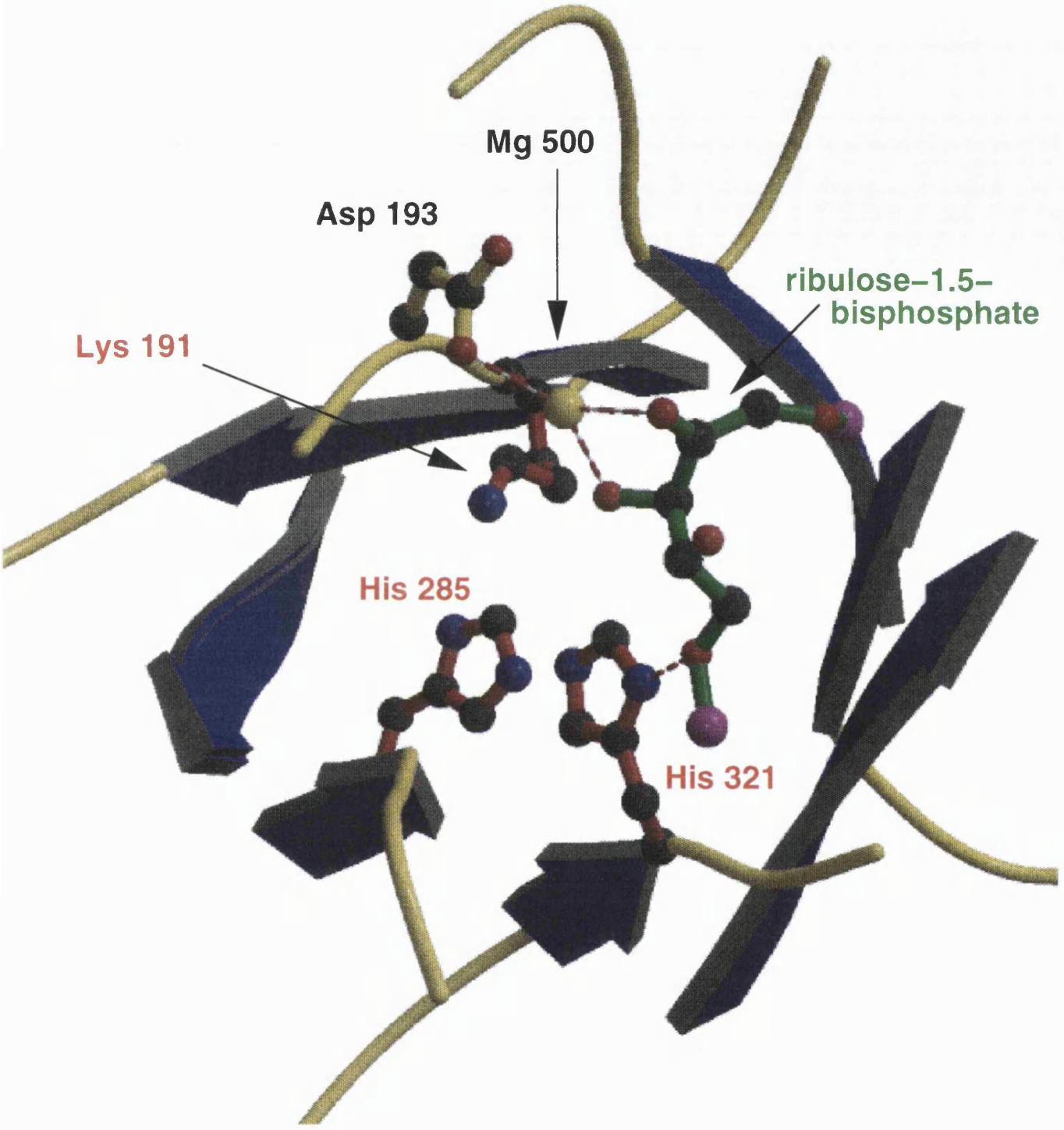


Figure 8.6: A 3D representation of the active site of RUBISCO (Lundqvist & Schneider, 1991) showing the 3 residues His 285, His 321 and Lys 191 (red bonds) that have the same conformation as the active site residues of ribonuclease A.

residues with respect to the substrate is different. The equivalent residues also have different roles; His 12 and His 119 in ribonuclease A are both acid/base catalysts and Lys 41 stabilises the transition state. This is interesting as it shows that catalytic residues can have different roles depending on their structural and chemical environments.

8.3.2 Ribonuclease T_1

RNase T_1 is isolated from the fungus *Aspergillus oryzae*. The enzyme has 2 isoforms containing Lys or Gln at position 25 of the polypeptide chain, denoted Lys^{25}–RNase T_1 and Gln^{25}–RNase T_1.

Specificity and catalytic mechanism

RNase T_1 is specific for the purine nucleotide guanosine (as opposed to pyrimidines in RNase A) and is strictly limited to hydrolysis at 3'–phosphate groups in RNA. The reason for this specificity is not fully understood.

The reaction mechanism is analogous to that of RNase A; firstly, the transesterification of RNA to yield oligonucleotides with terminal guanosine 2',3'–cyclic bisphosphate, and secondly, the hydrolysis of the 2',3'–cyclic bisphosphate to yield guanosine 3'–monophosphate. The X–ray crystal structure of RNase T_1 has been solved to 1.8Å resolution by Koepke *et al.* (1989) and a 3D representation of its active site complexed with the inhibitor guanylyl–2'5'–guanosine is shown in Figure 8.7. The catalytic residues shown are His 40/Glu 58, which takes a proton from H_2O in the first step and His 92 which donates a proton to the leaving group and activates the water molecule used in hydrolysis. The guanosine specific recognition occurs in the loop Tyr 42–Asn 43–Asn 44–Tyr 45–Glu 48.

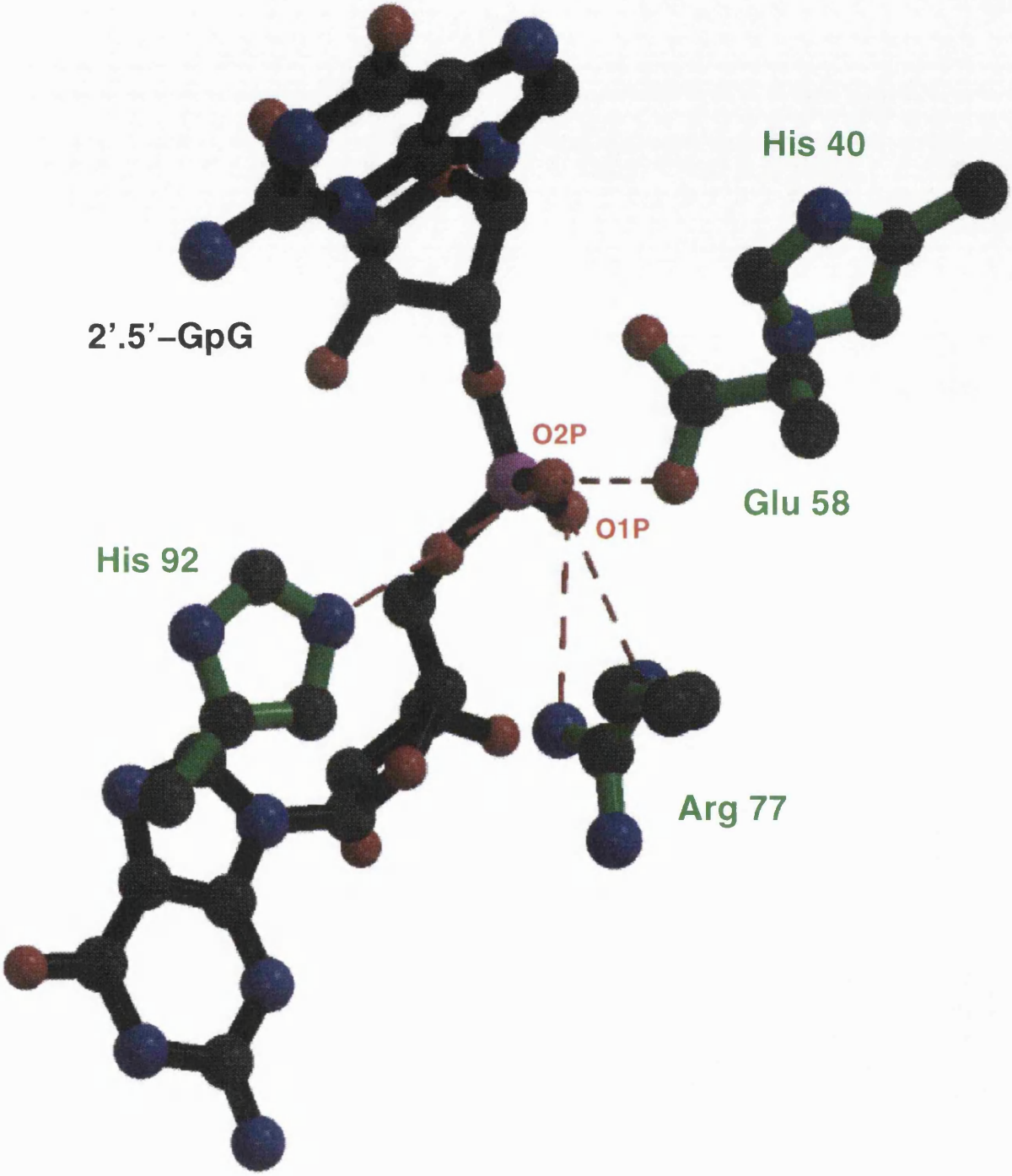


Figure 8.7: A 3D representation of the active site residues of RNase T_1 (Koepke *et al.*, 1989) with the inhibitor guanylyl–2'5'–guanosine.

1fus 1.08	1fut 1.28	1rga 1.28	1rgcB 1.40	1rgcA 1.92	1rgk 1.78	1rgl 0.66	1rls 1.71	1rms 1.56
1rn1A 1.53	1rn1B 1.53	1rn1C 0.62	1rnt 0.99	2rnt 0.47	3rnt 0.76	6rnt 0.41	7rnt 0.90	8rnt 1.78
9rnt 1.77	1trpA 0.85	1trpB 1.00	1trqA 0.87	1trqB 0.83	**2aae**	**5rnt**		

Table 8.5: A summary of the ribonuclease T_1 PDB structures and their *rms* deviations from the ribonuclease T_1 consensus template. Those PDB codes in bold are missed by the RNase T_1 template using a 3.0Å distance cut–off.

