Multicofactor proteins: Structure, prediction, function

By

Stephen J. Hearnshaw

A thesis presented for the degree of Doctor of Philosophy at the University of East Anglia

April 2011

Declaration

I declare that the work contained in this thesis, submitted by me for the degree of Ph.D., is my own original work, except where due reference is made to the authors, and has not been submitted by me for a degree at this or any other university.

Stephen J. Hearnshaw

Acknowledgements

Firstly, I would like to thank my supervisor, Dr. Andrew Hemmings, for his invaluable support and guidance throughout my study. I would also like to thank Liang Zhou for providing me with the CopZ protein samples and Cornelius Friedrich for providing me with the SoxF protein sample.

I acknowledge access to the Diamond beamlines I02, I03 and I04; the ESRF beamline BM14 at Grenoble, France and the SRS station 10.1 where the vast majority of the diffraction datasets were collected.

Finally, I would like to thank my family for all their support and encouragement throughout the course of my Ph.D. and especially my girlfriend Alix for her unwavering support and patience.

Abstract

The current methods available for protein structure prediction are quite unsuitable for cofactor containing proteins, as the cofactors themselves are not taken into account during the prediction methodologies, which can seriously affect the quality of the overall prediction. One of the primary aims of this thesis is to begin to solve this problem.

This project has two distinct areas; (1) the development of methodologies for the prediction of cofactor rich proteins, namely multiheme proteins and (2) the experimental structural determination of two cofactor containing proteins; the flavocytochrome c sulfide dehydrogenase SoxF from *Paracoccus denitrificans* and the copper chaperone protein CopZ from *Bacillus subtilis*.

The multiheme protein structure prediction methodology developed in this work builds its models around the packing of the hemes, which have found to be conserved within protein families. The methodology has had some successes and shown significant improvements over the existing tools currently available to the wider scientific community.

High resolution structures for CopZ have been determined with different packings of Cu(I) to CopZ monomers, namely a dimer containing a tetranuclear Cu(I) cluster and a trimer containing a trinuclear Cu(I) cluster. The trimeric CopZ structure has led to the generation of models for the prediction of Cu(I) transfer in *Bacillus subtilis* between CopZ and its target protein CopA via a hetero-trimeric complex.

The structure determination of SoxF had shed new light on the nature of the active site of this class of sulfite dehydrogenases, through which I have put forward a method for this proteins observed function as a catalyst for the reactivation of the sulfur oxidising, sox cycle, which is responsible for oxidising inorganic sulfur species to sulfate.

Contents

Chapter 1 - Introduction

1.1 Heme Proteins

1.1.1 Introduction

A heme is a prosthetic group that consists of an iron atom bound in the centre of a porphyrin ring, proteins that contain such groups are known as hemoproteins. These proteins have diverse biological functions including the transport of diatomic gases, chemical catalysis and electron transport. The iron in the heme group serves as a source or sink of electrons during electron transfer or redox chemistry due to the its ability to exist in its ferrous (Fe^{2+}), ferric (Fe^{3+}) or ferryl (Fe^{4+}) state. Hemoproteins achieve this remarkable functional diversity by modifying the environment around the heme groups.

1.1.2 Types of heme

Several different heme types exist in nature, the most common of which are shown in figure 1.1. Of these the most abundant is the B-type heme (Figure 1.1B), this heme is found in hemoglobin and myoglobin, as well as the peroxidase family of enzymes. Generally, B-type hemes are attached to the surrounding protein matrix by a single coordination bond between the heme iron and an amino acid side chain. For hemoglobin, myoglobin and the peroxidases this is a histidine residue, however, in heme-thiolate proteins, such as cytochrome P450, the residue is a cysteine **[Omura *et al* 2005]**.

C-type hemes are similar in structure to B-type hemes, the only differences being that the two vinyl side chains at ring positions two and three are covalently bound to the protein matrix through thioether linkages from cysteine residues (Figure 1.1C). In addition to these covalent bonds, the heme iron is generally coordinated by two conserved amino acid side chains. The cytochrome *c* electron transfer proteins are an example of hemoproteins that contain C-type hemes, the fifth heme iron ligand is always provided by a histidine residue and if a sixth is present it is generally provided by a methionine residue (in the case of class one cytochrome *c* proteins **[Bushnell *et al* 1990]**) or another histidine residue (in the case of class three cytochrome *c* proteins **[Czjzek *et al* 1994]**).

The A-type heme (Figure 1.1A) differs in structure from the B and C type hemes by incorporating an isopropanoid chain at ring position two and oxidising the methyl side chain at ring position eight into a formyl group. As with B-type hemes, A-type hemes are generally attached to the protein matrix by a coordination bond between the heme iron and a conserved amino acid side chain. An example of a protein with this type of heme is cytochrome *c* oxidase, the last protein in the electron transport chain, it receives four electrons (donated by cytochrome *c*) and transfers them to one oxygen molecule reducing it to two water molecules. This process also involves the translocation of four protons across the membrane, creating a proton motive force that ATP synthase uses to synthesise ATP, it is also thought that the formyl group and isopropanoid side chain play important roles in energy conservation during this reaction **[Papa *et al* 1998]**.

O-type hemes (Figure 1.1D) are structurally homologous to A-type hemes, the only difference being the methyl group at ring position eight has not been oxidised to a formyl group. An O-type heme has been isolated in the *Escherichia coli* enzyme ubiquinol oxidase, where it was found to reduce oxygen in a similar manner to the A-type heme **[Abramson *et al* 2000]**.

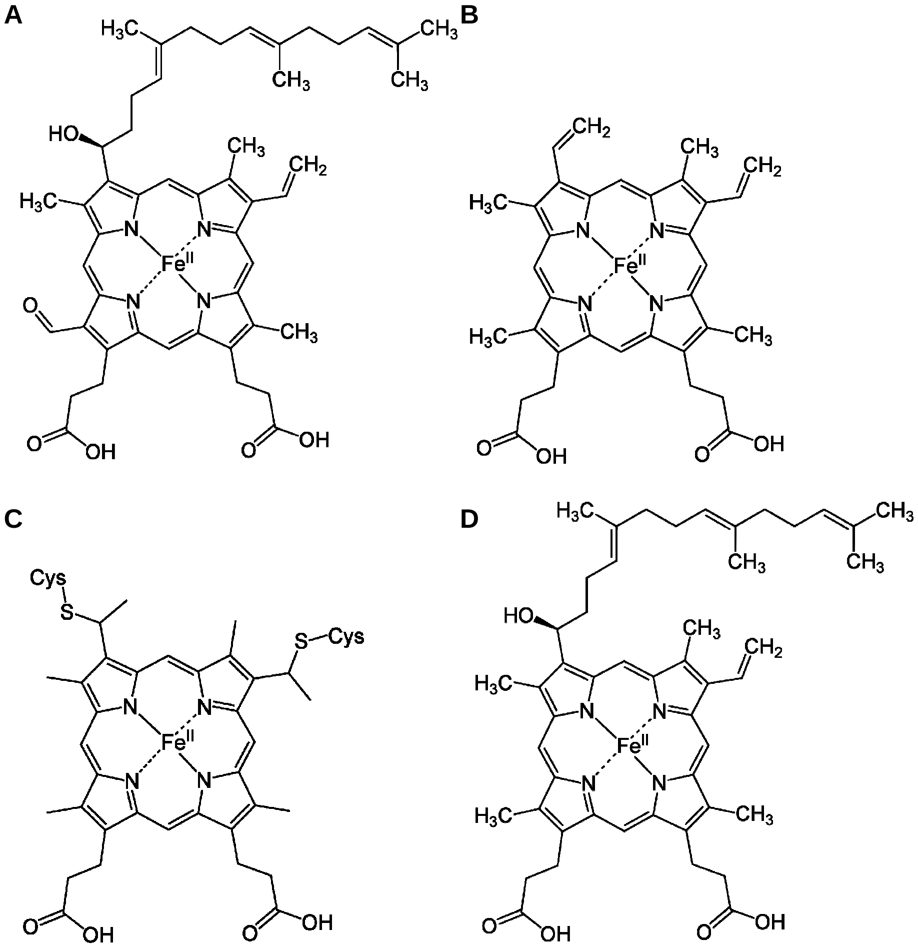


Figure 1.1 – Structures of the most common heme types, showing; **(A)** an A-type heme, **(B)** a B-type heme, **(C)** a C-type heme and **(D)** an O-type heme.

The heme iron is able to exist in three oxidation states, the most common being the ferrous (Fe^{2+}) and ferric (Fe^{3+}) states, with the ferryl (Fe^{4+}) state less common. Several heme proteins, including cytochrome *c*'s, peroxidases and cytochrome P450's are able to access more than one of these iron oxidation states during their functional processes. It is the heme iron's ability to undergo such redox chemistry and electron transfer that leads to the wide variety of functions in hemoproteins.

1.1.3 Heme synthesis and degradation

The basic protoheme (B-type heme) is synthesised in a seven step process involving successive enzymatic reactions (Figure 1.2), beginning with the universal tetrapyrrole precursor δ-aminolevulinic acid (ALA) created by the condensation of glycine and succinyl-CoA via δ-aminolevulinic acid synthase, which is the rate limiting enzyme for this pathway **[Anderson *et al* 2001]**.

Two molecules of ALA are condensed by ALA dehydratase to form porphobilinogen (PBG). PGB deaminase catalyses successive condensations of PGB, initiated by the elimination of the NH_2 group, until the linear tetrapyrrole hydroxymethylbilane is formed. This intermediate is converted by uroporphyrinogen III synthase to the macrocyclic uroporphyrinogen III, which is a precursor for vitamin B_{12} and siroheme biosynthesis **[O'Brian *et al* 1996]**. Next uroporphyrinogen III decarboxylase converts all four acetyl side chains to methyl side chains, forming coporphyrinogen III, before coporphyrinogen III oxidase converts the propionyl groups at ring positions 2 and 4 to vinyl groups, forming protoporphyrinogen IX, oxidation of this intermediate adds more double bonds via the action of protoporphyrinogen IX oxidase, yielding protoporphyrin IX. Protoporphyrin IX is the point at which the heme and chlorophyll biosynthesis pathways diverge **[O'Brian *et al* 1996]**, with chlorophylls adding an Mg atom to their porphyrin centres and hemes adding an Fe^{2+} ion via the ferrochelatase enzyme to create the protoheme.

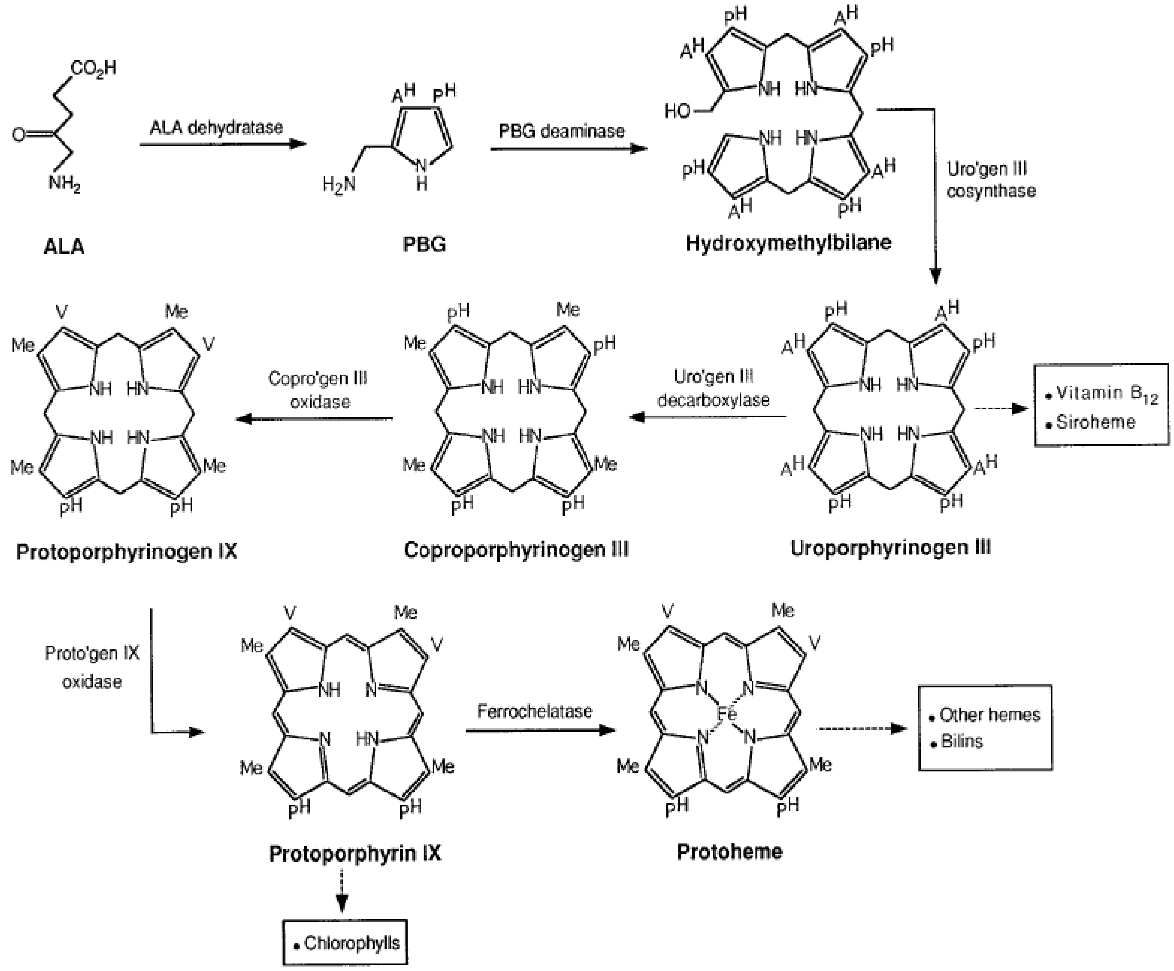


Figure 1.2 – The generic heme synthesis pathway, beginning with the universal tetrapyrrole precursor δ-aminolevulinic acid (ALA). The solid arrows correspond to the enzymatic reactions of the seven steps and the dashed arrows correspond to multi-step reactions leading to other tetrapyrrole derivatives. In addition to the substrates shown, coproporphyrinogen oxidase and protoporphyrinogen oxidase require O_2 in aerobic systems and another oxidant in anaerobic systems, and ferrochelatase requires ferrous iron. Abbreviations; ALA = δ-aminolevulinic acid, PBG = prophobilinogen, uro'gen = uroporphyrinogen, copro'gen = coproporphyrinogen, proto'gen = protoporphyrinogen, Me = methyl, A^H = acetyl, P^H = propionyl and V = vinyl. Figure taken from **[Anderson *et al* 2001]**.