Residue	Res. Number	Atom	x	y	z
His	92	N^{ϵ}_2	9.7	2.5	-2.1
Glu	58	C^{δ}	3.4	2.8	-2.6
Glu	58	O^{ϵ}_2	3.6	3.3	-1.5
Glu	58	O^{ϵ_1}	4.3	2.3	-3.2
His	40	C^{β}	-1.5	-0.1	0.0
His	40	C^{γ}	0.0	0.0	0.0
His	40	N^{δ_1}	0.8	-1.1	0.0
His	40	C^{δ_2}	0.8	1.1	0.0
His	40	C^{ϵ_1}	2.1	-0.7	0.0
His	40	N^{ϵ_2}	2.1	0.6	0.0

Table 8.6: The coordinates of the functional consensus templates of ribonuclease RNase T_1

The consensus template

The atoms used to generate a consensus template were the sidechain of His 40, His 92 N^{ϵ_2} and Glu 58 C^{δ}, O^{ϵ_1}, O^{ϵ_2} taken from the RNase T_1 X–ray crystal structure 1*rn*1 (Arni *et al.*, 1992); the distance cut–off was set at 3.0Å. Table 8.5 gives the dataset of RNase T_1 PDB codes and their *rms* deviations from the resultant consensus template whose coordinates are given in Table 8.6. There are two structures in the ribonuclease T_1 dataset, 2*aae* (Zegers *et al.*, 1992) and 5*rnt* (Lenz *et al.*, 1991), whose active site residues are not identified by the consensus template. 2*aae* has its His 40 mutated to a lysine, such a mutant will not be found

using this approach. *5rnt*'s structure is refined to the relatively low resolution of 3.2Å and has the inhibitor guanosine–3',5'–bisphosphate bound to its active site.

Template search through the PDB

There are two potential ribonuclease active sites found when the ribonuclease T_1 was matched against the 95% non–identical by sequence dataset of PDB structures. These are narbonin from *Vicia narbonesis*, *1nar* (Hennig *et al.*, 1992; Hennig *et al.*, 1995) and hemerythrin *2hmq* (Holmes & Stenkamp, 1991).

Narbonin is a storage globulin found in all legume plant seeds. It is a member of a family of plant seed proteins of different size and structural organisation. The protein has a TIM–barrel like fold with an eight stranded parallel β–barrel surrounded by a ring of seven α–helices. The majority of TIM–barrel proteins in the PDB are enzymes; narbonin is an exception because there has been no enzyme activity found. Divergent evolution may have led narbonin to lose its enzymatic activity but retain the TIM–barrel fold.

Figure 8.8 is a 3D representation of the structure of narbonin, showing the location of ribonuclease T_1 like catalytic residues of His 133, Glu 132 and His 234; these have an *rms* deviation of 1.9Å from the ribonuclease T_1 consensus template. The residues are at the C–terminal of the β–barrel and this is the position of the active site in other TIM–barrel proteins. In fact, Hennig *et al.*, 1995 have implicated Glu 132 as part of a salt bridge complex that protects the potential ligand binding site, yet salt bridges are found in other structurally and functionally important areas of TIM–barrel enzymes.

When BLAST was used to search of the SWISS–PROT sequence database there were no narbonin sequences located, however Table 8.7 gives a list of narbonin like sequences present in the OWL (Bleasby *et al.*, 1994) database. It also gives the equivalent residues to Glu 132, His 133 and His 234 in *1nar*. There is little

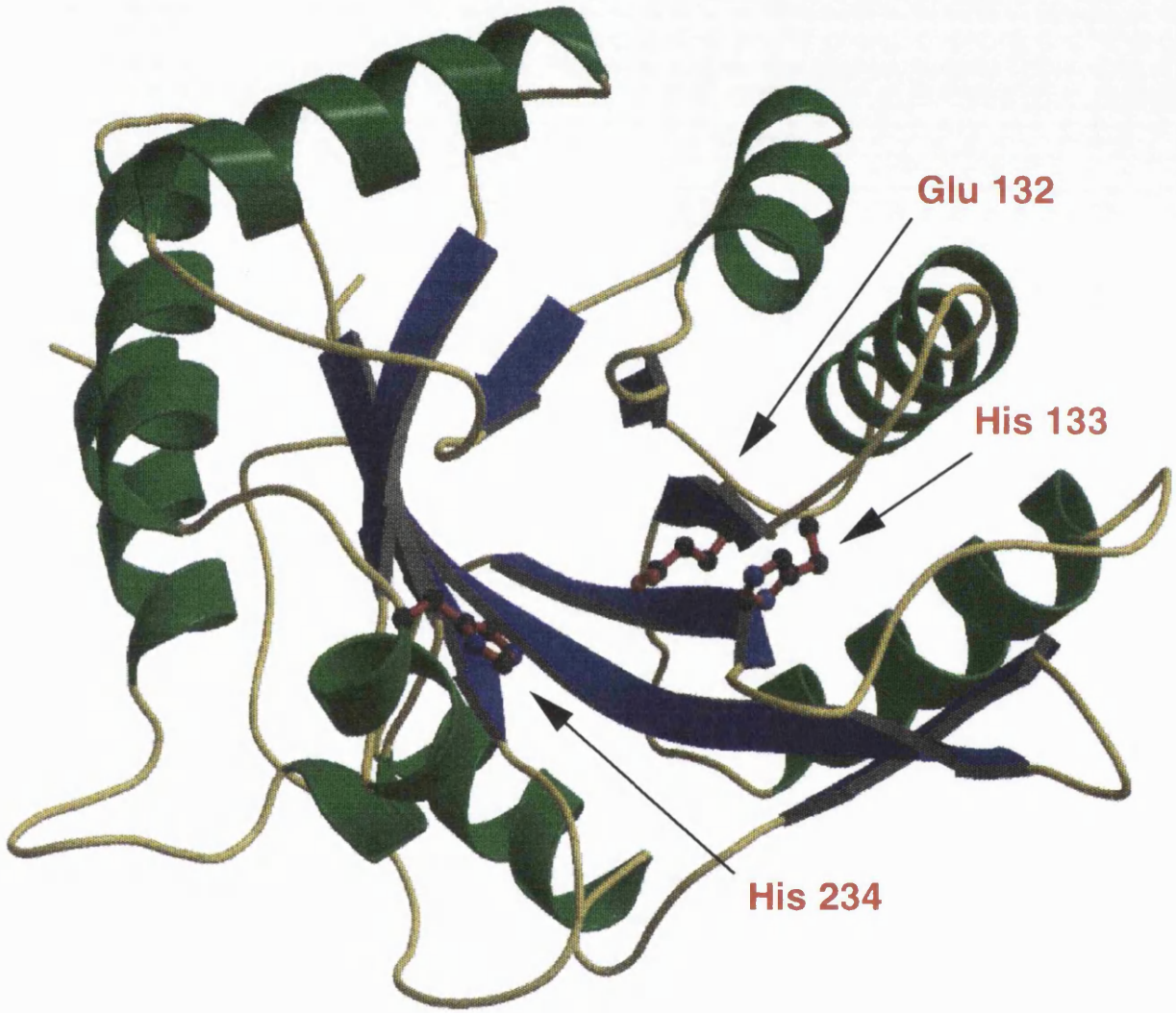


Figure 8.8: A 3D representation of the plant seed protein narbonin 1*nar* (Hennig *et al.*, 1992; Hennig *et al.*, 1995). Also shown are the Glu 132, His 133 and His 234 residues which adopt a similar conformation to the catalytic residues found in ribonuclease T_1.

sequence identification code	species	Glu 132	His 133	His 234
S49848	probable narbonin - jack bean	E	H	N
S49878	probable narbonin - soybean	E	H	A
S49897	hypothetical narbonin-like 2S protein (clone pVFNA4) - fava bean	E	Y	N
S49880	hypothetical narbonin-like 2S protein - fava bean	G	N	N
S44031	narbonin (clone pNaG2) - Vicia narbonensis	E	H	T
S44032	narbonin (clone pNaN21/pNaC18) - Vicia narbonensis	E	H	T
S44033	narbonin - Vicia pannonica	E	H	T
S50159	narbonin - Vicia pannonica	E	H	I
VFNDSA1	V.faba mRNA for nodulin homologous to narbonin. - fava bean.	G	N	N
VFNDSA2	Vicia faba mRNA for nodulin homologous to narbonin. - fava bean.	E	H	H

Table 8.7: A list of narbonin sequences found in the OWL database (Bleasby *et al.*, 1994) with the residues found at the equivalent positions of the Glu 132, His 133 and His 234 residues in *1nar* (Hennig *et al.*, 1992; Hennig *et al.*, 1995).

conservation of these residues in these sequences and this, along with the relatively high *rms* deviation of the narbonin residues from the ribonuclease T_1 consensus template, indicate that the potential catalytic residues probably do not have any functional importance. Of course, the possibility exists that narbonin is related by divergent evolution to an enzyme of ribonuclease activity.

Hemerythrin and myohemerythrin are oxygen–binding proteins found in marine invertebrate phyla. Hemerythrin and myohemerythrin subunits consist of four parallel α–helices and this provides the amino acid sidechain ligands for the binuclear iron oxygen bridged metal centre. The structure of hemerythrin has been solved to 1.66Å resolution (*2hmq*) by Holmes & Stenkamp, 1991. Figure 8.9 is a 3D representation of hemerythrin showing the potential catalytic residues and the two hemerythrin bound irons; its *rms* distance from the RNase T_1 is 1.66Å. In fact, all three residues are involved in coordinating the iron atoms; Glu 58 and His 73 to Fe1 and His 25 to Fe2.