The protoheme can be converted into other heme types by further enzymatic reactions. To create an O-type heme, farnesylation occurs at the vinyl group on ring position two of the protoheme, replacing it with a farnesyl group. This is thought to be carried out by a protoheme IX farnesyltransferase, coded for by the *cyoE* gene in *E.coli* **[Saiki *et al* 1993]**. The O-type heme is in fact an intermediate step in the production of an A-type heme, which is produced when the methyl group at heme position eight is oxidised to create a formyl group. This process is thought to occur via an initial hydroxylation of the position eight methyl group by a three-component monooxygenase consisting of Cox15p, ferredoxin and ferredoxin reductase, the resultant alcohol would then be further oxidised to the formyl group **[Barros *et al* 2002]**. A C-type heme is formed when a covalent attachment is made between the vinyl groups of the protoheme and the cysteine residues of the heme coordinating CXXCH motif of an apocytochrome *c* protein. This reaction is catalysed by a cytochrome *c* heme lyase, encoded by the CYC3 gene **[Moraes *et al* 2004]**.

The process of heme degradation is initiated by a family of enzymes known as heme oxygenases (HO) that catalyse oxidative degradation of ferric hemes to biliverdin IX, Fe^{2+} and carbon monoxide (CO), using NADPH as the reducing agent (Figure 1.3). In mammals biliverdin is further reduced to the potent antioxidant bilorubin by the action of biliverdin reductase, since bilirubin is toxic at high concentrations it is subsequently bound to glucuronic acid and excreted. The iron released by HO activity is normally recycled to keep up with the bodies daily iron requirement, and the CO has been identified as a factor in neuroendocrine regulation, a protective agent in hemorrhagic shock and a modulator of vascular tone **[Unno *et al* 2007]**.

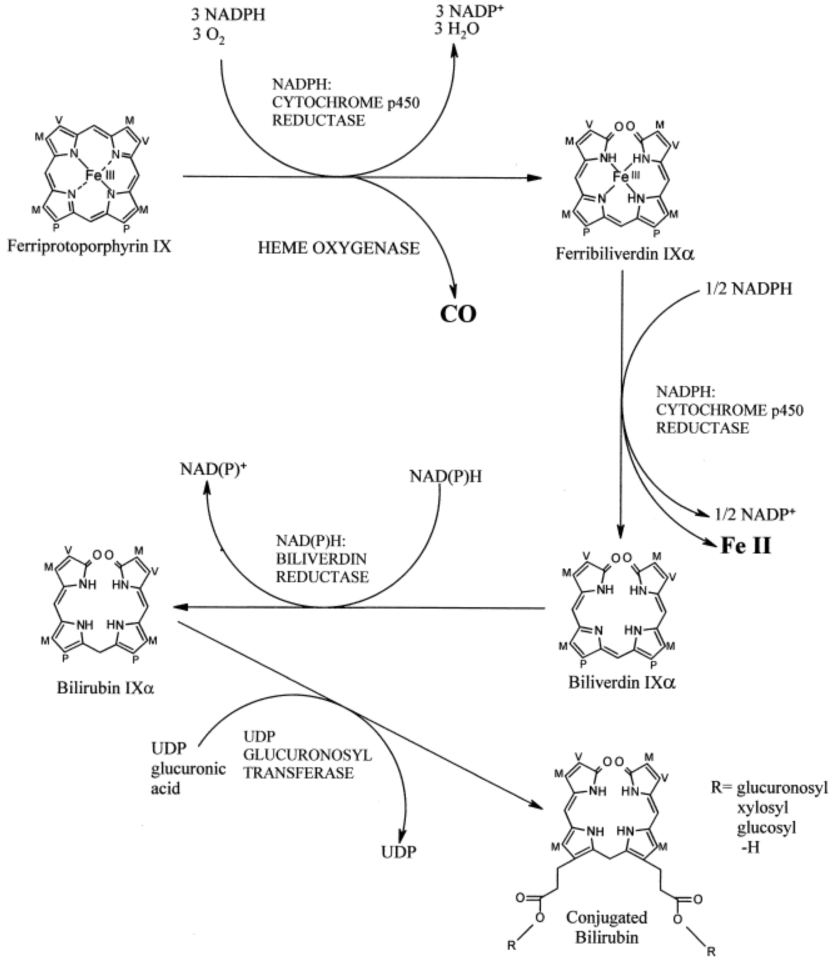


Figure 1.3 - Heme degradation. Heme oxygenase, catalyses the rate limiting step in heme metabolism. Both heme oxygenase enzymes (HO-1 and HO-2) oxidise ferric heme (ferriprotoporphyrin IX) to the bile pigment biliverdin-IXa (BV), in a reaction requiring 3 moles of molecular oxygen. NADPH:cytochrome p-450 reductase, reduces the ferric heme iron as a prerequisite for each cycle of oxygen binding and oxygen activation. The cleavage of the heme ring frees the a-methene bridge carbon as CO, and generates the biliverdin-iron complex (BV-Fe^{3+}). An additional NADPH dependent reduction releases Fe^{2+} from BV and the BV is reduced to BR by NAD(P)H:biliverdin reductase. Abbreviations: M = methyl, V = vinyl and P = propionate. Figure taken from **[Ryter and Tyrrell 2000]**.

Two forms of HO enzyme were discovered in the 1980's, they were called HO-1 and HO-2 respectively, and both have very different regulatory mechanisms for their production **[Maines *et al* 1986]**. Since then a third enzyme, related to HO-2 (≈90%), has been discovered that is thought to potentially have a heme-dependent regulatory role in the cell, although it has poor heme catalytic activity **[Mccoubrey *et al* 1997]**. HO-1 is induced by chemical agents and conditions that cause oxidative stress, including; heat shock, ischemia, GSH-depletion, radiation, hypoxia, hyperoxia, and cellular transformations and disease states. HO-2 is not induced by such stimuli; in fact the only chemical inducers of HO-2 identified to date are adrenal glucocorticoids **[Maines 1997]**.

Hemes have the ability to regulate their synthesis and degradation through feedback mechanisms to maintain intracellular heme levels. For example, the δ-aminolevulinic acid synthase enzyme (ALAS1), responsible for the production of ALA, has three heme regulatory motifs (HRMs) that consist of five amino acid residues ([Arg, Lys, or Asn]-Cys-Pro-[Lys or a hydrophobic residue]-[Lue or Met]) that are able to bind hemes. The binding of hemes to these HRMs prevents ALAS1 from undergoing translocation into the mitochondria where heme synthesis occurs, thus inhibiting heme synthesis **[Furuyama *et al* 2007]**. Hemes can also control their intracellular levels through transcriptional regulation of the HO-1 gene via the transcriptional repressor Bach1. Bach1 is a transcriptional repressor that is able to bind to the **MA**f **R**ecognition **E**lement (MARE) as a hetero dimer with a small maf family protein, subsequently repressing transcription, however, if the small maf family protein forms a hetero dimer with Nrf2 transcription is stimulated **[Sun *et al* 2002]**. Like ALAS1, back1 contains HRMs that hemes can bind with; this interaction prevents Bach1 from binding with the MARE site and thus prevents it from repressing the transcription of HO-1, leading to more HO production and heme degradation during periods of high heme concentration **[Suzuki *et al* 2004]**.

1.1.4 Structure and function of hemoproteins

Hemoproteins come in many forms and have many functions; these functions are dependent on the type of heme, heme iron oxidation state changes and the structure of the apoprotein itself, and can range from catalysis, to gas transport and channel proteins. This section will discuss how the heme type can affect protein function and how hemes are coordinated in hemoproteins. The principles arising from this discussion will be illustrated by examples of different hemoproteins.

1.1.4.1 How heme type can affect function

The most obvious difference between B and C type and A and O type hemes is the 17-carbon farnesyl group found in A and O type hemes that replaces the vinyl group found in B and C type hemes, this hydrophobic side chain has been identified as functionally important in several hemoproteins. Wang *et al* examined the effects of changing heme types in heme-copper oxidases (HCOs) **[Wang *et al* 2005]**. They found that HCOs remained mostly active after a substitution between A and O type hemes, however, replacing an A or O type heme with a B-type heme caused the HCOs to loose their activity, suggesting the 17-carbon farnesyl group played an important part in enzyme function. The general consensus for the function of this farnesyl group is as an anchor for keeping the heme in the correct position in the enzyme. It has also been proposed that this farnesyl group could be an essential part of the active site hydrogen bonding network which, along with internal water molecules, bridges the gap between tyrosine 288 and threonine 359 hydroxyl groups in the K-pathway of *Rhodobacter sphaeroides* **[Cukier *et al* 2004]**. Wang *et al* also experimented with replacing a B-type heme with an O-type heme in an engineered heme-copper site in myoglobin, they found that this change reduced the heme reduction potential by approximately 20 mV **[Wang *et al* 2005]**.

1.1.4.2 Heme coordination

The coordination of hemes within hemoproteins depends on several factors including; heme type, protein sequence and protein function. Heme type is important since hemes are coordinated differently depending on their structure. For example, A and B type hemes tend to be attached to the apoprotein through a single coordination bond between their iron centre and a conserved amino acid residue (Figure 1.4B), although they can also have two coordinating Fe-ligand bonds (Figure 1.4A & C). In contrast, C-type hemes are often attached to the apoprotein by two coordination bonds between their iron centre and two conserved amino acid residues (one of which is always a histidine), as well as these bonds they can also be coordinated by up to two thioether linkages from cysteine residues at ring positions two and four of the heme (Figure 1.4D).

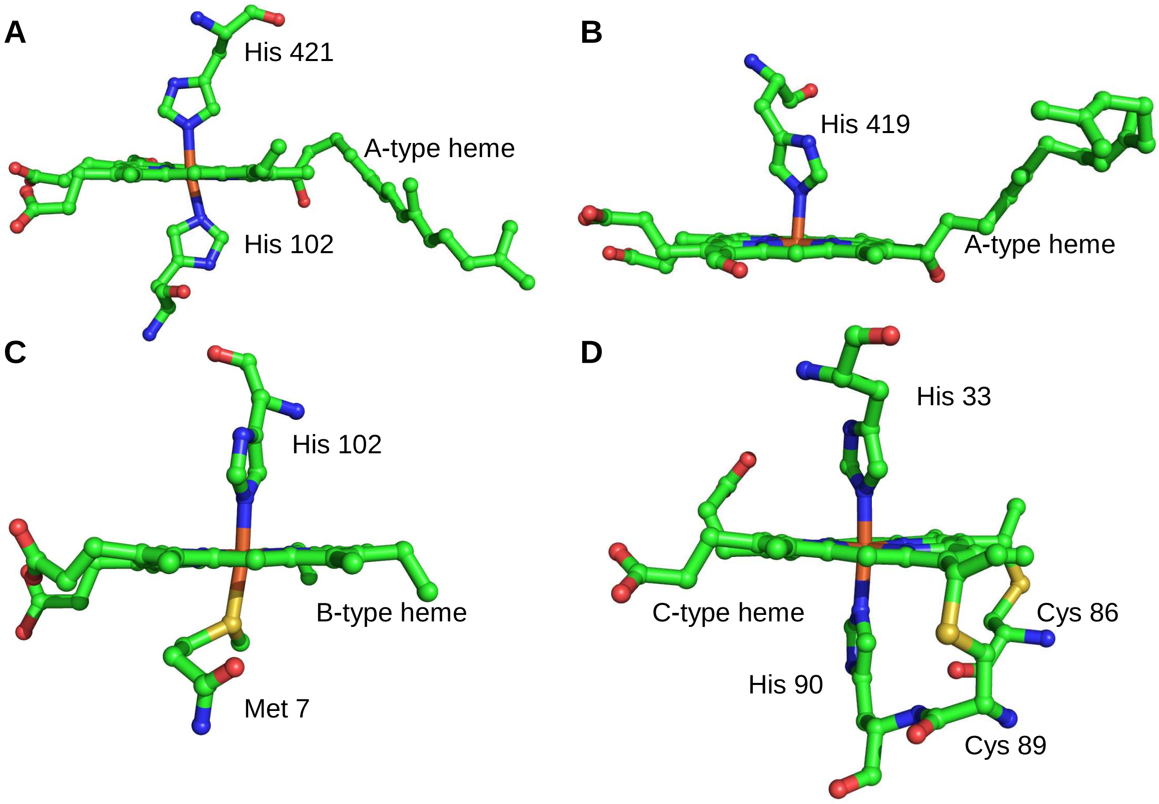


Figure 1.4 – Heme coordination in hemoproteins. Showing an A-type heme coordinated by two histidine residues **(A)** and one histidine residue **(B)** (PDB ID: 1M56 **[Svensson *et al 2002*]**), a B-type heme coordinated by a histidine residue and a methionine residue **(C)** (PDB ID: 1QQ3 **[Arnesano *et al* 2000]**) and a C-type heme coordinated by two histidine residues and two thioether linkages from cysteine residues **(D)** (PDB ID: 1AQE **[Aubert *et al* 1998]**).

The protein sequence is important in heme coordination because of the conserved residues that bind with the heme iron; the sequence must contain at least one histidine residue in close proximity to the heme group to act as an axial ligand, depending on the heme type another residue(s) may also be needed to act as a second axial ligand (e.g. histidine, methionine, tyrosine, etc) or to form thioether linkages. An example of sequence importance is the CXXCH binding motif used to bind C-type hemes, providing two cysteine residues for thioether linkages and a histidine for heme iron coordination **[Allen *et al* 2005]**, as a result this motif is very highly conserved amongst C-type heme binding hemoproteins.

Protein function is important in heme coordination because it will determine where in the protein the heme needs to be situated, i.e. if the heme is part of the catalytic active site for an enzyme it will need to be located around the active site where it can have access to the substrate, in these cases the heme would usually have only one axial ligand to expose the iron for performing redox reactions. The ligands that bind the heme can have an effect on the individual heme group or the protein as a whole. For example, carbon monoxide binding in the heme pocket of myoglobin results in a conformational relaxation of the protein **[Neinhaus *et al* 2002]**. A study by Das *et al*, using a *de novo* protein S824C, illustrated how ligand binding to heme groups can shift the hemes redox potential and that this shift responds differently to different ligands. They found that the binding of imidazole based ligands produced a negative shift in the hemes redox potential, whereas the binding of pyridine based ligands produced a positive shift in the hemes redox potential **[Das *et al* 2006]**. Similar results have been found with analysis of myoglobin, where binding of imidazole produced a negative shift in heme redox potential of approximately 50 mV **[Zhang *et al* 2003]**.

1.1.4.3 Hemoprotein functions – Heme enzymes

A well studied class of heme enzymes are the peroxidases, these enzymes are responsible for oxidising various biological substrates via the creation of high valent iron-oxygen intermediates by utilising an oxygen atom from hydrogen peroxide (H_2O_2) **[Poulos 2006]**. The first step in this catalytic process is the oxidation of the Fe^{3+} and porphyrin ring from the resting compound using hydrogen peroxide, creating an Fe^{4+} ion and a porphyrin π-cation radical, collectively known as compound I. The next step is a substrate oxidation by compound I, resulting in a one electron reduction of the porphyrin π-cation radical to a normal porphyrin containing an Fe^{4+} ion, this is collectively known as compound II. The final step is another one electron reduction resulting from substrate oxidation, reducing the Fe^{4+} ion back to Fe^{3+}, reforming the resting compound (Figure 1.5) **[Hersleth *et al* 2006]**.

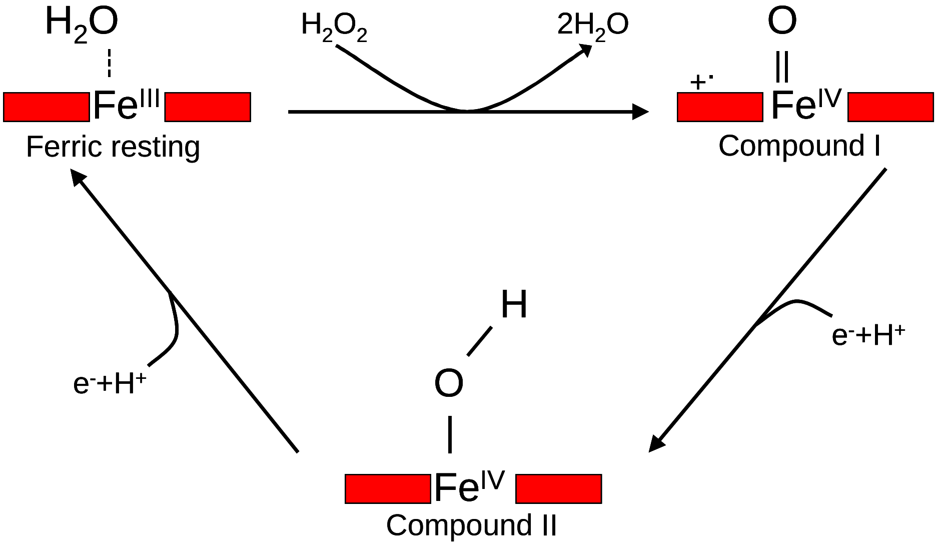


Figure 1.5 – The peroxidase catalytic cycle. The porphyrin ring is represented by red boxes on both sides of the Fe ion, and the porphyrin **π-**cation radical by + **.** **[Hersleth *et al* 2006]**

An example of a substrate oxidised by a heme peroxidase is ferulic acid, a plant cell wall protein that undergoes oligomerisation in the presence of horseradish peroxidase (HRP) and H_2O_2. HRP-catalysed oxidation of monomeric ferulic acid radicals leads to the formation of decarboxylated dehydrodimers that can be further oxidised by an additional ferulic acid monomer to form trimeric ferulic acid radicals **[Oudgenoeg *et al* 2002]**. The structure for HRP in a complex with ferulic acid was released in 1999 **[Henriksen *et al* 1999]**, in which the heme containing active site can be identified (Figure 1.6A) as well as how the ferulic acid substrate enters the active site (Figure 1.6B).

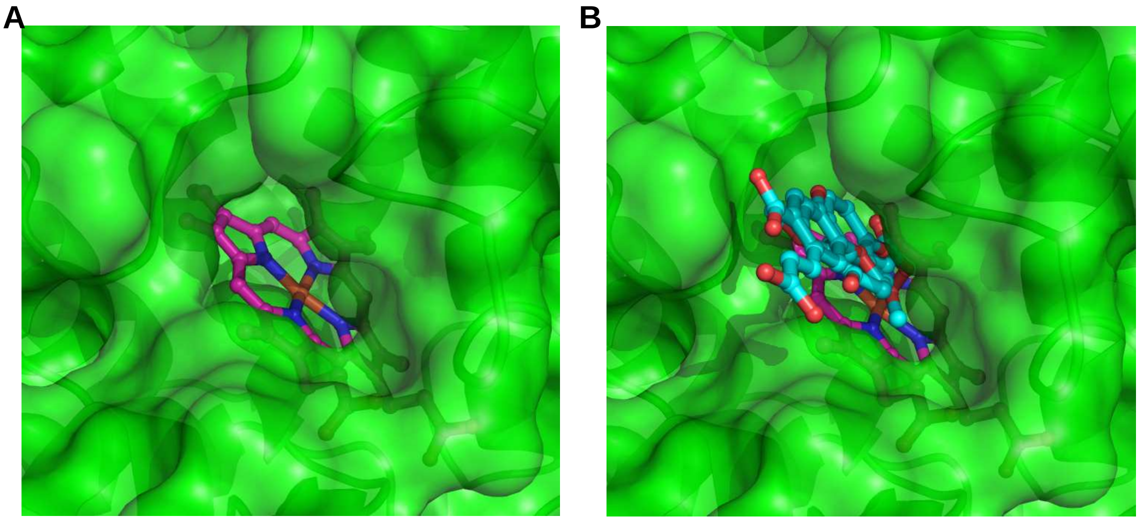


Figure 1.6 – The active site of horseradish peroxidase (PDB ID: 6ATJ) **[Henriksen *et al* 1999]**. Showing; **(A)** just the heme in the active site and **(B)** the active site with the heme and ferulic acid substrate present. The heme group is coloured magenta and ferulic acid substrate cyan.

Another example of a class of heme containing enzymes is the cytochrome P450s, the majority of which act as versatile monooxygenases. These enzymes are capable of catalysing many different reactions, including; the hydroxylation of alkanes to alcohols, conversion of alkenes to epoxides, arenes to phenols, sulfides to sulfoxides and sulfones, and the oxidative split of C-N, C-O, C-C or C-S bonds. The basic structure of all cytochrome P450s are relatively similar and all contain a well conserved heme-binding core, however the ability of cytochrome P450s to catalyse the reactions of many substrates of different conformations and charges mean the protein must be flexible to allow them to bind **[Zhao *et al* 2006]**. Structural studies of cytochrome P450s have shown the substrate is buried when bound to the active site, therefore the protein must be able to perform opening and closing motions to allow the substrate access to the active site **[Poulos 2005]**. The ability of cytochrome P450s to undergo this change in conformational state has been identified; Scott *et al* found that in mammalian cytochrome P450 some active site residues have the ability to move almost 19 Å, with the Ile114 residue being displaced by 18.9 Å **[Scott *et al* 2004]**.

1.1.4.4 Hemoprotein functions – Gas transport

A well studied gas transport hemoprotein is hemoglobin (Hb). Hb is the respiratory protein for the red blood cells, it allows them to carry oxygen from the lungs to the rest of the body, where the oxygen is exchanged for carbon dioxide and returned back to the lungs. Hb is a 64,500 Da heterotetrameric protein made up of two α and two β subunits that are 141 and 146 amino acid residues in length respectively (Figure 1.7A). Each subunit contains one heme group and can bind one molecule of oxygen when the heme iron is in the ferrous state. In this state the iron is bound to the heme through the four nitrogen's of the porphyrin ring, and coordinated in the protein by a histidine residue, this accounts for five of the irons six possible ligands, with the sixth being able to reversibly bind with an oxygen molecule (Figure 1.7B).

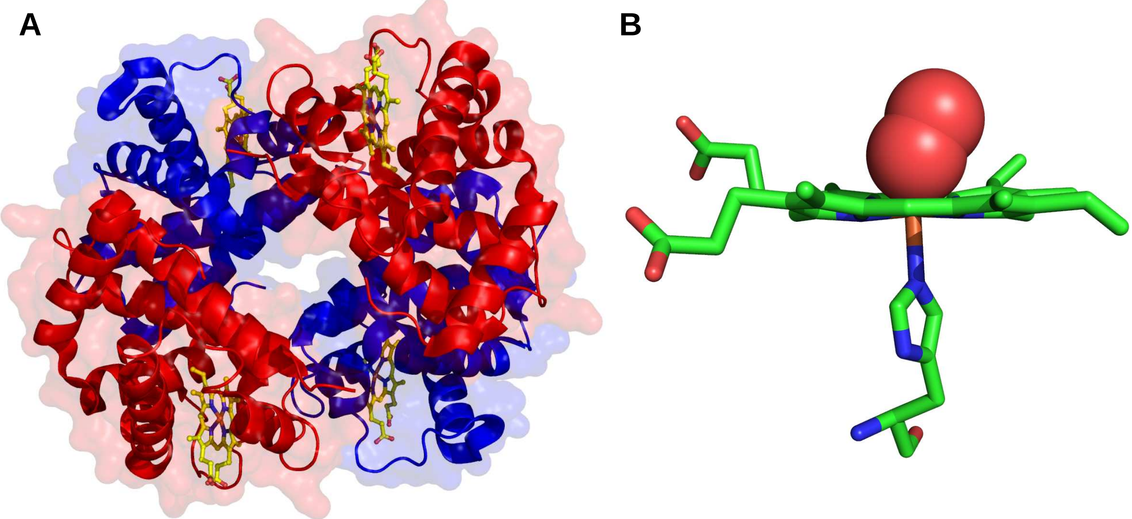


Figure 1.7 – The overall structure of hemoglobin and the binding of oxygen to the heme iron. **(A)** Shows the overall structure of hemoglobin (PDB ID: 2HHB **[Fermi *et al* 1984]**) with the α-subunits marked in red, the β-subunits in blue and hemes in yellow. **(B)** Displays how the heme and oxygen molecules interact, showing the heme group, iron coordinating histidine residue and oxygen molecule (red spheres).

Hb has the ability to exist in two states that are in rapid equilibrium; a "tense" state (T-state) with a low affinity for oxygen and a "relaxed" state (R-state) with a high affinity for oxygen **[Monod *et al* 1965]**. As well as being able to bind hemes, Hb has other binding sites that are able to bind alternative ligands, such as protons and chloride. It is hemes ability to bind protons that has been proposed as a trigger for conformational change between the T and R states, the binding of protons to R-state Hb reduces its affinity for oxygen, via a thermodynamic relationship that Wyman **[Wyman 1967]** termed "linked function", causing the oxygen to be released and carbon dioxide to be bound via the Bohr effect **[Tsuneshige *et al* 2002]**. Perutz proposed that the physical reason for the change in Hb conformation was the position of the heme iron atoms with respect to the plane of the hemes porphyrin ring **[Perutz 1972]**. In the T-state the iron is coordinated by five bonds and protrudes from the heme plane. Upon further ligation, via the binding of oxygen, the iron moves towards the heme plane, pulling on the proximal histidine in the process, breaking the $\alpha_1\beta_2$ and $\alpha_2\beta_1$ interactions formed during the T-state, resulting in Hb switching to its R-state.

1.1.4.5 Hemoproteins – Gas sensing\Transcriptional regulation

CooA is a carbon monoxide (CO) sensing heme protein, which upon sensing CO activates transcription of the *coo* operon; the genes responsible for metabolism of CO in *Rhodospirillum rubrum* **[Poulos 2006]**. CooA is a homodimeric protein, with each subunit containing 222 amino acid residues and one B-type heme that reversibly binds CO when the heme iron is in its ferrous state. CooA exists in two forms; an inactive form, where the heme in each subunit is coordinated by the His77 residue from that subunit and the Pro2 residue from the opposite subunit, and an active form where the hemes are coordinated by the His77 residue and bound to CO (Figure 1.8) **[Puranik *et al* 2004]**.

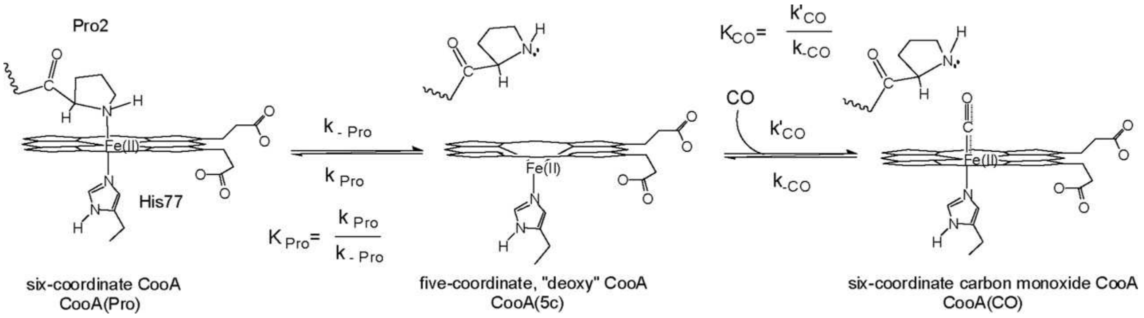

Pro2
His77
Fe(II)
k_{-Pro}
k_{Pro}
$K_{Pro}= \frac{k_{Pro}}{k_{-Pro}}$
$K_{CO}= \frac{k'_{CO}}{k_{-CO}}$
CO
k'_{CO}
k_{-CO}
six-coordinate CooA
CooA(Pro)
five-coordinate, "deoxy" CooA
CooA(5c)
six-coordinate carbon monoxide CooA
CooA(CO)

Figure 1.8 – Schematic for CO binding to a B-type heme in CooA. Figure taken from **[Puranik *et al* 2004]**.