A search of the SWISS–PROT (March 1995) database with the sequence of

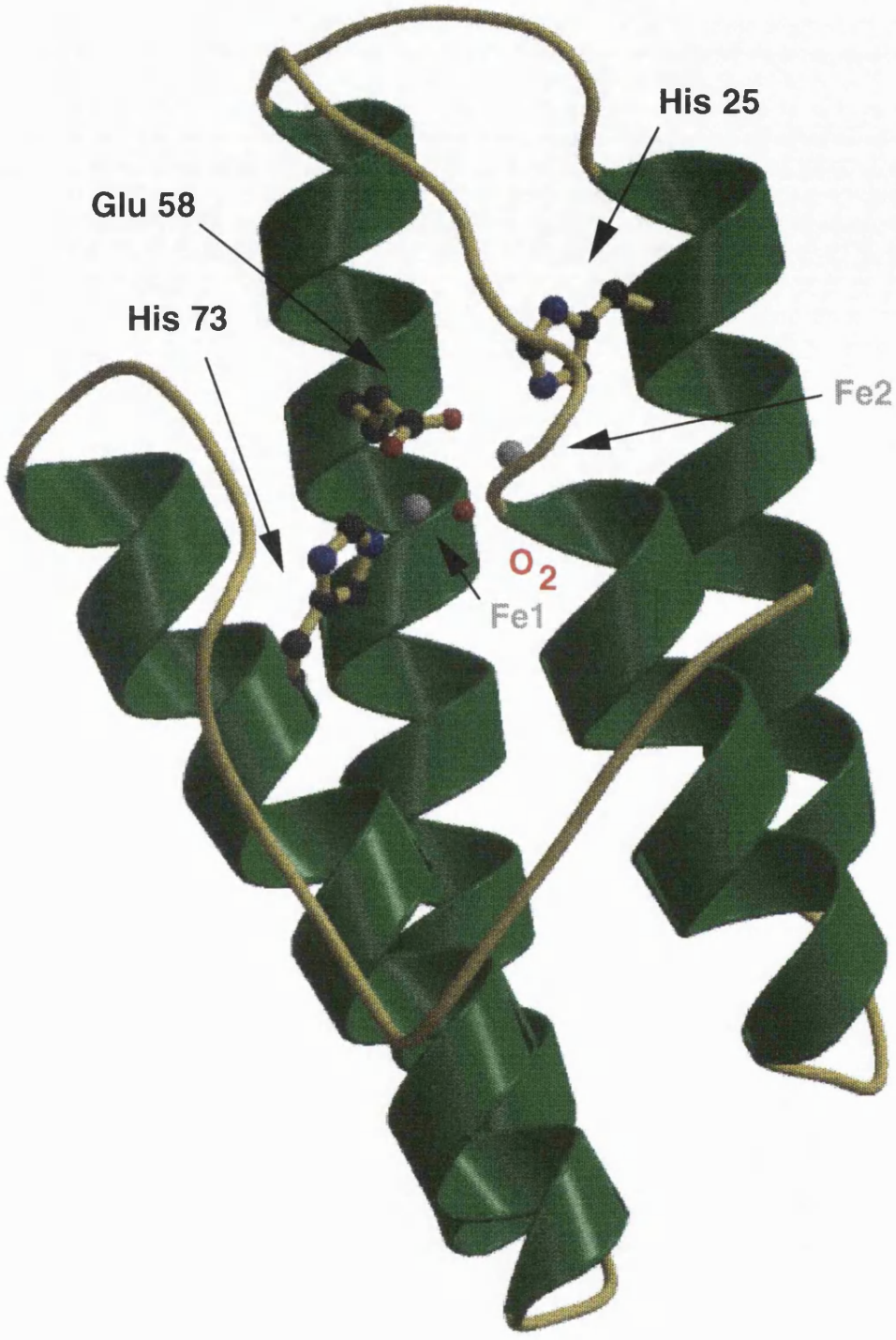


Figure 8.9: A 3D representation of the oxygen binding protein hemerythrin. Both the hemerythrin bound iron and the His 25, Glu 58 and His 73 residues which adopt a similar conformation to the catalytic residues found in ribonuclease T_1 are shown. Also shown is the bound oxygen.

2hmq gives 10 hemerythrin/myohemerythrin sequences with sequence identity ranging from 40–90% with *2hmq*. The His, Glu, His residues are conserved in all cases except one where the Glu is mutated to a Gln.

The role of His 25, Glu 58 and His 73 in hemerythrin is to coordinate the functional iron atoms; the ribonuclease like architecture of these conserved residues has probably occurred by chance and is not some form of divergent evolution.

8.3.3 Comparison of ribonuclease A and T_1 active sites

Having created two templates for RNase A, T_1 it is interesting to compare them. Figure 8.10 is a 3D representation of the two consensus templates and the atoms have been superimposed according to their proposed role in the catalytic mechanism. The general bases, His 12 and Glu 58; the electrostatic stabilising groups Lys 41 and His 40 and the general acid catalysts His 12 and His 92 of RNase A and RNase T_1 respectively all superimpose indicating convergent evolution.

8.3.4 Ribonuclease H

This is a unique ribonuclease that is specific for the RNA strand of a DNA/RNA complex. The crystal structure of RNase H from *Escherichia coli* has been determined to 1.48Å resolution by Katayanagi *et al.* (1992); it has an α/β like tertiary fold. Though its function is yet to be elucidated, RNase H activity has been found in prokaryotes and eukaryotes.

Site directed mutagenesis experiments have implicated various residues in the binding of the substrate. However, as yet, no clear picture of the catalytic residues and mechanism is available.

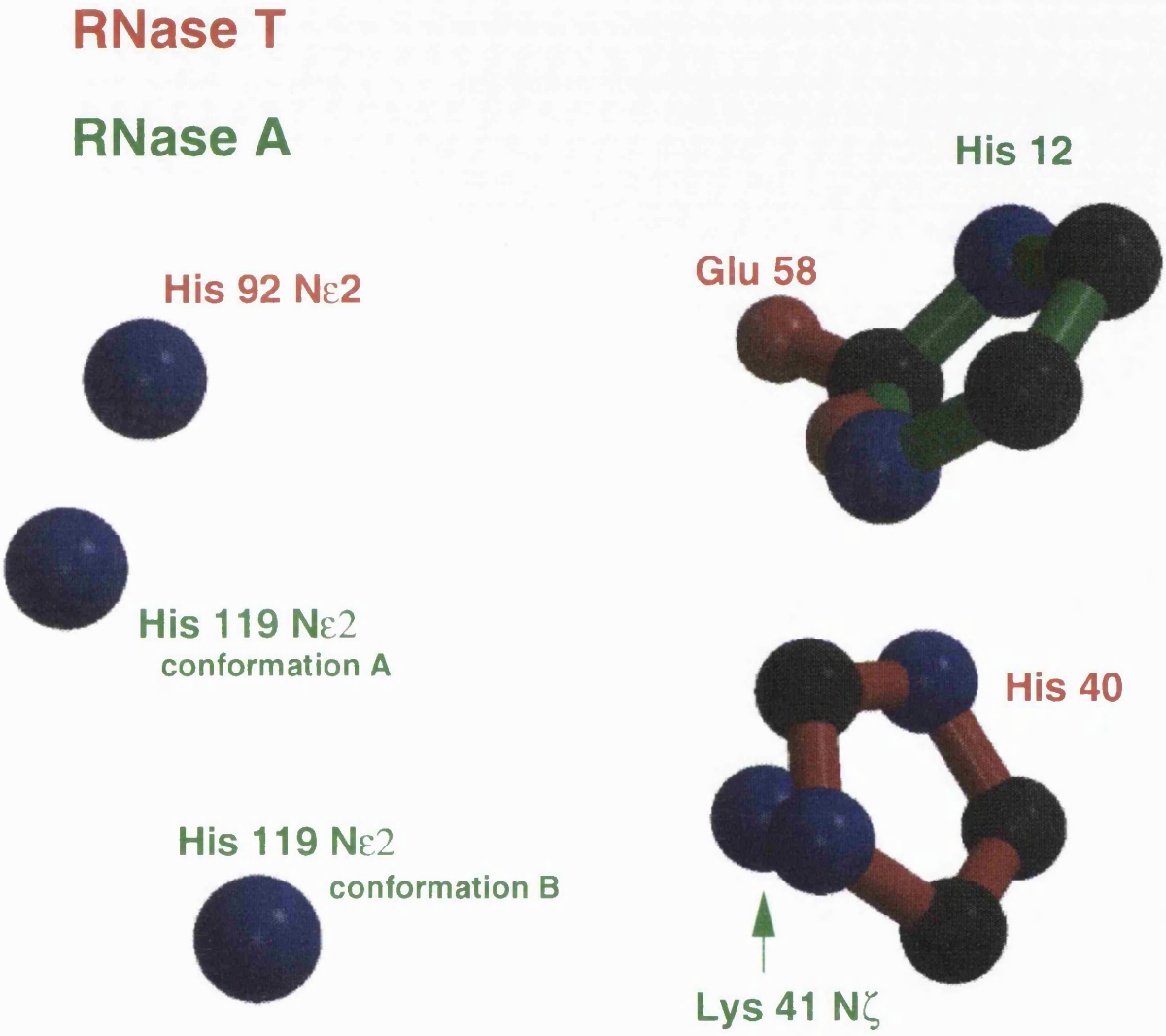


Figure 8.10: A diagram showing the relative conformations of the consensus template atoms of RNase A, T_1. The residues and atoms of the templates have been superimposed according to their proposed chemical role in the catalytic mechanism.

8.3.5 Barnase

Barnase is an endonuclease produced and excreted by *Bacillus amyloliquefaciens*; it has an $\alpha+\beta$ tertiary fold. It is believed to have the same chemical mechanism as pancreatic RNase A: a 2',3' cyclic intermediate is formed by transesterification and then it is hydrolysed.

The consensus template

Two residues have been implicated in the catalytic mechanism, the acid/base catalyst Glu 73 (equivalent to Glu 58 in RNase T_1) for the transesterfication and His 102 for hydrolysis. These two residues were used to create the consensus template. In fact, it was necessary to create two templates, using the sidechain of His 102 and Glu 73 C^{δ}, $O^{\epsilon 1}$, $O^{\epsilon 2}$ with the seed templates from *1bse* (Buckle *et al.*, 1993) and *1ban* (Serrano *et al.*, 1992) and a distance cut–off of 3Å; the coordinates and *rms* deviations of each of the structures from their respective templates are given in Table 8.8 and Table 8.9. Figure 8.11 is a diagram illustrating the distirbution of the Glu 73 residue with respect to the His 102 residue for all the barnase structures in Table 8.9. There are two distinct clusters which occurs because structures in the *1bse* group have the natural barnase inhibitor barstar bound to the active site. This conformational change that barstar induces may also be its mechanism of inhibition.

Template search through the PDB

There were over 100 hits located when the barnase consensus templates were parsed against the representative structures in the PDB. This has occured because there are only two residues in the consensus templates, increasing the chance that the relatively common His and Glu residues will occur in these positions. In addition, these residues lie around 9Å apart in the active site and are bisected by

Residue	Res. Number	Atom	x	y	z
		1bse template			
Glu	73	C^{δ}	3.4	7.1	-5.0
Glu	73	O^{ϵ}_{2}	2.8	6.7	-4.1
Glu	73	O^{ϵ_1}	4.0	6.5	-5.5
		1ban template			
Glu	73	C^{δ}	2.0	8.7	4.0
Glu	73	O^{ϵ}_{2}	2.3	7.5	4.1
Glu	73	O^{ϵ_1}	-1.5	9.2	2.9
		His template residue			
His	40	C^{β}	-1.5	-0.1	-0.0
His	40	C^{γ}	0.0	0.0	0.0
His	40	N^{δ_1}	0.8	-1.1	0.0
His	40	C^{δ_2}	0.8	1.1	0.0
His	40	C^{ϵ_1}	2.1	-0.7	0.0
His	40	N^{ϵ_2}	2.1	0.6	0.0

Table 8.8: The coordinates of the consensus templates created for barnase using the seed coordinates of *1bse* (Buckle *et al.*, 1993) and *1bgs* (Guillet *et al.*, 1993).