A crystal structure of the inactive form of CooA has been solved by Lanzilotta *et al* that indicates how the hemes are coordinated by the His77 and Pro2 residues from each chain (Figure 1.9A) **[Lanzilotta *et al* 2000]**. However, no wild type structure for a transcriptionally active form of CO bound CooA has yet been reported. Komori *et al*, have produced a crystal structure for an imidazole-bound CooA, with imidazole taking the place of a CO molecule (Figure 1.9B) **[Komori *et al* 2007]**. This imidazole bound from was still found to be transcriptionally inactive; however the effect of a change in heme coordination on the overall structure of the protein can be clearly seen (Figure 1.9C vs. Figure 1.9D).

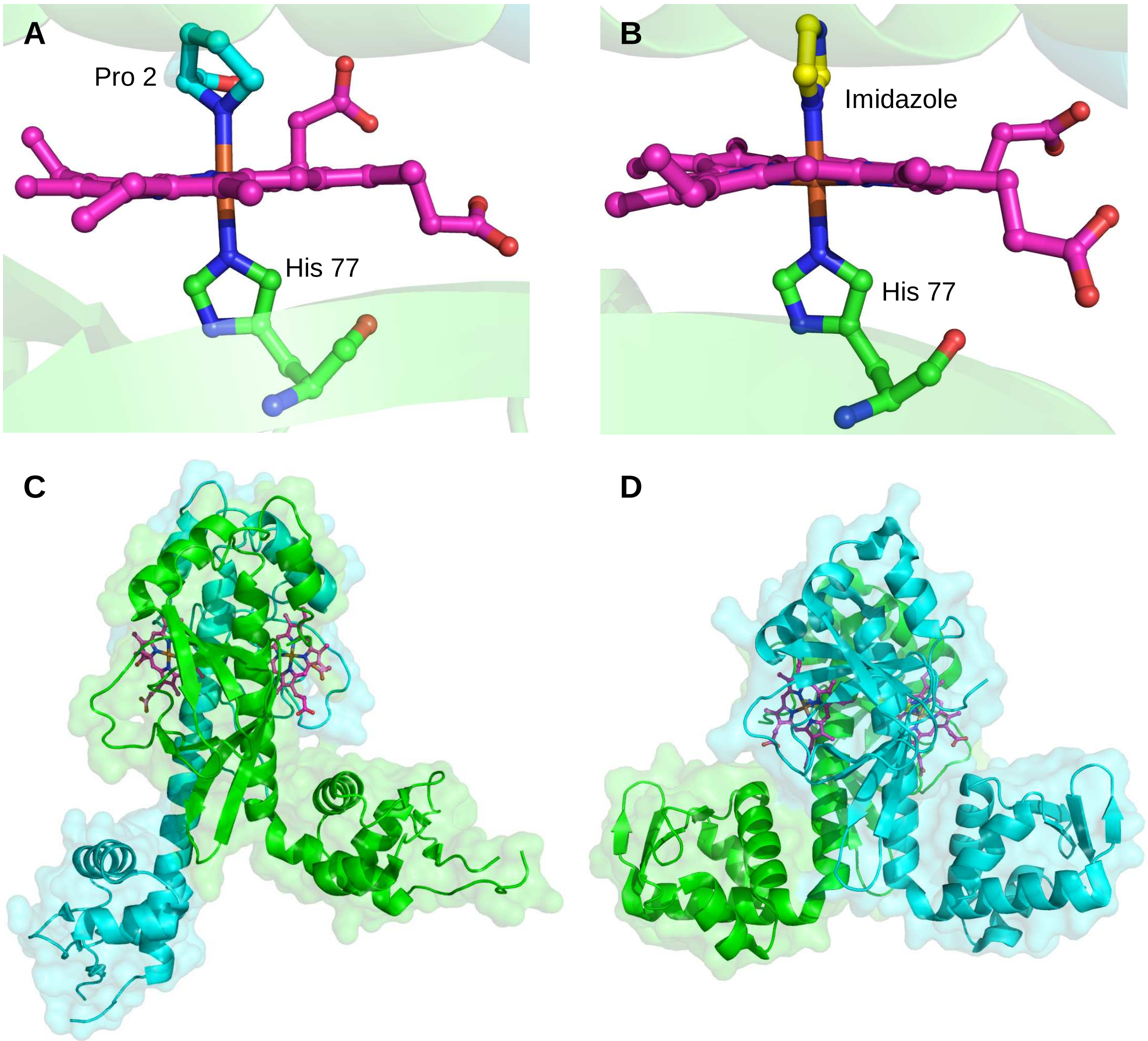


Figure 1.9 – The coordination of the heme groups and overall structures for the inactive and imidazole bound forms of CooA (PDB ID: 1FT9 **[Lanzilotta *et al* 2000]** (inactive) & 2FMY **[Komori *et al* 2007]** (imidazole bound)). Shown are; **(A)** the His-Pro coordinated heme from the inactive CooA, **(B)** the His-Imidazole coordinated heme from the imidazole bound CooA, **(C)** and **(D)** the overall structures of inactive and active CooA respectively, demonstrating the large conformational change caused by changes in heme coordination. Residues are coloured by chain ID (A=green, B=Cyan), hemes are coloured magenta and imidazole yellow.

Komori *et al* postulated the reason for the inactivity of the imidazole bound from of CooA is likely to be due to hydrogen bonding between the carbonyl oxygen atom of Met5 and N^{ε} of imidazole, stabilising the complex and restricting movement that CO is likely to be able to induce **[Komori *et al* 2007]**. Although without a refined structure of a transcriptionally active form of CO bound CooA it is not possible to ascertain the specific interactions that occur.

Borjigin *et al* have crystallised a CooA mutant where Asn127 and Ser128 have been converted to Leucine **[Borjigin *et al* 2007]**. This form of CooA was also found to be transcriptionally inactive in the presence of CO, but as with the Komori structure, does contain a significant domain movement of approximately 20 Å.

1.1.4.6 Hemoproteins – Electron transfer

Electron transfer hemoproteins function either directly by mediating transport of electrons through them, as seen in the membrane bound cytochrome *bc*$_1$ complex, or by storing electrons in their heme groups and moving the entire protein, as seen in cytochrome *c*. Both the cytochrome *bc*$_1$ complex and cytochrome *c* occur widely in eukaryotic and prokaryotic respiratory and photosynthetic electron transfer chains, including the mitochondrial electron transport chain; where they are responsible for the transfer of electrons across the mitochondrial inner membrane which is coupled with the pumping of protons across the same membrane (the cytochrome *bc*$_1$ complex), and the transport of electrons from complex III to complex IV of the electron transport system (cytochrome *c*) **[Crofts *et al* 2006]**.

The cytochrome *bc*$_1$ complex (Figure 1.10) is a dimeric membrane bound protein complex that contains three catalytic subunits in each monomer; a cytochrome *b* protein with two B-type hemes (one high spin and one low spin), a cytochrome *c*$_1$ protein with one C-type heme and Rieske iron sulphur protein with one iron sulphur cluster. The complex is responsible for the transfer of electrons between two mobile electron carriers across the mitochondria inner membrane; from ubiquinol (QH_2), located in the matrix, to cytochrome *c*, located in the inner membrane space. This movement of electrons also creates a proton motive force, capable of driving ATP synthesis **[Trumpower 1990]**. This electron transport and proton motive force occur via a Q-cycle mechanism. In this mechanism two separate ubiquinone binding sites, called Q_o (quinoloxidising site, located on the inner membrane space side of the membrane) and Q_i (quinonereducing site, located on the matrix side of the membrane) are responsible for feeding electrons into the system from QH_2. The first electron of QH_2 is transferred to a soluble cytochrome *c* electron carrier via the iron sulphur cluster of the Rieske iron sulphur subunit and C-type heme of the cytochrome *c*$_1$ subunit. The second electron is transferred from the Q_o site to the low spin B-type heme. From heme it moves within the membrane to reduce the high spin B-type heme, which in turn reduces a ubiquinone molecule bound to the Q_i site. During one complete Q-cycle, one molecule of ubiquinol is oxidised to ubiquinone, two molecules of cytochrome *c* are reduced, two protons are consumed on the matrix side of the membrane and four protons are released on the inner membrane space side of the membrane, thus generating a proton motive force **[Rich 2004]**.

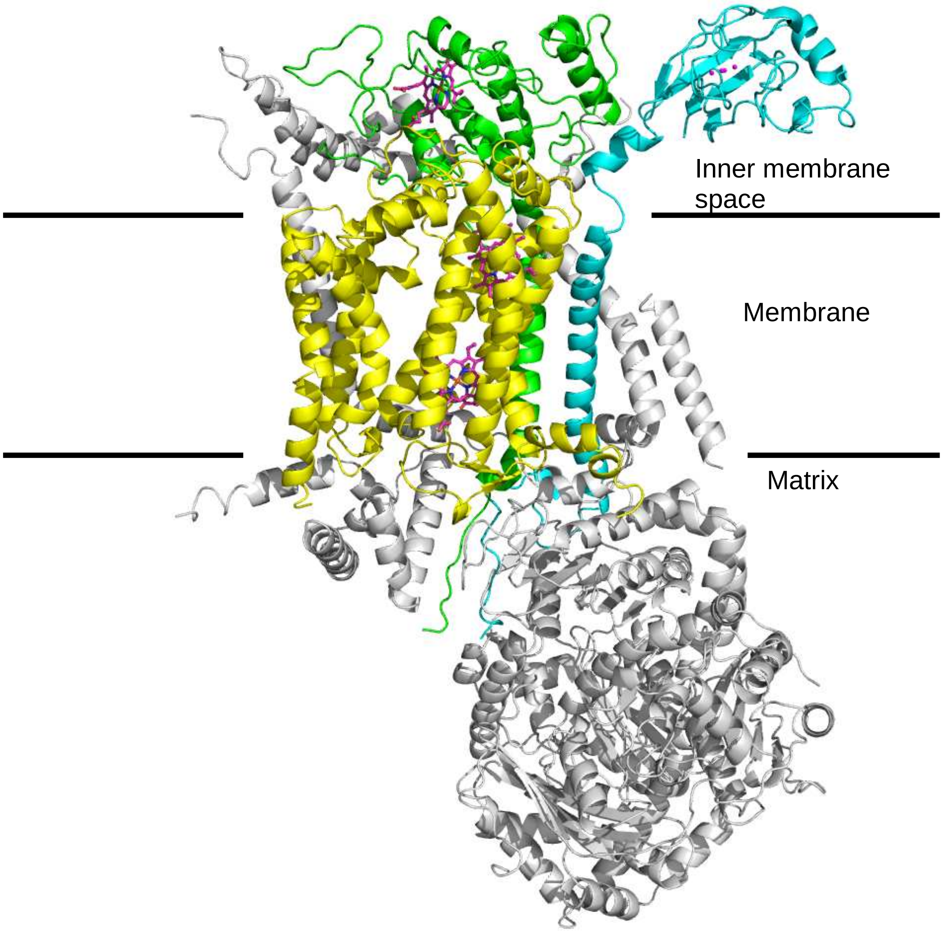


Figure 1.10 – A monomer of the Bovine cytochrome bc1 complex (PDB ID: 1BE3 **[Iwata *et al* 1998]**). The three catalytic subunits are coloured cyan for the Rieske iron sulphur subunit, green for the cytochrome *c*1 subunit and yellow for the cytochrome *b* subunit. The cofactors are coloured magenta and the remaining protein grey.

As previously mentioned, cytochrome *c* can act as an electron transport protein involved in the electron transport chain of many eukaryotic and prokaryotic respiratory and photosynthetic electron transfer chains. It is responsible for transferring electrons from complex III (the cytochrome bc_1 complex) to complex IV (a cytochrome *c* oxidase) of the electron transport chain. At complex IV cytochrome *c* oxidase removes four electrons from four molecules of cytochrome *c* and transfers them to dioxygen (O_2), producing two molecules of water, this reaction is coupled with a proton motive force that pumps four protons across the membrane from the matrix to the inner membrane space (in the case of the electron transport chain from mitochondria) **[Stiburek *et al* 2006]**.

1.1.4.7 Hemoproteins – Channel proteins

An example of a channel hemoprotein is the calcium dependant BK channel, a transmembrane protein responsible for the control of trafficking potassium ions (K^+) across the membrane **[Poulos 2006]**. The BK channel contains a conserved heme binding motif, it has been found that the binding of hemes to this motif profoundly inhibits the K^+-transporting action of the protein by inhibiting the mechanisms that open the channel **[López-Barneo and Castellano 2005]**. A proposed mechanism for this action suggests that the binding of heme to the BK channel alters the conformation of the protein, preventing the activating calcium ions (Ca^{2+}) from binding with the protein, therefore causing it to stay closed **[Horrigan *et al* 2005]**. The presence of BK channels in mitochondria (the site of heme synthesis) and the potential for heme regulation of these channels **[Tang *et al* 2003]** provides more evidence for the self regulating ability of hemes.

1.1.5 Heme motifs

The covalent attachment of hemes to their apoproteins enables them to be tightly bound, increasing the stability of the protein and allowing clusters or chains of hemes to be formed, which is thought to allow for fast electron transfer between the heme groups **[Page *et al* 2003]**. This is due to the protein folding in such a way that the edge-to-edge distances between heme groups and other electron donors/acceptors (e.g. radical forming amino acids such as tyrosine or tryptophan) are kept to a minimum, allowing for faster electron transfer, since smaller distances result in lower activation energies and faster rates for electron transfer **[Moser *et al* 2006]**.

These heme clusters can be highly conserved, even in proteins that are totally unrelated in amino acid sequence or polypeptide fold. In multiheme proteins many hemes are found to be arranged relative to neighbouring hemes in characteristic ways. These commonly-observed packings of hemes will be referred to in this thesis as heme pair motifs, with the most common pairs containing hemes in either an offset parallel arrangement (parallel stacking pair motif, Figure 1.11A) or perpendicular to each other (di-heme elbow motif, Figure 1.11B), previously observed by Inês *et al* **[Inês *et al* 2006]**.

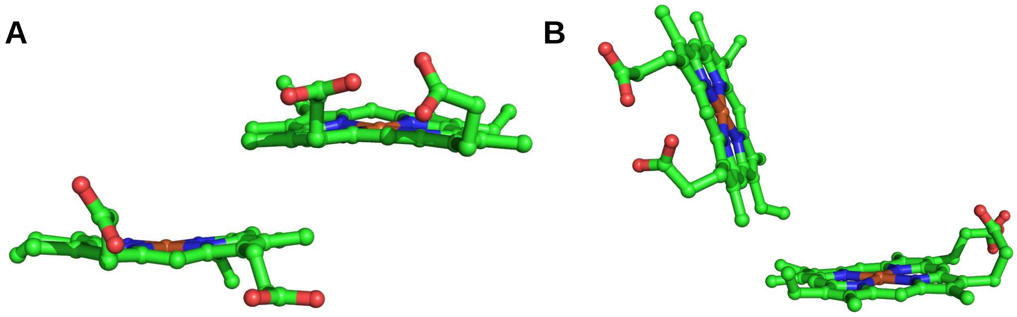


Figure 1.11 – Structures of the most common heme pair motifs **[Inês *et al* 2006]**. **(A)** The offset parallel arrangement of the parallel stacking pair motif. **(B)** The perpendicular arrangement of the di-heme elbow motif.

Inês *et al* have also shown that smaller heme motifs can be used to construct larger heme motifs; an example of this is the cytochrome c_3 family **[Inês *et al* 2006]**. The cytochrome c_3 family is populated by hemoproteins with four distinct motifs; cytochrome c_3, cytochrome c_7, 9-heme cytochrome and 16-heme cytochrome *c*. Proteins incorporating the c_3 motif are involved in intramolecular electron transfer and contain a tetraheme domain (Figure 1.12A). The c_7 motif consists of a 3-heme domain (Figure 1.12A), with a heme substructure homologous to three of the four hemes in the c_3 motif and contains proteins with metal ion reducing properties. The 9-heme motif contains two repeats of the c_3 motif, with a linking heme between them (Figure 1.12B) and is populated by proteins that are believed to take part in the periplasmic assembly of proteins involved in the mechanism of hydrogen cycling, receiving electrons from tetraheme c_3 proteins. The 16-heme cytochrome *c* motif is an amalgamation of the other motif of the c_3 like family, containing a 9-heme cytochrome motif, bound to a c_3 motif, bound to a c_7 motif (Figure 1.12C).

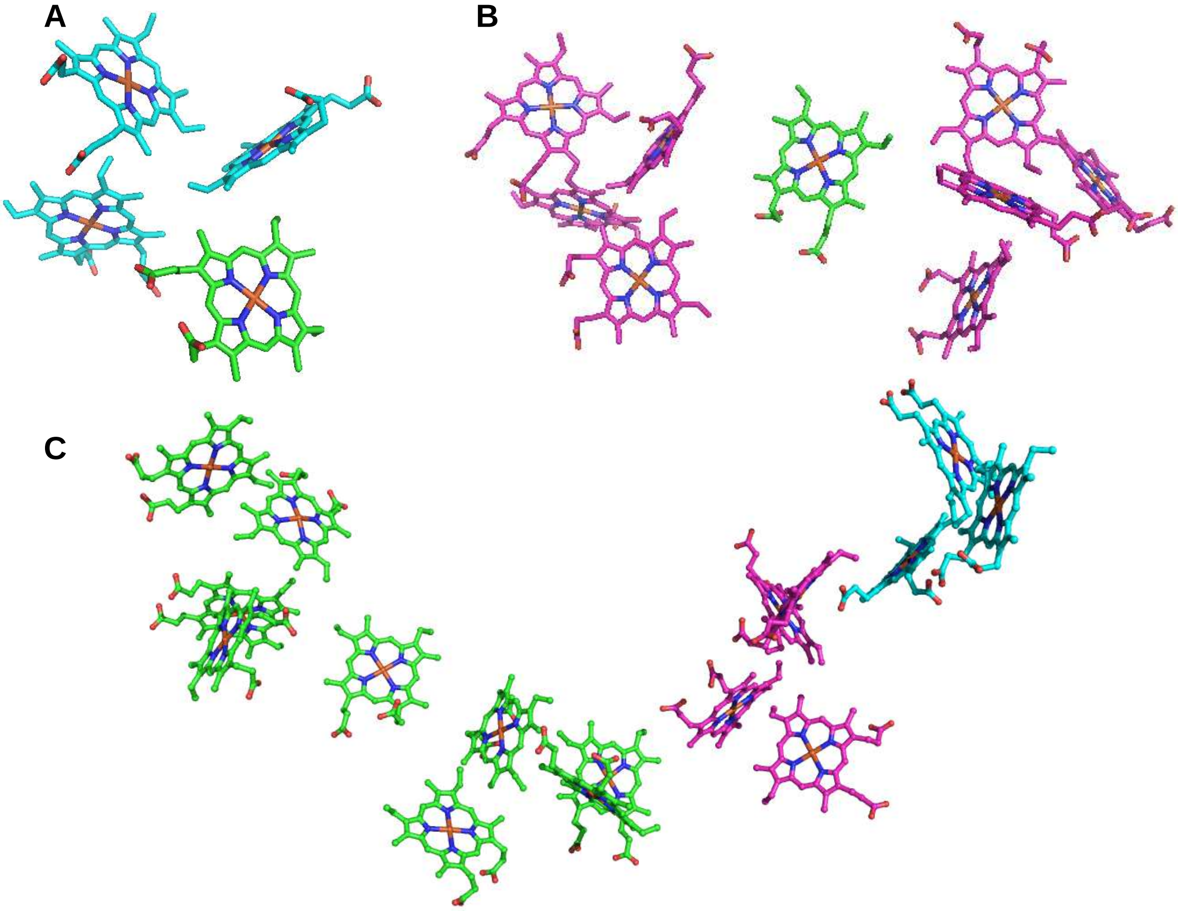


Figure 1.12 – The heme substructure of the motif of the cytochrome c_3 family, showing; **(A)** The tetraheme c_3 motif, with the c_7 motif contained within it coloured cyan, **(B)** the 9 heme domain (with the c_3 motifs coloured magenta) and **(C)** the 16 heme motifs (with the 9 heme motif coloured green, the c_3 motif magenta and c_7 motif cyan)

1.2 Copper proteins

1.2.1 Introduction

Copper is an essential trace metal and cofactor for many proteins and it's involved in many important cellular processes, such as enzymatic reactions **[Pufahl *et al 1997*]** and electron transport **[Brown *et al* 2002]**. This is largely because copper is able to exist in multiple oxidation states *in vivo* **[Rae *et al* 1999]**. For example, copper plays an important role in the action of the Cu, Zn superoxide dismutase enzyme, where it acts as an electron carrier **[Pufahl *et al 1997*]**. This enzyme is responsible for breaking down superoxide O_2^- into O_2 and H_2O_2, a vital process since an excess of superoxide species within cells has been linked to oxidative damage to proteins, lipids and DNA, as well as an acceleration of age-dependent skeletal muscle atrophy **[Muller *et al* 2006]**. However, copper is also potentially extremely toxic due to the formation of reactive free radical species via the Fenton reaction **[Halliwell and Gutteridge 1990]**. Therefore, cellular processes are needed to control the concentration and oxidation state of copper within the cell.

1.2.2 Copper homeostasis systems

Saccharomyces cerevisae Atx1 and the transporting P-type ATPases associated with it was the first copper homeostasis system to be identified **[Lin and Culotta 1995]**. Atx1 is required for the transport of Cu(I) into the *trans*-Golgi network, responsible for trafficking protein towards the cell wall and beyond in eukaryotic cells. Atx1 delivers copper to the Cu(I) transporting ATPase Ccc2, which in turn transfers the Cu(I) across the membrane into the *trans*-Golgi network. From here the Cu(I) is incorporated into the milti-copper oxidase Fet3, which is located on the cell membrane and required for high-affinity iron uptake into the yeast cell **[Askwith *et al* 1994]**. Atx1 deletion mutants of *Saccharomyces cerevisae* show a deficiency in iron, due to the lack of copper incorporation into Fet3 **[Lin *et al* 1997]**.

The *cop* operon is an example of a copper chaperone system that plays a role in copper homeostasis within the bacterial cell. This operon has been well studied in the Gram-positive bacterium *Enterococcus hirae* and *Bacillus subtilis.* The *E.hirae cop* operon consists of four genes; *copA* and *copB* that code for P-type ATPases, *copY* that codes for the transcriptional regulator of the *cop* operon and *copZ* that codes for an Atx1-like copper chaperone. The P-type ATPases are copper pumps, proposed to be involved in Cu(I) uptake into the cell (CopA) and Cu(I) secrction out of the cell (CopB) **[Multhaup *et al* 2001]**. The transcriptional regulator CopY exists as a homodimer under conditions of normal copper concentration, in which it binds to two distinct 28 basepair sequences in the promoter region of the *cop* operon, thus inhibiting transcription of the genes **[Strausak and Solioz 1997]**. The DNA binding conformation of the CopY homodimer is stabilized by a Zn(II) ion bound by four cysteine residues in a tetrahedral environment, under conditions of elevated Cu(I) this Zn(II) ion is displaced by Cu(I), resulting in the conversion of CopY from a DNA-binding form to a non-binding form, thus releasing it from the promoter region and inducing transcription of the *cop* operon **[Cobine *et al* 2002]**. The copper chaperone CopZ is required for the delivery of Cu(I) to CopY **[Cobine *et al* 1999]** and has also been shown to interact with the Cu(I) uptake ATPase CopA, which is thought to result in Cu(I) loading of CopZ **[Multhaup *et al* 2001]**.

The proposed mechanism for this system is as follows; the extracellular reductase CorA reduces Cu(II) to Cu(I) **[Solioz and Stoyanov 2003]**, Cu(I) is taken into the cell through CopA where it is transferred to CopZ, from here Cu(I) is donated to the CopY repressor bound to the promoter region of the cop operon, releasing the Zn(II) from CopY, allowing CopY to detach from the promoter and for transcription of the *cop* operon to commence (Figure 1.13) **[Magnani and Solioz 2005]**.

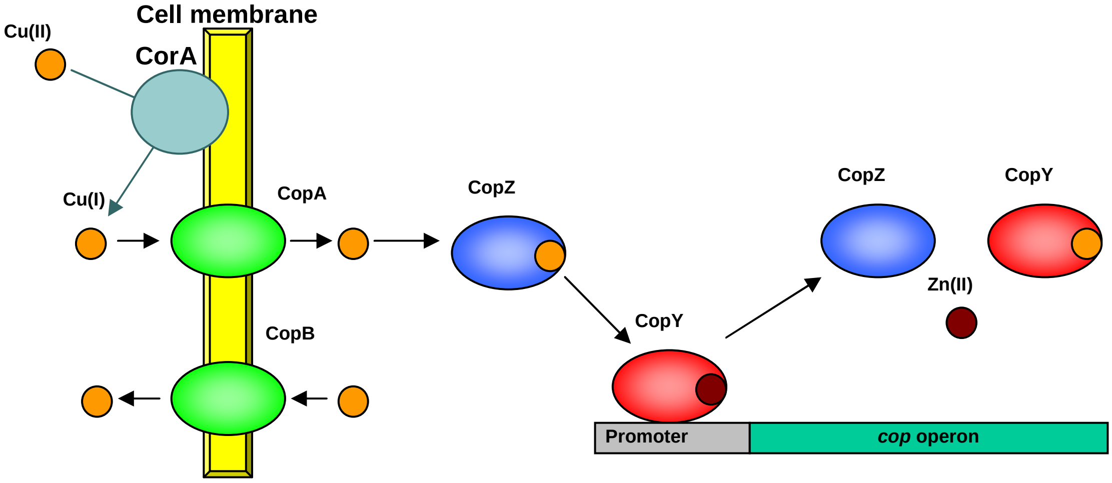


Figure 1.13 – A schematic of the *Enterococcus hirae cop* operon. The extracellular reductase CorA reduces Cu(II) to Cu(I) **[Solioz and Stoyanov 2003]**. Cu(I) is taken into the cell via the P-type ATPase CopA, Cu(I) is then transferred to the CopZ chaperone, which in turn donates it to the transcriptional regulator CopY. This releases the Zn(II) ion bound to CopY, releasing CopY from the promoter region and thus allowing transcription of the *cop* operon **[Magnani and Solioz 2005]**

The cop operon of *B.subtilis* consists of only two genes; *copA* that codes for a Cu(I)-transporting P-type ATPase and *copZ* that codes for an Atx1-like copper chaperone. CsoR has been identified as one of the Cu(I)-sensing repressors that regulate transcription of the *cop* operon **[Smaldone and Helmann 2007]**, YcnK has also been identified as a transcriptional regulator and YcnJ has been identified as the protein responsible for the influx of copper into the cell **[Chillappagari *et al* 2009]** (Figure 1.14). Inactivation of CopA led to an enhanced sensitivity to environmental copper, suggesting that CopA is responsible for Cu(I)-export in *B.subtilis* **[Radford *et al* 2003]**. Inactivation of

CopZ also resulted in an increase in copper sensitivity, it also resulted in a significant decrease in cellular copper, suggesting that CopZ may act as a cytoplasmic store of Cu(I) under normal conditions **[Radford *et al* 2003]**.