1bse template						
1bgsA 0.71	1bgsB 0.90	1bgsC 0.58	1brsA 1.30	1brsB 1.09	1brsC 0.72	1onc 1.66
1ban template						
1banA 1.01	1baoA 0.94	1bnsA 0.69	1bsbA 0.62	1bscA 0.91	1bscB 1.14	1bseA 0.92

Table 8.9: The PDB codes for the barnase structures that are represented by the *1bse* and *1ban* templates.

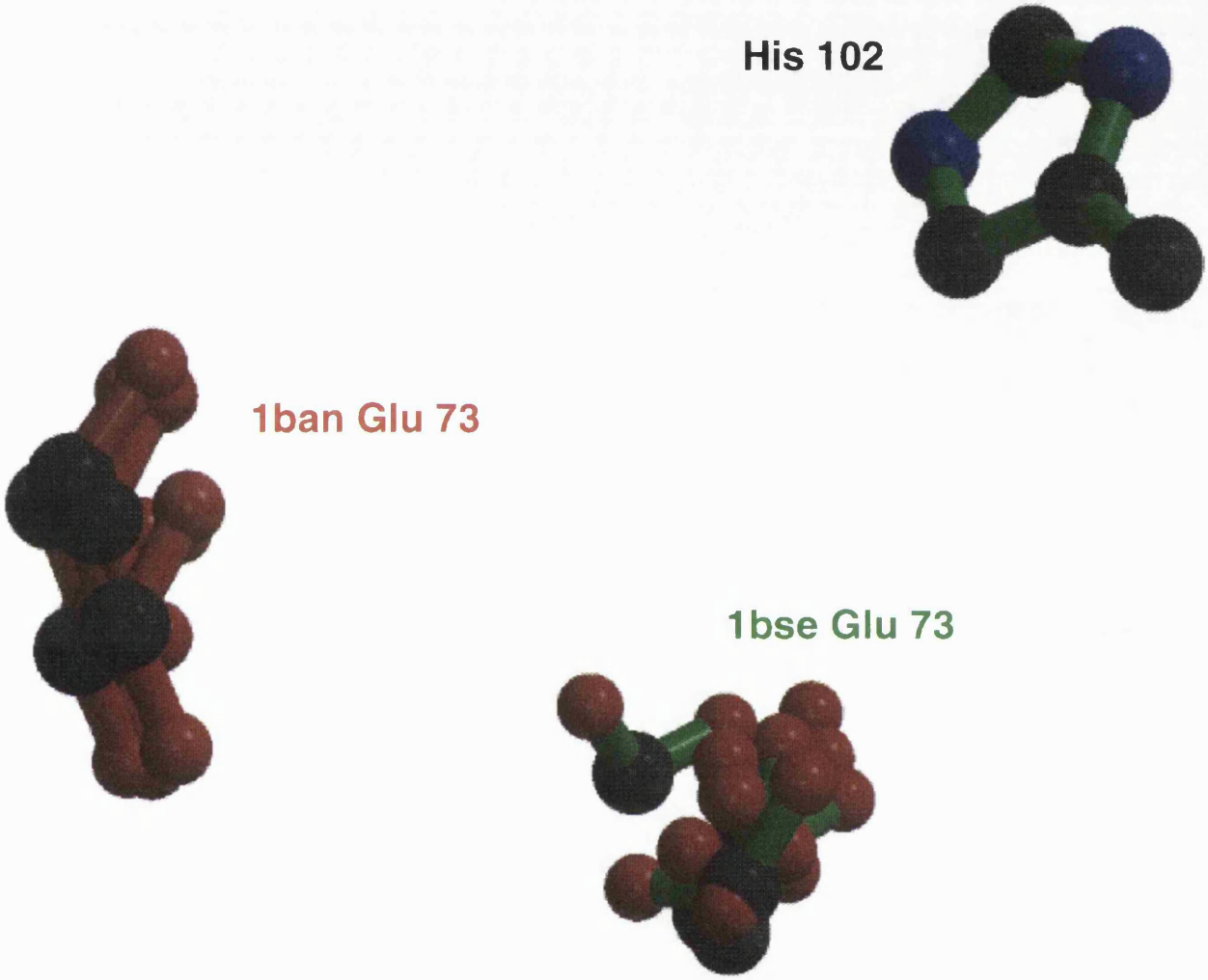


Figure 8.11: A diagram showing the distribution of all the barnase Glu 73 residues with respect to the His 102 residue for barnase. The 1*bse* Glu 73 residues (green bonds) all originate from structures that have the natural barnase inhibitor barstar bound to their active site. The 1*ban* Glu 73 structures have no inhibitors in their active sites.

the substrate. In non–catalytic regions of proteins, this space would be occupied by other residues.

8.4 Lysozyme

The second group of enzymes considered in this chapter are lysozymes. Lysozyme kills certain bacteria by cleaving the polysaccharide component of their cell wall (Imoto *et al.*, 1972). This polysaccharide is made of two components: N–acetylglucosamine (NAG) and N–acetylmuramate (NAM). The PDB has lysozyme structures originating from both prokaryotes and eukaryotes and, though they have low sequence identity, they all have a similar $\alpha+\beta$ tertiary fold (Remington & Matthews, 1978) and may have evolved from a common precursor. They can, however, be divided into eukaryotic and prokaryotic groups on the basis of the geometry of their catalytic residues.

The crystal structure of chicken hen white lysozyme was first determined by Blake *et al* (1965) to 2Å resolution; it was then refined further over the following years. These data were used to aid identification of the catalytic residues and propose a catalytic mechanism for the hydrolysis of the substrate. There are six binding sites for the NAG–NAM polymer in the active site of the enzyme, identified by the letters A to F. The cleavage site lies between the D and E sites (Figure 8.12). There are two catalytic residues located here, Glu 35 and Asp 52 in mammalian and avian lysozyme and Glu 11 and Asp 20 in the prokaryotic lysozyme from bacteriophage T4. Figure 8.13 is a 3D representation of the active site of the bacteriophage T4 lysozyme 148*l* (Kuroki *et al.*, 1993) with a bound substrate cleaved from the cell wall of *E. coli*. The catalytic residues, about 7Å apart, are bisected by the substrate and the bond cleaved would be at the N2 atom. Glu 11 and Asp 20 are seen in the vicinity of this bond. Glu 11 is an

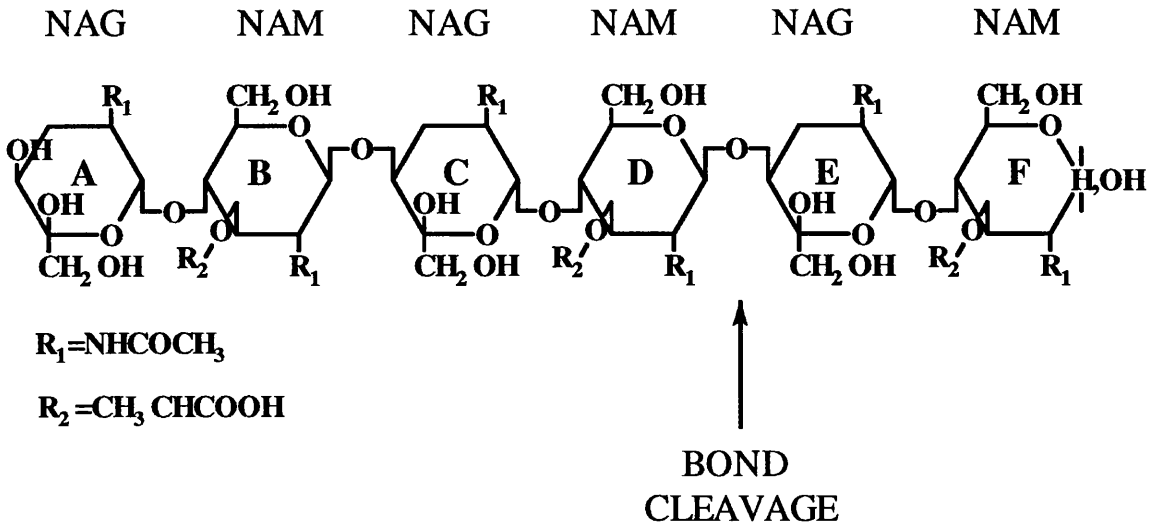


Figure 8.12: A diagram of the NAG–NAM substrate of lysozyme; 6 sugars fit into the binding sites A-F. Cleavage is between subsites D and E

acid/base catalyst, it donates a proton to the glycosidic oxygen atom linking the NAM and NAG in sites D and E, producing a carbonium ion intermediate. This intermediate is stabilised by Asp 20. In addition, the NAG sugar unit in site D is also distorted into a half chair form which promotes the formation of the carbonium ion intermediate. The sugars in sites E and F then diffuse away from the active site and a water molecule hydrolyses the carbonium ion intermediate.

8.4.1 Eukaryotic: Mammalian and avian lysozyme

There are 76 lysozyme structures in the PDB from several eukaryotic species. These can be divided into two main groups, avian and mammalian. Figure 8.14 is a 3D representation of the conformation of the Glu 35 and Asp 52 catalytic residues from all the members of these groups. The diagram has been divided into mammalian and avian lysozymes for clarity. In all cases the Glu residues have been superimposed so the relative conformation of the Asp sidechain can be compared. Although the conformation of the Asp sidechains are similar, there