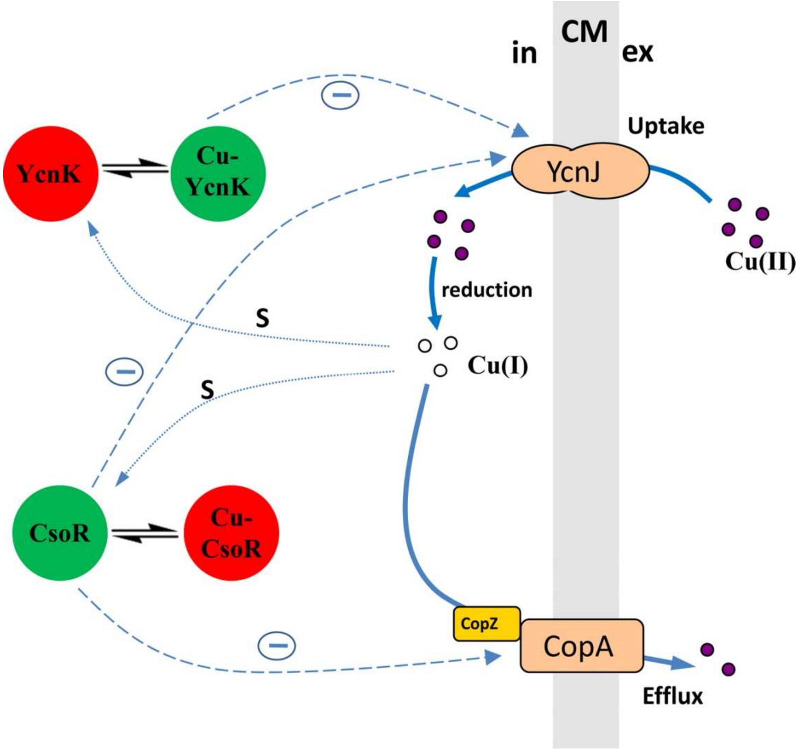


Figure 1.14 – A schematic for copper homeostasis in *Bacillus subtilis*. Cu(II) is taken into the cell via YcnJ, where it is reduced to Cu(I). Depending on their association with copper, the transcriptional repressors YcnK and CsoR are either activated (Green) or inactivated (Red) by this influx of copper. Resulting in the increase or decrease in CopZ production to remove copper via CopA. The negative regulation of components (–) is indicated with dashed arrows. Copper sensing (s) is indicated with dotted arrows. CM = cytoplasmic membrane; in = intracellular and ex = extracellular. This figure was taken from **[Chillappagari *et al* 2009]**.

1.2.3 Atx1-like copper chaperone structures and copper transfer mechanism

Atx1 is a 72 residue polypeptide that forms a ferredoxin-like βαββαβ-fold, where the antiparallel β strands form a β-sheet, on which the two α helices are found in an open-faced β-sandwich formation **[Rosenzweig *et al* 1999]**. The structures of the CopZ copper chaperones from *E.hirae* and *B.subtilis* share the same ferredoxin-like βαββαβ-fold and therefore have homologous structures (Figure 1.15), with c-α RMSDs of 1.7 and 2.0 Å with the Atx1 structure respectively **[Wimmer *et al* 1999, Banci *et al* 2001]**.

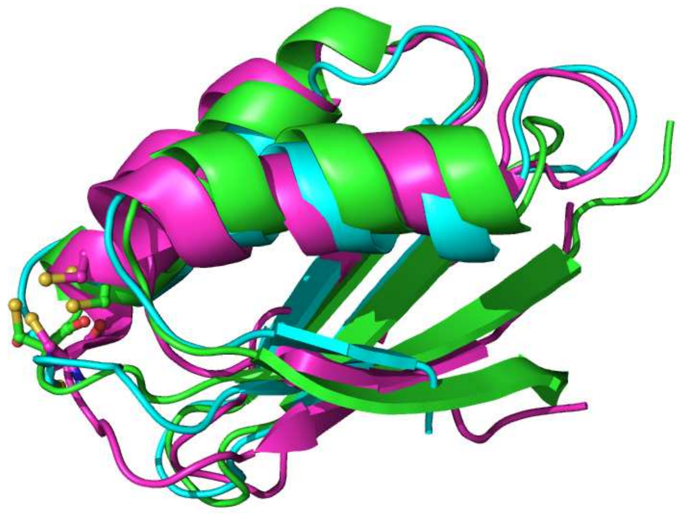


Figure 1.15 – Structures for the copper chapperones; Atx1 from *Saccharomyces cerevisae* (green), CopZ from *Enterococcus hirae* (cyan) and CopZ from *Bacillus subtilis* (magenta), showing the homology between the three structures. The polypeptide chain is displayed in cartoon format and the copper binding cysteine residues in stick format.

A mechanism for how copper is transferred between Atx1 and its target P-type ATPase Ccc2 has been proposed (Figure 1.16) **[Pufahl *et al* 1997]**, as the structures of the copper chaperones from the CopZ structures are so similar, it is reasonable to assume that CopZ would employ a similar technique when transferring copper to its target P-Type ATPase CopA..

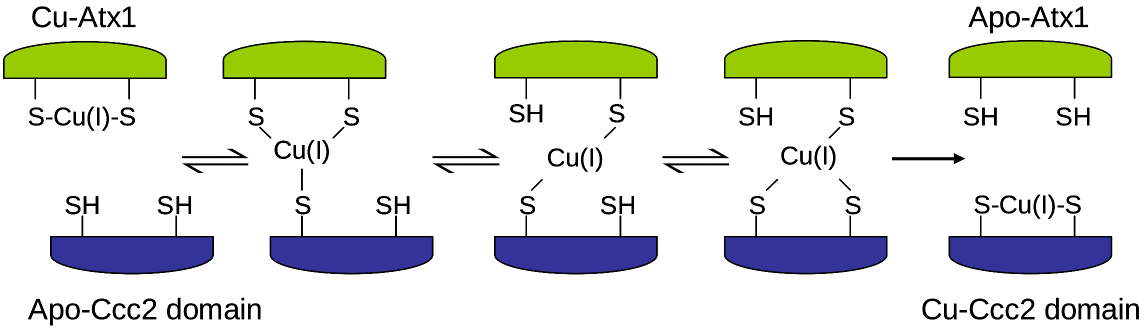


Figure 1.16 – The proposed mechanism for copper transfer between Atx1 and Ccc2, involving two and three-coordinate Cu(I) bridge intermediates. This mechanism is likely to hold true for copper transfer between CopZ and CopA.

1.2.4 Copper related diseases

Menkes disease and Wilson disease are two conditions that rise from disruption in copper homeostasis processes. In humans, under normal conditions, the copper chaperone HAH1 binds and delivers copper to P-type ATPases that are located in the membrane of the trans-Golgi network and deliver copper to the secretory pathway for metalation of cuproenzymes. The P-type ATPases that are associated with the Menkes and Wilson diseases are ATP7A and ATP7B respectively **[Klomp *et al* 1997, Hung *et al* 1998]**. Under normal conditions, when intracellular copper concentration increases these proteins export excess copper outside the cell **[Lutsenko *et al* 2007]**. It is mutations in these proteins that lead to Menkes and Wilson diseases **[Bull and Cox 1994]**, due to the bodies inability to distribute copper correctly.

1.3 Protein structure prediction

1.3.1 Introduction

The overall aim of protein structure prediction is the creation of three-dimensional protein structures from their amino acid sequence, in essence predicting a proteins tertiary structure from its primary structure. Protein structure prediction has the potential to be useful for processes such as, designing new drugs or creating novel enzymes, giving it high importance in both the medical and biotechnology industries. A reliable method for computationally predicting protein structure has also become more important in recent years with the completion of large scale DNA sequencing projects, such as the human genome project. These have produced massive amounts of sequence data, but as yet yielded relatively few experimentally determined protein structures due to the time consuming and relatively expensive nature of X-ray crystallography and NMR spectroscopy, as well as these methods not being successful with all proteins, especially membrane proteins **[Qain *et al* 2007, Lacapère *et al* 2007]**.

The task of creating a protein structure from an amino acid sequence is not an easy one and is made more difficult by the sheer number of possible protein structures for any given sequence, a limited understanding of how the amino acid sequence folds into a native protein, and the massive amounts of computing power needed for some prediction methods. The current methods for structure prediction fall into two main categories; comparative protein modelling and *de novo* protein modelling. In basic terms comparative protein modelling or homology modelling uses previously solved structures as templates for structure prediction and *de novo* protein modelling attempts to build three-dimensional protein models “from scratch” based on physical principles.

1.3.2 Homology modelling

Homology modelling works on the principle that the structural conformation of a protein is more highly conserved than its amino acid sequence, therefore subtle changes in sequence identity result in only minor changes in the overall structure **[Lesk and Chothia 1986]**. Homology modelling software takes an amino acid sequence as an input and uses it to search for homologues of that sequence from proteins with experimentally solved structures in a structural database, such as the Protein Data Bank (PDB) **[Berman *et al* 2000]**. This search for conserved sequences is carried out by sequence alignment software such as BLAST **[Johnson *et al* 2008]**; BLAST is able to reliably identify protein segments with a sequence identity greater than 30%. For lower sequence identities, methods such as PSI-BLAST and hidden Markov models, as used in the SAM (Sequence Alignment and Modelling) package **[Karplus *et al* 1998]**, provide a more reliable result due to their use of profile analysis.

Once the structural segments are identified, a model is assembled and assessed for its accuracy. In the case of models where an experimentally defined structure exists, accuracy is measured by comparing the prediction with the experimentally refined model using a root mean square deviation (RMSD), which measures the distance between corresponding atoms in the two superimposed structures. However, RMSD analyses alone are not full proof, a small change in just one part of the protein, such as a hinge joining two domains or a loop, can cause two similar structures to appear very different. An alternative method is the Local Global Alignment (LGA) method **[Zemla 2006]** that uses the longest continuous segments (LCS) and global distance test (GDT) algorithms to determine the accuracy of the modelled structure **[Zemla 2003]**. The LCS algorithm identifies the longest continuous segments of residues in the target deviating from the model by not more than specified α-carbon RMSD cut-off. The GDT algorithm identifies in the target the sets of residues deviating from the model by no more than a specified α-carbon distance cut-off using many different superimposed structures.

In cases where there is no experimentally defined structure, statistical potentials or force field-based energy calculations must be used to assess the accuracy of the model. Statistical potentials are based on observed residue-residue contact frequencies among proteins of known structure in the PDB, assigning an energy score to each possible pairwise interaction between amino acids, these pairwise interaction scores are combined into a single score for the entire model **[Melo *et al* 2002]**. Force field based energy calculations aim to assess the atomic interactions that are physically responsible for the stability of the protein structure, these calculations are performed using a molecular mechanics force field and take into account covalent, van der Waals and electrostatic interactions **[Moult 1997]**. Since the force fields are firmly based on the principles of physics, the force field based analyses for model assessment has the potential to be highly accurate, although more accurate force fields will be needed before this is the case **[Misura 2005]**. A further problem is that many proteins are too large for these calculations to be practical with the current algorithms and levels of available computing power, in these cases statistical potentials are a viable alternative.

1.3.2.1 Statistical potentials

Statistical potentials are calculated from known protein structures and are used to quantify the observed preference for the different residues or atom types to be exposed to the solvent, or to form pairwise or higher order interactions with each other. Statistical potentials can be used in; assessment of experimentally determined or computationally predicted proteins structures, *de novo* protein structure prediction **[Chiu and Goldstein 2000]**, threading **[Panchenko *et al* 2000]**, detection of native-like protein confirmations **[Vendruscolo *et al* 2000]** and the prediction of protein stability **[Gilis and Rooman 1996]**.

There are a number of methods for calculating statistical potentials including; distance-dependent, contact, accessible surface and main chain dihedral angle potentials. Each method calculates the occurrences of their given variable (pairwise contact, φ/ψ angle, etc) by statistical examination of the native variables present in the database of structures contained within the PDB **[Berman *et al* 2000]**. Distance-dependent potentials can be calculated using the following equation **[Melo and Feytmans 1997]**:

$$E_k^{ij}(l) = RT\ln[1 + M_{ijk}\sigma] - RT\ln[1 + M_{ijk}\sigma\frac{f_k^{ij}(l)}{f_k^{xx}(l)}]$$

Where $M^{ij}_k\sigma$ is the number of occurrences for the interaction type pair ij separated by k residues in sequence, $f^{ij}_k(l)$ is the relative frequency of occurrence for the interaction type pair ij at sequence separation k in the class of distance l, and $f^{xx}_k(l)$ is the relative frequency of occurrence for all the interaction type pairs at sequence separation k in the class of distance l.

Contact potentials can be calculated using the following equation **[Melo and Feytmans 1997]**:

$$E(i, N_j) = -RT\ln\frac{N_{obs}(i,k)}{\sum_k N_{abs}(i,k)/N_{cbin}}$$

Where i is the interaction types (amino acids or binary profiles), N_i is the contact number of the interaction centre i. $N_{obs}(i,k)$ is the number of observed contacts of interaction centre i with other interaction centres at k'th bin and N_{cbin} is the number of contact bins. A contact is defined by the Cα-Cα distance of two interaction centres within 8 Å. The number of contact bins is set to 25. In the rare occasions of more than 25 contacts, the statistics is included in the bin for 25 contacts **[Melo and Feytmans 1997]**.

The accessible surface of an interaction centre is defined as the number of interaction centres within a sphere around the central interaction centre, the distance range of the potential is used for the radius of the sphere. Accessible surface potentials can be calculated using the following equation **[Gilis and Rooman 1996]**:

$$E^i(l) = RT\ln[1 + M_i\sigma] - RT\ln[1 + M_i\sigma\frac{f^i(r)}{f^i_{ref}(r)}]$$

Where M_i is the frequency of the interaction centre type i in all the burial classes $f^{\,i}(r)$ is the relative frequency of occurrence of the interaction centre type I in the burial class r and $f^{\,i}{}_{ref}(r)$ is the reference state **[Gilis and Rooman 1996]**.

The dihedral angle potential can be calculated using the following formula **[Melo and Feytmans 1997]**:

$$E(i,\phi_i,\varphi_i) = -RT\ln\frac{N_{obs}(i,\phi_i,\varphi_i)}{\sum_{\phi_i,\varphi_i} N_{obs}(i,\phi_i,\varphi_i)/N^2_{bin}}$$

Where i is the amino acid type, Φ_i, Ψ_i are the torsion angles of the specific amino acid i. The torsion potential is the logarithm of the number of observed occurrence of the amino acid i at torsion angles of Φ_i, Ψ_i [$N_{obs}(i, \Phi i, \Psi i)$] normalized by the averaged occurrence. Each torsional angle is divided into 36 bins, therefore, N_{bin} is equal to 36 **[Melo and Feytmans 1997]**.