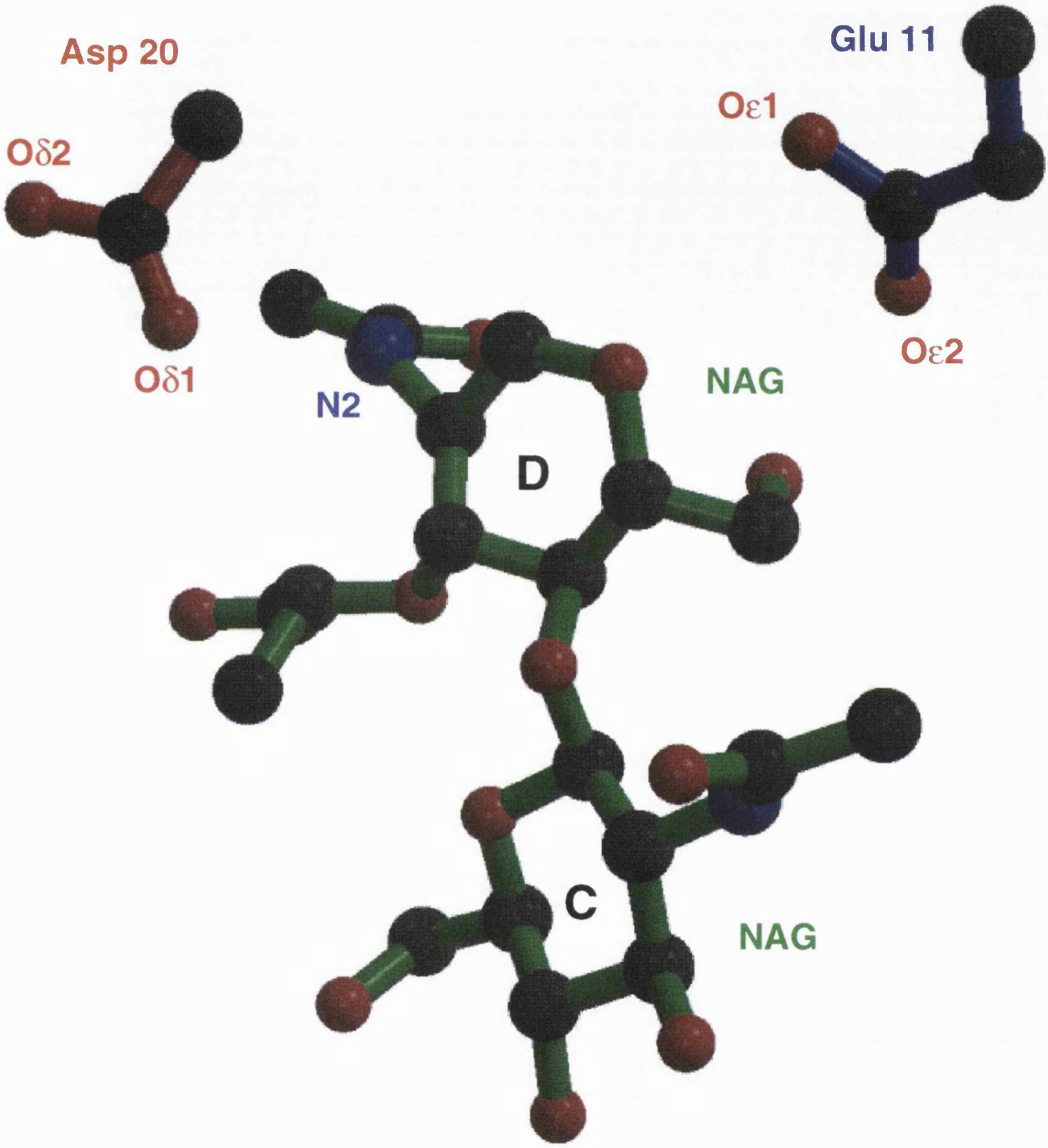


Figure 8.13: A 3D diagram showing the relative orientation of the catalytic residues Glu 11 and Asp 20 with respect to a substrate analogue bound to the active site of 148*l* (Kuroki *et al.*, 1993). The bond cleaved would be at the N2 atom shown in the diagram.

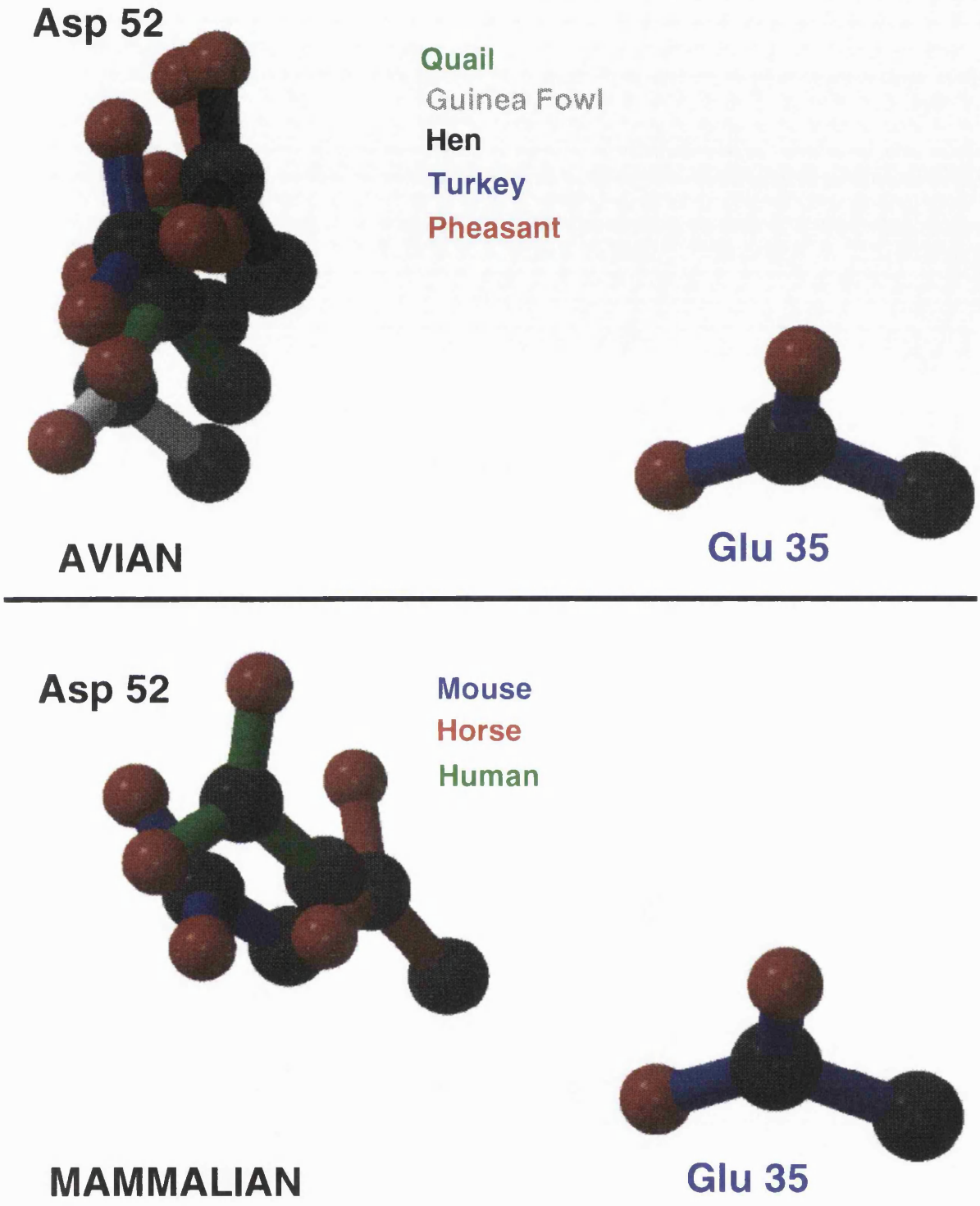


Figure 8.14: The relative conformations of the eukaryotic lysozyme catalytic residues Asp 52 and Glu 35. They are divided into two main groups, avian and mammalian. In all cases the Glu residues have been superimposed so the relative conformation of the Asp can be compared.

Residue	Residue Number	Atom	x	y	z
Asp	52	C^{γ}	4.0	-4.3	4.1
Asp	52	O^{δ_1}	4.4	-3.8	5.1
Asp	52	O^{δ_2}	4.4	-4.3	3.1
Glu	35	C^{β}	-2.2	-1.3	-0.3
Glu	35	C^{γ}	-1.5	0.1	-0.0
Glu	35	C^{δ}	0.0	0.0	0.0
Glu	35	O^{ϵ_1}	0.6	-1.0	0.0
Glu	35	O^{ϵ_2}	0.6	1.1	0.0

Table 8.10: Coordinates of the consensus template describing the active site of mammalian lysozymes present in the PDB.

is no obvious clustering of the functional Asp O^{δ} atoms. Without performing a comprehensive analysis of the individual PDB structures, it is not clear whether the structural differences occur because of an inherent difference in the active site structures of the lysozymes or the types of inhibitors bound to the active site.

Of the members of these groups hen and human lysozymes have 44 and 17 PDB structures respectively available. A consensus template was constructed for the hen lysosyme dataset using the seed template atoms of the Glu 35 sidechain and Asp 52 C^{γ}, O^{δ_1} and O^{δ_2} from 135*l* (Harata *et al.*, 1993). This hen consensus template was then used to search for the Glu 35 and Asp 52 catalytic residues in all the avian and mammalian PDB codes with a cut–off of 2.0Å, resulting in a consensus template which describes both the avian and mammalian lysozymes; the coordinates are given in Table 8.10. Table 8.11 is a summary of the results. Those PDB codes in bold type are not picked out by the consensus template, in fact 14 of these originate from hen lysozymes. These structure are either NMR structures (Smith *et al.*, 1992) or structures by Diamond *et al.*, 1975. A separate consensus template was constructed using the seed template of 6*lyt* (Diamond *et al.*, 1975) and Figure 8.15 shows that these catalytic residues cluster at a

Avian lysozymes														
Hen														
132l	1hel	1hem	1hen	1heo	1hep	1heq	1her	1hew	1lma	1lsa	1lsb	1lsc	1lsd	1lse
1lsf	1lsm	1lsn	1lysA	1lysB	2lym	3lym	4lym	4lytA	4lytB	5lyt	6lyt	2lzt	1rcmA	1rcmB
1hwa	**1lym**	**1lyz**	**1lzt**	**2hf1**	**2lyz**	**3lyt**	**3lyz**	**4lyz**	**5lyz**	**6lyz**	**7lyz**	**8lyz**	**1laa**	
Pheasant														
1ghlA	1ghlB													
Quail														
2ihl														
Turkey														
135l	1lz2	1lz3	2lz2	3lz12										
Guinea Fowl														
1hhl														
Mammalian lysozymes														
Human														
133l	134l	1lhh	1lhi	1lhj	1lhk	1lhl	1lhm	2lhm	3lhm	1lz1	1lz4	1lz5	1tay	1tby
1tcy	1tdy													
Mouse														
1fdl	**2hfm**	**3hfm**	2iff											
Horse														
1eql	**2eql**													

Table 8.11: Summary of the eukaryotic lysozyme structures found in the PDB. The PDB codes in bold are those lysozymes whose catalytic residues are not identified by the template using a distance cut–off of 2.0Å.

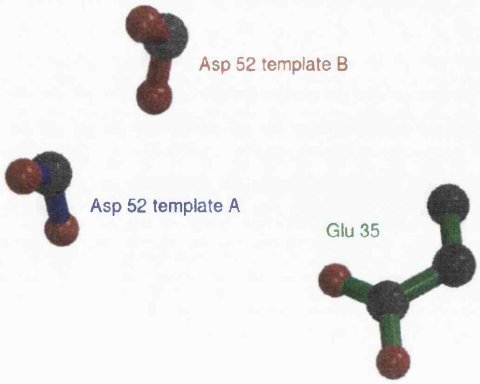


Figure 8.15: The two Asp 52 conformations with respect to the catalytic Glu 35 found for hen lysozyme. Template A is that in Table 8.10 while template B was derived from the structures of Diamond *et al.* (1975).

different position with respect to the main group. In fact the Diamond *et al.*, 1975 structures are refinement models of the original Blake *et al* (1965) lysozyme structures. These had substrate analogs and inhibitors bound and this indicates that the active site conformation changes upon binding of inhibitors to lysozyme. In addition there may be differences in the crystallisation conditions for these structures and this would affect the molecular packing in the crystal form.