1.3.2.2 Errors in homology models

Large scale errors in protein structures created using homology modelling techniques tend to be a result of poor template selection or poor sequence alignment, removing human error from the equation, these problems do not have a straightforward solution as they are often the result of not having an available template structure, and can therefore only realistically be solved by large scale *de novo* modelling. Serious local errors in homology modelled protein structures frequently form where there are gaps in the template structure; these gaps are most common in loops. The modelling of loops not present in the template structures can be performed using database methods that work well with short loops (<5 residues), or *de novo* methods that can handle longer loops, but struggle with anything larger than 12 residues **[Rohl *et al* 2004, Xiang 2006]**. The other major source of local errors is the prediction of amino acid side chain conformations, this is partly due to the fact that many side chains in crystal structures are not in their optimal rotameric state as a result of crystal packing. Our current ability to accurately predict side chains is limited and mainly caused by misaligned residues and/or backbone shifts that need to be either accurately modelled in the initial prediction or refined simultaneously to improve side chain predictions **[Ginalski 2006]**.

1.3.2.3 Loop and Side chain prediction

The basic goal of loop prediction is to ascertain the conformation of a loop that is fixed at both ends by the protein backbone. Loop prediction can be made using either database or *de novo* methods. Database methods work by searching for segments of protein with known 3D structures that fit with the two exposed ends of the protein backbone, sequence similarity is then applied to determine which protein segment is the most likely loop. This method has been found to work adequately for loops up to 5 residues in length before the predictions become unreliable **[Fidelis *et al* 1994]**. The *de novo* method involves the generation of large numbers of randomly chosen candidate conformations, once generated an energy function that utilises CHARMM molecular mechanics force field **[MacKerell *et al* 1998]** is applied to find the most likely conformation. This method has been found to work adequately for loops up to 12 residues in length before the predictions become unreliable; however, it is thought that a more accurate energy function will lead to more accurate loop modelling **[Fiser *et al* 2000]**. Some of the highest accuracy has been achieved by Jacobson *et al*, who achieved a 1.0 Å RMSD deviation for 8 residue loops with a computer intensive approach that combined OPLS all-atom energy function, efficient methods for loop building and side-chain optimisation, and the hierarchical refinement protocol **[Jacobson *et al* 2004]**.

Accurate prediction of amino acid side chains is best achieved when the backbone structure itself is known to a high degree of accuracy. The majority of side chain predicting programs are based on rotamer libraries that contain the side chain torsional angles for the preferred conformations of specific side chains, this has become a valid method of prediction since computational prediction through energy functions is impractical due to the sheer number of possible conformations, and that the most frequently observed conformation tends to the be the most energetically favourable. These libraries have been improved by incorporating protein backbone data; backbone-dependant rotamer libraries use backbone φ and ψ angles to help determine side chain conformation. This is possible due to significant correlations found between side chain dihedral angle probabilities and backbone φ ψ angles **[Dunbrack and Karplus 1995]**. The major advantage of backbone-dependent libraries is they increase computer efficiency, since bad rotamers that clash with the backbone have already been removed. It has been shown that using a detailed rotamer library based on conformations taken from known structures, rather than idealised bond lengths and angles, can yield accuracies of 0.62 Å RMSD deviation for core residues **[Xiang and Honig 2001]**.

1.3.3 Threading

Threading is an alternative method of protein modelling. The essential difference between threading and homology modelling is that where homology modelling attempts to align a query sequence to a target sequence, threading attempts to align a query sequence to a structural segment or fold (Figure 1.17).

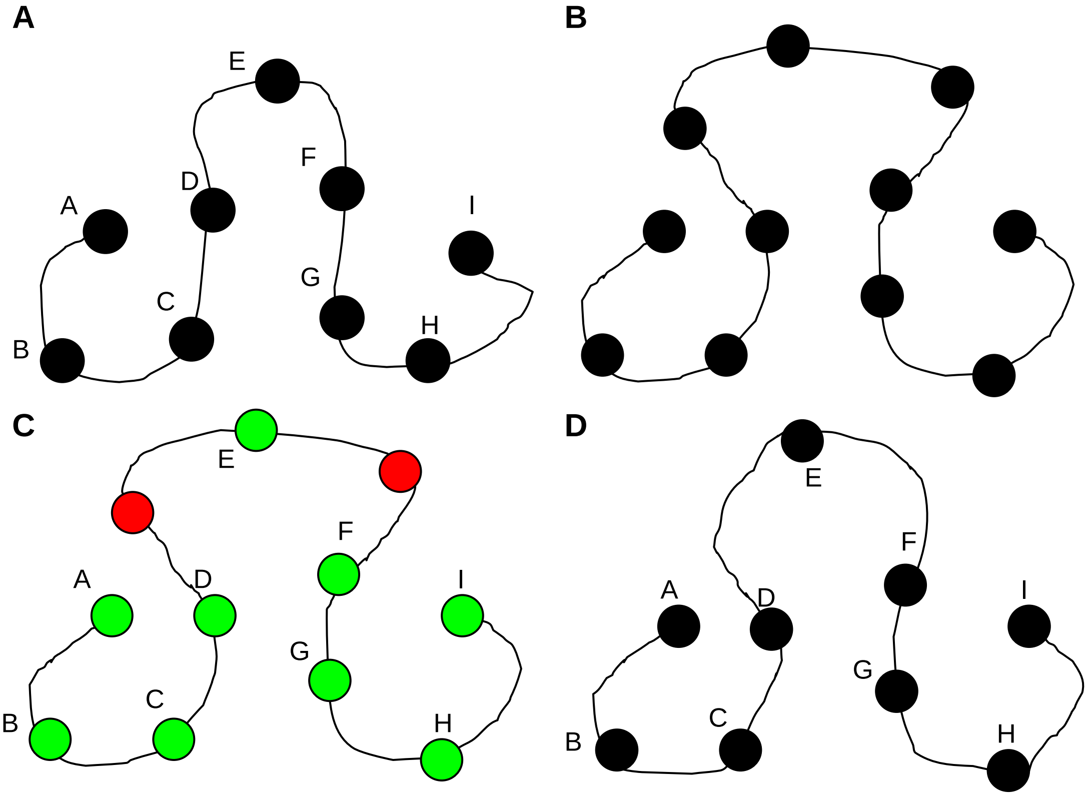


Figure 1.17 – The basic principle of threading. The sequence “ABCDEFGHI” fits a protein fold as shown in **(A)**, although this is unknown. **(B)** Is the template structure taken from a fold library that has been deemed the closest match for the target structure and the sequence will be threaded onto it. **(C)** Shows how the sequence best fits the template structure with matches coloured in green and gaps in red. **(D)** Shows the predicted structure from the threading method.

The rationale behind threading is based on the observation of the limited number of folds found in nature and that amino acids preference for different structural environments provides enough information to choose between these folds. The term “threading” was first introduced by Jones *et al* in 1992 **[Jones *et al* 1992]**. The basic principle is that a target sequence is threaded through the backbone structures of a collection of template proteins from a fold library and a “goodness of fit” score calculated for each based on an energy function, with the prediction with the lowest free energy value being taken as the result. Threading methods therefore incorporate characteristics from homology modelling (the sequence alignment aspect) and *de novo* modelling (predicting structure based on low-energy conformations in the target protein) to create their protein models, with the two main problems to overcome being; how to calculate the energy and how to "thread" a sequence through a fold. The energy of each threading alignment can be calculated as the sum of the energy of all pairwise interactions using the following equation **[Mirny *et al* 2000]**:

$$E^s = \sum_{ij=1}^{L} U(\xi_i, \xi_j) \Delta(r_i^s, r_j^s)$$

Where L is the length of the query sequence, s denotes alignment, and r^s_i is a coordinate of the *i*th group in this alignment (usually the α or β carbon atom). Δ corresponds to the cutoff distance for contact potential that determines which groups are interacting (this is usually 7.5-9 Å between α or β carbon atoms of the two interacting residues), ξ_i corresponds to the type of amino acid at the *i*th position in the query sequence and *U* is a 20x20 matrix of interaction energy parameters between all types of amino acids **[Mirny *et al* 2000]**. However, this is not the only method of calculating the free energy for a structural alignment, the potential of mean force is another type of interaction function for protein threading and can be calculated using the following formula **[Hayward 2006]**:

$$E(r) = -K_b T \ln \sum_i p(r, \xi_i)$$

Where *E(r)* is the potential of mean force, K_b is Boltzmanns constant, T is the temperature and $p(r,\xi_i)$ is the probability of amino acid *i* occurring at a distance of *r*.

Just as there are different methods for calculating the energy function, there are also different methods for aligning the query sequence with the target structure. The simplest method is to use protein sequence alignment between the query and target sequences, the problem with this method is its non-physical approach, i.e. it does not incorporate structural information, and that the observed sequence variability in otherwise similar structures can make the results unreliable. More sophisticated methods incorporate structural factors into their alignment predictions and can use pairwise interactions or mean force potentials to aid their predictions **[Xu and Xu 2000]**. These methods work by constructing a matrix which gives the score that every sequence residue would have if it were placed in each position of the template structure. A dynamic programming algorithm is then used to trace back through the matrix for the lowest energy pathway that keeps the query sequence intact, but is allowed to insert gaps if necessary, although any gaps inserted are subject to a gap penalty. The size of the gap penalty differs between alignment methods, but the general rule is that the penalties for inserting gaps into core structures (α-helices, β-sheets, etc) are much greater than the penalties for inserting gaps into turns and loops **[Lathorp and Smith 1996]**.

The first program to implement a threading method, called THREADER, was released in 1994 **[Jones 1994]**. At the first CASP (Critical Assessment of techniques for protein Structure Prediction) in 1995 it was the most successful method for fold recognition **[Lemer *et al* 1995]**. Over the years many more threading methods have been proposed **[Xu and Xu 2000, Rost *et al* 1997]**, with the main emphasis on finding more accurate alignment algorithms, utilising larger fold libraries or coming up with novel methods for energy calculations.

Despite advances in threading there are limitations in the method that are still causing problems. If there is an unknown fold in the query protein that does not appear in the fold library used by your specific threading method it is very unlikely that a successful model will be produced. Even predictions where the template structure is similar to the native structure can produce high energy models with a small "energy gap" (the difference in energy between optimal and random alignments, a large gap means a fully folded, low energy structure) **[Mirny and Shakhnovich 1998]**. It has also been suggested that more unique folds in the fold library can make detection more difficult, by increasing the likelihood of random errors **[Rost *et al* 1997]**. This conclusion was reached by testing a set of 89 proteins against three different fold libraries, with 723, 449 and 403 chains respectively. The percentage of correctly detected first hits was inversely proportional to the size of the dataset with accuracies of 29 %, 31 % and 33 %, respectively. Another current limitation in threading techniques is the complexity of the models created. Advances in computing should be able to allow more complex models that can take side-chain size, shape and charge into account; this would allow search models to eliminate templates that would produce unfeasible side-chain packing, improving search focus **[Lovell *et al* 2000]**.

1.3.4 *De novo* protein modelling

The aim of *de novo* protein modelling, also known as *ab initio* modelling, is to build three dimensional protein models from scratch, that is to say, based wholly on physical principles rather than direct comparisons with previously solved structures. *De novo* methods work by either attempting to mimic protein folding or by applying a stochastic method that investigates all possible solutions and uses global optimisation to find the structure with the lowest free energy.

Stochastic methods are likely to have limited success with current levels of computing power due to Levinthal's paradox **[Levinthal 1968]**, which observes that if a protein is folded by randomly attempting all possible conformations the time needed to do so would be astronomical due to the sheer number of possible conformations. For example, a protein made up of 150 amino acids would have around 10^{300} different conformations. Since in nature many small proteins fold spontaneously on a millisecond or even microsecond time scale, Levinthal proposed that a random conformational search does not occur in folding and the protein must, therefore, fold by a directed process.

Folding methods are a more likely source of *de novo* protein structure prediction, but are not without their problems that need to be tackled before they can become viable prediction methods, such as, the thermodynamic question of how an amino acid sequence forms the native protein structure from the interatomic forces acting on it **[Dill *et al* 2007]**. Although some progress has been made in this field, with the prediction of novel small proteins such as Top7 **[Kuhlman *et al* 2003]**, the key challenges, including better understanding of the relative strengths of intermolecular and solvation forces still remain. Other significant stumbling blocks are the efficiency of the algorithms used for folding calculations and the availability of the necessarily huge computing power needed to perform them. That is not to say there have not been successes. The IBM Blue Gene group were able to fold a 20 residue mini-protein “Trp-cage” to an accuracy of approximately 1 Å RMSD **[Pitera and Swope 2003]** using 92 nanoseconds of replica exchange molecular dynamics. Zagrovic *et al* **[Zagrovic *et al* 2002]** have been able to fold the 36 residue α-helical protein from the villin headpiece to an accuracy of 1.7 Å RMSD using the Folding@home distributed computing system **[Pande 2000]**. Both these structures were solved using Molecular Dynamics (MD). These successes do indicate the immense amount of computing power needed to fold even the smallest protein with any degree of accuracy, however, it is becoming clear that *de novo* predictions are no longer and insurmountable challenge.