The catalytic residues from horse (Tsuge *et al.*, 1992) are also not located; Figure 8.14 shows the relative position of the Asp 35 is shifted with respect to the other mammalian Asp residues. Finally, 3 structures from mouse lysozyme are also not picked out by the consensus template search. In fact these lysozymes have been crystallised in a complex with an immunoglobulin molecule (Fischmann *et al.*, 1991) suggesting there was a conformation change around the active site of lysozyme in this Ig–lysozyme complex.

Residue	Residue Number	Atom	x	y	z
Asp	20	C^{γ}	7.5	-3.6	1.5
Asp	20	O^{δ_1}	4.4	-3.8	5.1
Asp	20	O^{δ_2}	4.4	-4.3	3.1
Glu	11	C^{β}	-2.2	-1.3	-0.3
Glu	11	C^{γ}	-1.5	0.1	-0.0
Glu	11	C^{δ}	0.0	0.0	0.0
Glu	11	O^{ϵ_1}	0.6	-1.0	0.0
Glu	11	O^{ϵ_2}	0.6	1.1	0.0

Table 8.12: Coordinates of the consensus template describing the active site of the prokaryotic T4 lysozymes present in the PDB.

8.4.2 Prokaryotic: Bacteriophage T4 lysozyme

T4 lysozyme is produced late in the infection of *Escherichia coli* by T4 bacteriophage. The structure was first determined by Matthews & Remington (1974). This enzyme has been used as a model to study the effects of mutations on protein stability and function and there are over 150 different mutant forms deposited in the PDB (e.g. Weaver & Matthews, 1987; Alber *et al.*, 1987).

A consensus template was constructed from the T4 structure *2lzm* (Alber *et al.*, 1987) using the same atoms as the mammalian template; the distance cut–off was set at 3.0Å. The coordinates of the resultant complex are given in Table 8.12.

Of the 165 T4 structures in the PDB, 20 are not located by the consensus template; these are listed in Table 8.13. All 'missed' structures have mutations of the active site residues or of regions around the active site, perturbing the conformation of the catalytic residues.

2011A	1021	1031	1041A	1071	1081	1091	1101	1111	1121	1131	1141	1151	2171	1181
1191	1201	2211	1221	1231	2241	1251	1261	1271	1281	1291	1301	1311	1371A	1381
1391	1401	1411	1421	1431	1441	1451	1461	1471	1481E	1551	1561	1581	1591	1601
1611	1621	1631	1641	1651	1661	1dyb	1dyc	1dyd	1dye	1dyf	1dyg	1100	1103	1104
1107	1108	1110	1111	1112	1115	1116	1117	1118	1119	1120	1121	1122	1123	1124
1125	1127	1128	1130	1132	1133	1134	1137	1138	1139	1140	1141	1142	1143	1144
1145	1146	1147	1148	1149	1150	1151	1152	1155	1156	1157	1158	1159	1160	1163
1169	1171	1172	1173	1174	1175	1176	1177	1179	1180	1182	1183	1185	1186	1187
1188	1189	1190	1191	1192	1193	1194	1195	1196	1198	1199	1lyd	1lye	1lyf	1lyg
1lyh	1lyi	1lyj	2lzm	3lzm	4lzm	5lzm	6lzm	7lzm	152l	157l	1dya	1101	1102	1105
1106	1109	1113	1114	1126	1129	1131	1153	1154	1181	1184	1197	2051	2161	

Table 8.13: The prokaryotic bacteriophage T4 lysozyme structures in the PDB. The two structures underlined were not identified by the T4 consensus template.

8.4.3 Comparison of prokaryotic and eukaryotic lysozymes

Due to the similarity of the tertiary fold of both the prokaryotic and eukaryotic lysozyme structures and since they both have a Glu and Asp residue as their catalytic residues, it would not be surprising if the conformation of the catalytic residues is also the same. Figure 8.16 is a 3D representation of the consensus templates from the prokarotic T4 lysozyme and eukaryotic lysozymes. The Glu residues have been superimposed and the Asp residues lie around 4.5Å apart. It is proposed that the general mechanism of catalysis of these two lysozymes is the same. The crystal structures represent the ground state conformation of the enzyme; there is known to be considerable distortion of the substrate and active site during the reaction course. This suggests that the catalytic residues of these two lysozymes may move into similar positions during the transition state of the reaction.

8.4.4 Template search through the PDB

When the prokaryotic and eukaryotic lysozyme consensus templates are searched through the representative structures of the PDB, there are about 100 hits for each of the templates. This large number of hits occurs for several reasons. Firstly

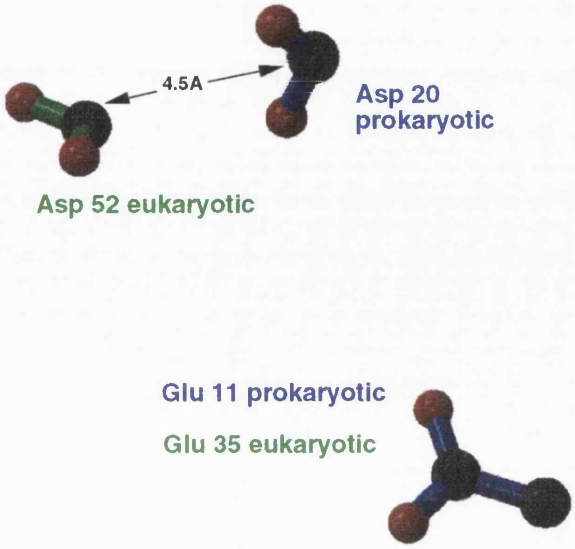


Figure 8.16: A 3D representation comparing the active site geometry of the catalytic residues from prokaryotic T4 lysozyme and eukaryotic lysozymes. The Glu residues of the two consensus templates have been superimposed allowing comparison of the catalytic Asp residues.

and most importantly, the consensus templates represent the ground state form of the active site. It is apparent that there is a conformational change of the catalytic residues in the transition state and there is no representation of this movement in the consensus templates. Secondly, in the active site of lysozyme, the substrate bisects the catalytic Glu and Asp residues. In non–catalytic regions of a protein, this space could easily be replaced by a residue, rendering the Glu and Asp residues non–catalytic. Thirdly, there are only two residues in the consensus templates, increasing the chance that the relatively common Glu and Asp residues will occur in these positions.

8.5 Is the conformation of catalytic residues unique to enzyme active sites?

In this chapter as well as chapters 3 and 5, after deriving an enzyme active site consensus template, we searched through a representative dataset of structures to see if this template was found in any other proteins in the PDB. We found that there were only a few hits located for those consensus templates with three or more residues or atoms (*e.g* the catalytic triad) but as expected there were far more for templates with only 2 residues or atoms (*e.g* lysozyme Asp 52–Glu 35).

This may occur simply because there is greater chance of two rather than three residues being located randomly in the same conformation in a given dataset of protein structures. To investigate this, we have taken 3 randomly picked non–catalytic Ser, His and Asp interactions and compared the number of hits located when these triads are used to search our representative dataset of PDB structures with those for the catalytic triad consensus template derived in chapter 3. Figure 8.17 shows the results of this test. The histogram in the top left represents the search with the catalytic triad consensus template derived from the seed template of 1*lpr* (Bone *et al.*, 1991) (chapter 3). The other 3 are non–catalytic interactions that have an *rms* deviation between 3Å and 6Å from the 1*lpr* consensus template; these are Ser 202 O^{γ}–His 205 sidechain–Asp 172 $O^{\delta 2}$ from chicken annexin 1*ala* (Bewley *et al.*, 1993), Ser 13 O^{γ}–His 503 sidechain–Asp 518 $O^{\delta 2}$ from cylcodextrin glycosyltransferase 1*cdg* (Lawson *et al.*, 1994) and Ser 104 O^{γ}–His 108 sidechain–Asp 201 $O^{\delta 1}$ from α–amylase 2*aaa* (Boel *et al.*, 1990). The bars shaded white in the histogram are hits located from the same protein family as the seed template. There are generally more hits located for the non–catalytic triads, however, as for the catalytic triad, none of the non–catalytic triads have matches below 1Å *rms* distance and very few are below 2Å. This suggests that any hits located when a

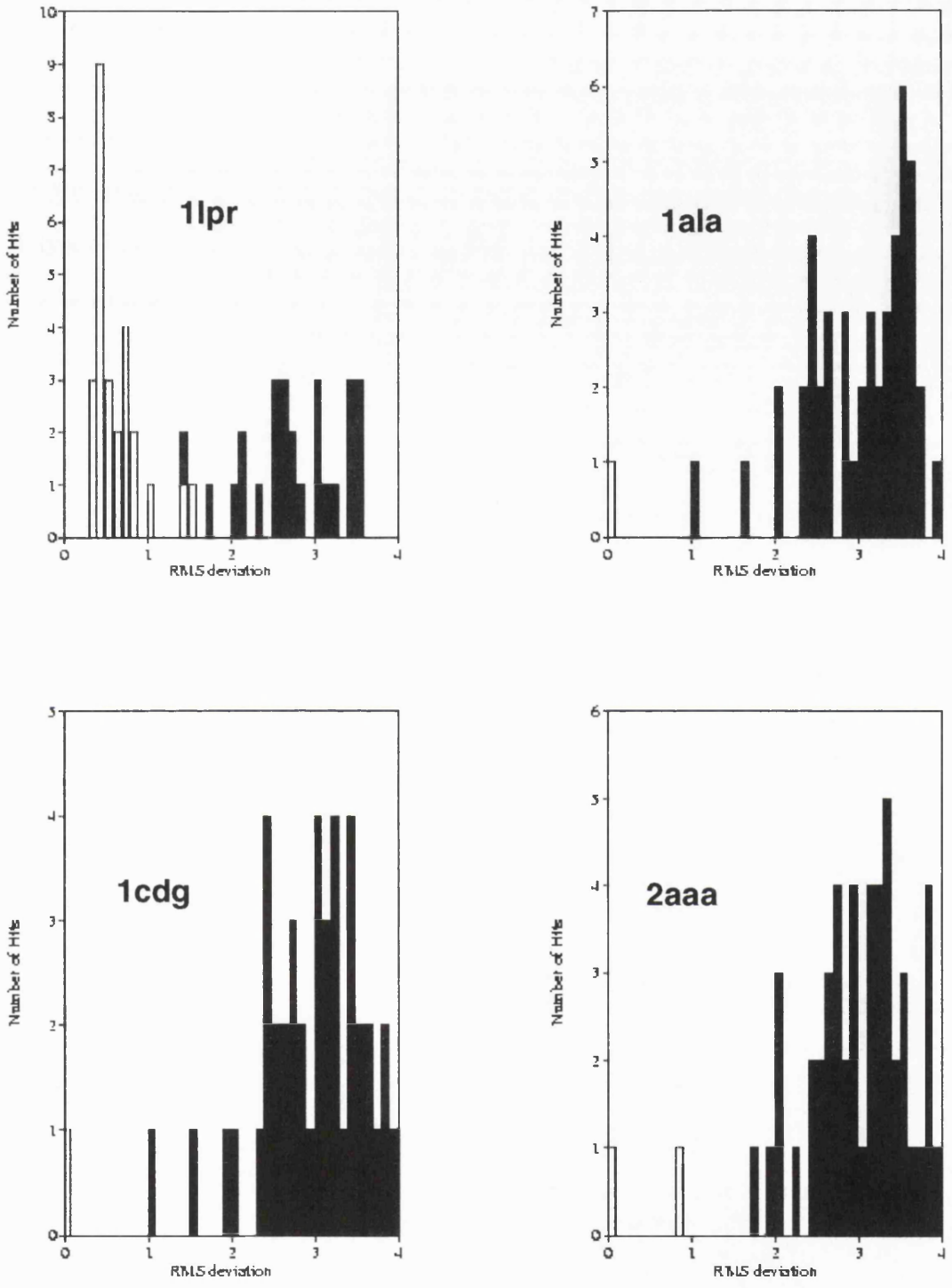


Figure 8.17: Histograms of the number of hits versus *rms* deviation from the respective templates when 4 Ser–His–Asp triads are used to search through the 95% by sequence non–homologous dataset. The *1lpr* triad is the catalytic consensus template for the serine proteinases and lipases. The other 3 are randomly chosen non–catalytic triads.

consensus template is used to search a representative dataset of proteins should be treated with caution and may not be of functional significance.

A similar test was carried out on a template consisting of only 2 residues; here we compared the 95% by–sequence non–homologous dataset search of the T4–lysozyme Asp 11–Glu 20 diad with 3 other randomly chosen non–catalytic Glu–Asp interactions. These diads were the sidechains of Glu 11–Asp 109 from cytochrome C550 155*c* (Timkovich & Dickerson, 1976), Glu 128–Asp 169 from foot–and–mouth virus 1*bbt* (Parry *et al.*, 1990) and Glu 172–189 from the elastase structure 1*ezm* (Thayer *et al.*, 1991). Figure 8.18 shows the number of hits located for searches with these templates. There are around the same number of hits located for the catalytic as the non–catalytic diads. In addition, for all 4 diads tested, the are no hits below 0.8Å from the respective consensus templates. Therefore if a hit is found below 1Å, it may well be significant.

This proves that as expected the number of hits located when searching a dataset of protein structures with a consensus template depends on the number of atoms and residues in that template. If hits are located below the chosen *rms* cut–off, it does not necessarily mean they are functionally significant but merely provides a possible starting point for further experimental investigation. Indeed, when such a hit is located, other factors should be considered, such as locality with respect to potential ligand binding site or accessibility to the protein surface.

These results are borne out by the template searches through the non–homologous protein dataset in chapters 3, 5 and 7. Though there were always several hits below the defined *rms* distance cut–off, there are no clear example of a hit revealing a new function for a protein. It should be noted that the functions of the vast majority of proteins in the PDB are understood and the PDB is by no means a representative dataset of all proteins in the genome. Only as the number of protein structures increases will we be able to appraise the power of such a

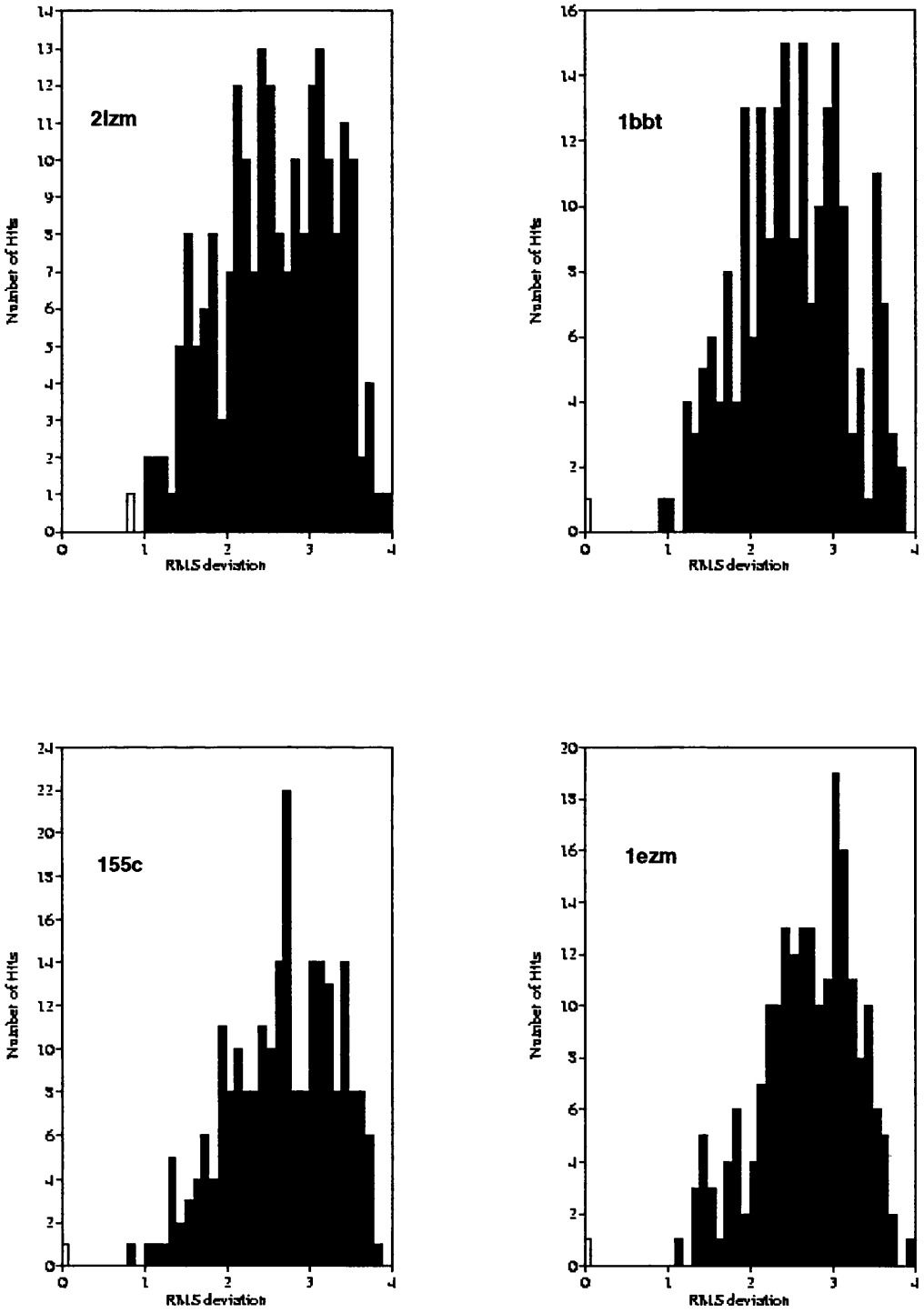


Figure 8.18: Histograms of the number of hits versus *rms* deviation from the respective templates when 4 Asp–Glu diads are used to search through the 95% by sequence non–homologous dataset. The *2lzm* diad is the catalytic consensus template for T4 lysozyme. The other 3 are randomly chosen non–catalytic Asp–Glu diads.

search method.

8.6 Conclusion

We have shown that defining a 3D consensus template can be a complex and time consuming process that requires comprehensive literature searches to identify the residues responsible for chemical catalysis. There is, as yet, no clear method for automatically defining the catalytic residues and atoms involved in an enzyme's catalytic machinery.

Other factors also complicate matters: we saw that the prokaryotic and eukaryotic lysozymes are related by either convergent or divergent evolution. This meant separate active site consensus templates were required for each of these groups. There are many types of ribonucleases depending on their origin, specificity for substrate and, in higher organisms, the organs from which they originate. These again are related by convergent evolution and lead to more than one ribonuclease consensus template. Furthermore, ribonuclease A exhibits more than one conformation of its catalytic His 119 residue. It is unclear which of these residue conformations is catalytically active so two templates were constructed, one for each conformer.

In addition, there is no clear way to validate a given consensus template. Searching a dataset of representative protein structures gives an indication of a consensus template's occurrence in other proteins but, as we saw for lysozyme, this leads to a large number of hits due to the nature of the derived consensus template.

There are at present over 200 distinct enzymes present in the January 1995 release of the PDB and this will rise around 5 fold by the turn of the century. Due to this, one way to create a database of 3D enzyme active site templates might

be to make the TESS program accessible over the WWW to enable experts on a particular enzyme to create and deposit a consensus template for a new enzyme structure.

This database would be useful for several reasons. Firstly, it will enable swift evaluation of new PDB structures as they are solved. Since the number of X–ray and NMR structures is expected to increase to around 50000 by the turn of the century, it is clear that the functions of these new structures need to be evaluated swiftly. As well as aiding structure based drug design, it would enable automatic searches for potential ligand binding sites in other proteins, possibly of unknown function. It would also give potential leads for protein engineering experiments, in designing novel enzymes, or enzymes which act on different substrates.

8.7 References

Alber T., Dao-Pin S., Nye J.A., Muchmore D.C. & Matthews B.W. (1987) T4 Temperature sensitive mutations of bacteriophage T4 lysozyme occur at sites with low mobility and low solvent accessibility in the folded protein *Biochemistry* **26** 3754–3758

Arni R.K., Pal G.P., Ravichandran K.G., Tulinsky A., Walz F.G. & Metcalf P. Jr. (1992) Three–dimensional structure of Gln 25–ribonuclease T1 at 1.84Å resolution: structural variations at the base recognition and catalytic sites *Biochemistry* **31** 3126–3135

Bairoch A. & Bucher P. (1994) PROSITE: recent developments. *Nucleic Acids Res.* **22** 3583-3589

Beintema J.J., Schuller C., Masachika I. & Carsana A. (1988) Molecular evolution of the ribonuclease superfamily *Prog. Biophys. molec. Biol.* **51**

165–192

Beintema J.J., Confalone E., Sasso M. P., & Furia A. (1990) In: Cuchillo M.C., Llorens R., Nogues M.V., & Pares X. (eds). *Structure mechanism and function of ribonucleases. Gerona* 275–281

Bewley M.C., Boustead C.M., Walker J.H., Waller D.A. & Huber R. (1993) Structure of chicken annexin V at 2.25Å resolution *Biochemistry* **32** 3923–3929

Blake C.C.F., Koenig D.F., Mair G.A., North A.C.T., Phillips D.C. & Sarma V.R. (1965) Structure of hen egg-white lysozyme, a three–dimensional fourier synthesis at 2Å resolution *Nature* **206** 757–780

Blackburn P. & Moore S. (1982) Pancreatic ribonuclease, In: Boyer P.D. ed., *The Enzymes* New York: Academic Press 317–433

Bleasby, A.J., Akrigg, D. & Attwood, T.K. (1994) OWL – A non-redundant, composite protein sequence database *Nucleic Acids Res.* **22** 3574–3577

Boel E., Brady L., Brzozowski A.M., Derewenda Z., Dodson G.G., Jensen V.J., Peterson S.B., Swift H., Thim L. & Woldike H.F. (1990) Calcium binding in α–amylases: an X–ray diffraction study at 2.1Å resolution of two enzymes from *Aspergillus Biochemistry* **29** 6244–6249

Bone R., Fujishige A., Kettner C.A. & Agard D.A. (1991) Structural basis for broad specificity in α–lytic protease mutants *Biochemistry* **30** 10388-10398

Borkakoti N. (1983) The active site of ribonuclease A from the crystallographic studies of ribonuclease–A–inhibitor complexes *Eur J. Biochem.* **132** 89–94

Borkakoti N., Moss D.S., Stanford M.J. & Palmer R.A. (1982) The refined structure of ribonuclease A at 1.45Å resolution *J. Crystallogr. Spectros. Res.*

14 467–471

Buckle A.M., Hendrick K. & Fersht A.R. (1993) Crystal structural analysis of mutations in the hydrophobic cores of barnase *J. Mol. Biol.* **234** 847–861

Crestfield A.M., Stein W.H. & Moore S. (1963) Alkylation and identification of the histidine residues at the active site of ribonuclease *J. Biol. Chem.* **238** 2413–2420

deMel V.S., Martin P.D., Doscher M.S. & Edwards B.F. (1992) Structural changes that accompany the reduced catalytic efficiency of two semisynthetic ribonuclease analogs *J. Biol. Chem.* **267** 247–256

Findlay D., Herries D.G., Mathias A.P., Rabin B.R. & Ross C.A. (1962) The active site and mechanism of action of bovine pancreatic ribonuclease 7. The catalytic mechanism *Biochem. J.* **85** 152–153

Fischmann T.O., Bentley G.A., Bhat T.N., Boulot T.N., Mariuzza R.A. Phillips S.E.V., Tello D. & Poljak R.J. (1991) Crystallographic refinement of the 3–dimensional structure of the FABD1.3–lysozyme complex at 2.5Å resolution *J. Biol. Chem.* **266** 12915–12920

Guillet V., Lapthorn A., Hartley R.W. & Maugen Y. (1993) Recognition between a bacterial ribonuclease, barnase, and its natural inhibitor, barstar *Structure* **1** 165–176

Harata K., Muraki M. & Jigami Y. (1993) Role of Arg 115 in the catalytic action of human lysozyme – X–ray structure of His 115 and Glu 115 mutants *J. Mol. Biol.* **233** 524–535

Hennig M., Schlesier B., Dauter Z., Pfeiffer S., Betzel C., Hoehne W.E. & Wilson K.S. (1992) A TIM–barrel protein without enzymatic activity? Crystal structure of Narbonin at 1.8Å resolution *Febs lett.* **306** 80–84

Hennig M., Pfeiffer S., Dauter Z., Wilson K.S., Schlesier B. & Hai Nong V. (1995) Crystal structure of narbonin at 1.8Å resolution *Acta Cryst.* **D51** 177–189

Holmes M.A. & Stemkamp R.E. (1992) Structures of met and azidomet hemerythrin at 1.66Å resolution *J. Mol. Biol.* **220** 723–737

Koepke J., Maslowska M., Heinemann U. & Saenger W. (1989) Three–dimensional structure of ribonuclease T_1 complexed with guanylyl–2'5'–guanosine at 1.8Å resolution *J. Mol. Biol* **206** 475–488

Imoto T., Johnson L.N., North A.C.T., Phillips D.C. & Rupley J.A. (1972) Vertebrate lysozyme. In Boyer P.D. (ed.) *The Enzymes* (3^{rd} edition) **7** 666–868 Academic Press.

Laskowski R.A., Luscombe N.M., Swindells M.B. & Thornton J.M. (1996) Protein clefts in molecular recognition and function *Protein Science* To be published.

Lawson C.L., Van Montfort R., Strokopytov B., Rozeboom H.J., Kalk K.H., De Vries G.E., Penninga D., Dijkhuizen L. & Dijkstra B.W. (1994) Nucleotide sequence and X–ray structure of cyclodextrin glycosyltransferase from *Bacillus circulans* strain 251 in a maltose–dependent crystal form *J. Mol. Biol.* **236** 590–600

Lenz A., Heinemann U., Maslowska M. & Saenger W. (1991) X–ray analysis of cubic crystals of the complex formed between ribonuclease T1 and guanosine–3',5'–bisphosphate *Acta. Crystallogr.* **B47** 521–527

Lichtarge O., Bourne H.R. & Cohen F.E. (1996) An evolutionary trace method defines surfaces common to protein families *J. Mol. Biol.* **257** 342–358

Lundqvist T. & Schneider G. (1989) Crystal structure of the complex of ribulose–1,5–bisphosphate carboxylase and a transition state analogue, 2–carboxy-D–arabinitol 1,5–bisphosphate *J. Biol. Chem.* **264** 7078–7083

Matthews B.W. & Remington S.J. (1974) The three–dimensional structure of the lysozyme from bacteriophage T4 *Proc. Natl. Acad. Sci., U.S.A.* **71** 4178–4182

Parry N., Fox G., Rowlands D., Brown F., Fry E., Acharya R., Logan D. & Stuart D. (1990) Structural and serological evidence for a novel mechanism of antigenic variation in foot–and–mouth disease virus *Nature* **347** 569–572

Peters K.P., Fauck J. & Frommel C. (1996) The automatic search for ligand binding sites in proteins of known three–dimensional structure using geometric criteria. *J. Mol. Biol.* **256** 201–213

Remington S.J. & Matthews B.W. (1978) Method to assess the similarity of protein structures, with applications to T4 bacteriophage lysozyme *Proc. Natl. Acad. Sci. USA* **75** 2180–2184

Santoro J., Gonzalez C., Bruix M., Neira J. L., Herranz J., & Rico M. (1993) High–resolution three–dimensional structure of ribonuclease A in solution by nuclear magnetic resonance spectroscopy *J. Mol. Biol.* **229** 722–734

Schneider G., Lindqvist Y. & Lundqvist T. (1990) Crystallographic refinement and structure of ribulose–1,5–bisphosphate carboxylase from *Rhodospirillum rubrum* at 1.7Å resolution *J. Mol. Biol.* **211** 989–1008

Smith L.J., Sutcliffe M.J., Redfield C. & Dobson C.M. (1993) Structure of hen lysozyme in solution *J. Mol. Biol.* **229** 930–944

Thayer M.M., Flaherty K.M., McKay D.B. (1991) Three–dimensional structure of the elastase of *Pseudomonas aeroginosa* at 1.5Å resolution *J. Biol. Chem.* **266** 2864–2871

Tsuge H., Ago H., Noma M., Nitta K., Sugai S. & Miyano M. (1992) Lysozyme from equine milk at 2.5Å resolution *J. Biochem* **111** 141–143

Varadarajan R. & Richards F.M. (1992) Crystallographic structures of ribonuclease S variants with nonpolar substitutions at position 12: packing and cavities *Biochemistry* **31** 12315–12327

Wallace A.C., Laskowski R.A. & Thornton J.M. (1995) LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions *Protein Engineering* **8** 127–134

Verscheuren H.G., Seljee F., Rozeboom HJ., Kalk K.H & Dijkstra B.W. (1993) Crystallographic analysis of the catalytic mechanism of haloalkane dehalogenase *Nature* **363** 693–698

Weaver L.H. & Matthews B.W. (1987) Structure of bacteriophage T4 lysozyme refined at 1.7Å resolution *J. Mol. Biol.* **193** 189–199

Zegers I., Maes D., Dao–Thi M., Poortmans F., Palmer R. & Wyns L. (1994) The structures of RNase A complexed with 3'–CMP and d(CpA): Active site conformation and conserved water molecules *Protein Science* **3** 2322–2339

Zegers I., Verhelst P., Choe H.W., Steyaert J., Heinemann U., Saenger W. & Wyns L. (1992) The role of histidine–40 in ribonuclease T1 catalysis: 3–dimensional structures of the partially active His40Lys mutant *Biochemistry* **31** 11317–11325

Chapter 9

Summary

The following is a summary of the main achievements of this thesis:

- A fully automated computer program called LIGPLOT has been developed that draws schematic diagrams of protein–ligand interactions.
- Using all serine–proteinase and lipase X–ray and NMR structures, we have proved that it is possible to define a 3D consensus template, in this case consisting of Ser, His and Asp, that is able to identify all catalytic Ser–His–Asp triads in the PDB with the exclusion of all other non–catalytic Ser, His and Asp interactions.
- A computer program called TESS has been developed that allows the automatic production of 3D consensus templates for any enzyme active site as long as the catalytic residues are known.
- We have shown that one 3D consensus template, consisting of Nu:–His–ELEC is able to identify the active site residues of the serine proteinases, lipases and the α/β hydrolase enzymes, with the exclusion of all other interactions.
- The orientation of the ligand binding sites of the enzymes identified with the Nu:–His–ELEC are varied. However, there is a clear relationship between the

orientation of the sidechain if the Nu: group and the ligand binding site.

- Analysis of metal–His interactions in the PDB has revealed that they are usually situated in functional sites. In addition, a triad of type metal–His–ELEC has been defined, however, unlike the Nu:-His–ELEC triad, it is structurally heterogeneous and we have been unable to separate catalytic metal–His–ELEC triads from other interactions. This illustrates the fact that metal binding sites are often distorted in structure.

- Taking ribonuclease and lysozyme as examples, it is clear that defining a 3D template is not a straightforward process. There may need to be more than one template defining an enzyme family; for example separate templates have been constructed for eukaryotic and prokaryotic lysozymes. Furthermore, residues can adopt more than one conformation and this may need to be taken into consideration. There is as yet no obvious way to fully automate the construction of 3D enzyme active site templates.