1.3.4.1 Molecular Dynamics

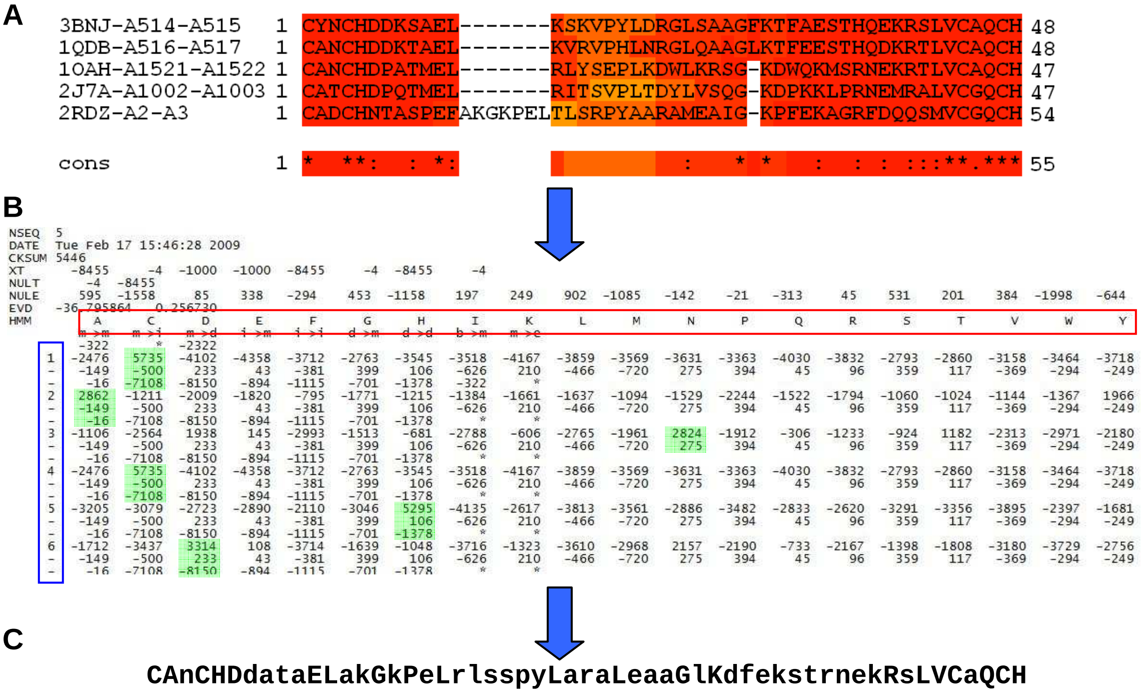
MD is a form of computer simulation where atoms and molecules are allowed to interact for a period of time under known laws of physics, providing a view of the motion of the atoms. In basic terms, the forces acting on each atom are calculated using “force fields” that take into account covalent, van der Waals and electrostatic interactions. The effects these forces will have on the position of the atoms is modelled using Newtown’s second law of motion (Force = mass x acceleration) and new positions for all atoms are calculated. The time step between each calculation (i.e. the amount of modelled time between each integration of Newton’s second law) is very small, generally in the order of femtoseconds (10^{-15} seconds), and the overall period of time modelled is typically in the order of picoseconds (10^{-12} seconds). The computing power needed to perform MD modelling is vast, to put it into context, the modelling of one nanosecond of real time life of a protein requires one million sets of calculations for each atom in the protein, meaning simulations of a few nanoseconds of a moderate side protein can take months to perform. The overall period of time modelled in MD calculations is important; to be able to draw valid conclusions the time span of the simulations must at least equal the time span of the kinetics of the natural process **[McDowell *et al* 2007]**.

Variations on MD techniques have been developed; one such is Replica Exchange Molecular Dynamics (REMD). REMD has been developed to overcome the problem of conventional MD simulation methods getting “trapped” in a large number of local minimum states **[Sugita and Okamoto 1999]**. To overcome this problem REMD performs a number of parallel simulations at different temperatures, with periodic exchanges of configuration. The effect of this is; if a particular simulation has become trapped at an energy minimum, it can escape via an exchange with a higher temperature conformation. A detailed description of the algorithms for this method can be found in Sugita *et al* **[Sugita and Okamoto 1999]**. Since each replica can be simulated using its own computer processor, REMD is well suited to running on parallel computers that can increase the speed of simulations, however this can also be a weakness, since REMD requires synchronisation between processors to perform the exchanges, therefore the simulation is limited to the speed of the slowest processor **[Rhee and Pande 2003]** making it unsuitable for large scale distributed computing, such as Folding@home **[Pande 2000]**. A solution to this problem has been put forward by Rhee and Pande **[Rhee and Pande 2003]**, who proposed having multiple replicas for each temperature level, eliminating the synchronisation needed in the original REMD method. This multiplexed-replica exchange molecular dynamics (MREMD) method is therefore able to make use of distributed computing, allowing Rhee and Pande **[Rhee and Pande 2003]** to simulate more than 200 microseconds of MD time, allowing their model protein (BBA5) to reach the folded state starting from the unfolded state, which was a first for an REMD-based simulation **[Rhee and Pande 2003]**.

1.4 Hidden Markov Models

1.4.1 Creation of Hidden Markov Models

A Hidden Markov model (HMM) is a statistical model for predicting the probability of a given sequence of events given prior knowledge of a past series of events and has application in speech recognition **[Rabiner 1989]**, gene prediction and sequence alignment. For example, given a multiple sequence alignment of a set of amino acid sequences, a HMM can be built from these sequences that gives the probability of finding a given residue type at each position of the sequence, based on the amino acid positions in the input sequences (Figure 1.18).



A

```
3BNJ-A514-A515   1 CYNCHDDKSAEL-------KSKVPYLDRGLSAAGFKTFAESTHQEKRSLVCAQCH 48
1QDB-A516-A517   1 CANCHDDKTAEL-------KVRVPHLNRGLQAAGLKTFEESTHQDKRTLVCAQCH 48
1OAH-A1521-A1522 1 CANCHDPATMEL-------RLYSEPLKDWLKRSG-KDWQKMSRNEKRTLVCAQCH 47
2J7A-A1002-A1003 1 CATCHDPQTMEL-------RITSVPLTDYLVSQG-KDPKKLPRNEMRALVCGQCH 47
2RDZ-A2-A3       1 CADCHNTASPEFAKGKPELTLSRPYAARAMEAIG-KPFEKAGRFDQQSMVCGQCH 54

cons             1 *  **:  : *:                 :   * *   :  : : :::**.*** 55
```

B

```
NSEQ  5
DATE  Tue Feb 17 15:46:28 2009
CKSUM 5446
XT      -8455     -4  -1000  -1000  -8455     -4  -8455     -4
NULT       -4  -8455
NULE      595  -1558     85    338   -294    453  -1158    197    249    902  -1085   -142    -21   -313     45    531    201    384  -1998   -644
EVD   -36.795864   0.256730
HMM        A      C      D      E      F      G      H      I      K      L      M      N      P      Q      R      S      T      V      W      Y
         m->m   m->i   m->d   i->m   i->i   d->m   d->d   b->m   m->e
         -322      *  -2322
     1  -2476   5735  -4102  -4358  -3712  -2763  -3545  -3518  -4167  -3859  -3569  -3631  -3363  -4030  -3832  -2793  -2860  -3158  -3464  -3718
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -369   -294   -249
     -    -16  -7108  -8150   -894  -1115   -701  -1378   -322      *
     2   2862  -1211  -2009  -1820   -795  -1771  -1215  -1384  -1661  -1637  -1094  -1529  -2244  -1522  -1794  -1060  -1024  -1144  -1367   1966
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -369   -294   -249
     -    -16  -7108  -8150   -894  -1115   -701  -1378      *      *
     3  -1106  -2564   1938    145  -2993  -1513   -681  -2788   -606  -2765  -1961   2824  -1912   -306  -1233   -924   1182  -2313  -2971  -2180
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -369   -294   -249
     -    -16  -7108  -8150   -894  -1115   -701  -1378      *      *
     4  -2476   5735  -4102  -4358  -3712  -2763  -3545  -3518  -4167  -3859  -3569  -3631  -3363  -4030  -3832  -2793  -2860  -3158  -3464  -3718
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -369   -294   -249
     -    -16  -7108  -8150   -894  -1115   -701  -1378      *      *
     5  -3205  -3079  -2723  -2890  -2110  -3046   5295  -4135  -2617  -3813  -3561  -2886  -3482  -2833  -2620  -3291  -3356  -3895  -2397  -1681
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -369   -294   -249
     -    -16  -7108  -8150   -894  -1115   -701  -1378      *      *
     6  -1712  -3437   3314    108  -3714  -1639  -1048  -3716  -1323  -3610  -2968   2157  -2190   -733  -2167  -1398  -1808  -3180  -3729  -2756
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -369   -294   -249
     -    -16  -7108  -8150   -894  -1115   -701  -1378      *      *
```

C

CAnCHDdataELakGkPeLrlsspyLaraLeaaGlKdfekstrnekRsLVCaQCH

Figure 1.18 – A multiple sequence alignment **(A)**, an extract from the hidden portion of the HMM **(B)** and the visible portion of the HMM **(C)** created from the sequence alignment. The area boxed in red highlights the residue types and the area boxed in blue highlights the residue number. The matrix of numbers refer to the scores calculated for the likelihood of finding each residue type at each position in the sequence, the highest scoring residues are highlighted in green.

1.4.2 Uses of HMMs

1.4.2.1 Using HMMs for protein analysis

The Protein families (Pfam) database is a biological example of the implementation of HMMs. Pfam is a comprehensive collection of nearly 12,000 conserved protein families and is used by; experimental biologists researching specific proteins, computational biologists who need to organise sequences, evolutionary biologists considering the origin and evolution of proteins and structural biologists for identifying interesting new targets for structure determination **[Finn *et al* 2010]**. HMMs are built for each protein family providing; information on the domains found within the family, a phylogenetic tree built from the sequences in the family and details of any structures that have been solved for proteins in each family.

Users are able to search the Pfam database with a target sequence of an unknown protein and Pfam will suggest possible functions and if present, structural templates, for the protein based on the domains it finds using its database of HMMs. It has been found that searching a database of HMMs rather than sequences gives more accurate results than a pairwise sequence search used by BLAST searches, and that the outputs from HMM based searches are easier to digest since the user is provided with a list of a few possible domains rather than a large number of homologous sequences, many of which will have the same domain **[Sonnhammer *et al* 1998]**.

1.4.2.2 Using HMMs for gene prediction

Advances in gene sequencing techniques has led to an increase in the volume of genetic data needing interpretation. Most commonly, genes have been identified by homology-based methods such as BLASTX **[Altschul *et al* 1990, Meyer *et al* 2008]**, however, these methods use searches against known databases, meaning they are unable to predict novel genes.

HMMs can be used in improve gene prediction from sequencing data, Rho *et al* developed the novel gene prediction method FragGeneScan, which combines sequencing error models and codon usages in a HMM to improve the prediction of protein-coding regions **[Rho *et al* 2010]**. They compared the results of their method with the non-HMM based methods Glimmer **[Delcher *et al* 1999]** and metagene **[Noguchi *et al* 2006]** (no longer in use) and found their HMM-based method not only out preformed the existing techniques for identifying genes with existing homologues, but was also able to identify novel genes with no homologues in existing sequence databases **[Rho *et al* 2010]**.

Another study of gene prediction was carried out by Yao *et al* who analysed *ab initio* gene prediction by testing 5 programs for the discovery of maize genes **[Yao *et al* 2005]**. The 5 programs tested were; FGENESH **[Salamov and Solovyev 2000]**, GeneMark.hmm **[Lukashin and Borodovsky 1998]**, GENSCAN **[Burge and Karlin 1997]**, GlimmerR **[Salzberg *et al* 1999]** and Grail **[Xu and Uberbacher 1997]**, of which three used HMM based methods (FGENESH, GeneMark.hmm and GENSCAN). Each program was tested with 10 different genes and found that the HMM based methods gave more correct predictions than the non-HMM based methods, more specifically that FGENESH was the best performer of the five **[Yao *et al* 2005]**.

1.4.3 HMM software

Several HMM software packages are available, the most popular of these being the HMMER **[Eddy 2998]** and SAM **[Hughey and Krogh 1996]** packages, these packages have the ability to build HMMs, search sequence databases using the HMMs and score the sequence hits they identify.

The HMMER package was developed chiefly by Sean Eddy and contains all the necessary HMM building and scoring programs relevant to homology detection. The HMMER package also contains a HMM calibration program (hmmcalibrate) that calibrates the HMM by scoring it against a set of random sequences and fitting an extreme value distribution to the resultant raw scores, this parameter is used to calculate E-values for alignments between the HMM and sequences of interest.

The SAM package was developed by the bioinformatics group at the University of California, and as with the HMMER package, contains all the necessary HMM building and scoring programs, as well as several scripts for running them. The SAM package does not contain a HMM calibration program, instead the HMM searching program calculates E-values directly using a theoretical function that takes the difference between the raw scores of the query sequence and its reverse as its argument for E-value calculation.

Studies have been carried out to compare these two packages **[Madera and Gough 2002, Wistrand and Sonnhammer 2005],**with the general consensus being that SAM is more sensitive when identifying HMM hits to sequences of interest, and that HMMER is faster (between 1 and 3 times faster) and has a more accurate scoring system.

1.5 Outline of the scope of the thesis

This thesis aims to improve upon existing protein structure prediction methodologies for multicofactor proteins, focusing on multiheme cytochromes to begin with, but the methods developed in this thesis are also likely to be applicable for other cofactor rich proteins that have sufficient structural data available. The prediction methods developed will be compared with existing tools available to the wider scientific community to demonstrate the specific advantages of the new methodology.

This thesis will also provide an insight into the copper transport mechanisms of the Atx1-like copper chaperone proteins by examining changes in monomer packing and copper cluster formation for the *Bacillus subtilis* copper chaperone CopZ in response to changing levels of available copper. This data will also be used to predict a structural complex for the mechanism of copper transfer between CopZ and the P-type ATPase CopA.

The final section of this thesis will examine the structural differences between native and product inhibited forms of the flavoprotein SoxF and what clues this information could provide to ascertain SoxFs role in the sulfur oxidising sox cycle of *Paracoccus denitrificans*. SoxF has been shown to have sulfide dehydrogenase activity and also shown to interact with the sox cycle intermediate transport complex SoxYZ, this thesis will attempt to solve the crystal structures of native and inhibited forms of SoxF and use them to hypothesise the nature and role of this interaction.

Chapter 2 - Analysis and Prediction of the Structures of Multiheme Cytochromes

2.1 Introduction

This chapter will analyse the distribution of heme groups in multiheme cytochromes with available crystal structures and examine the conserved heme motifs that arise from this analysis. The sequences and polypeptide structures that coordinate each of these conserved heme motifs will be extracted and aligned for the creation of Hidden Markov Models (HMMs). These HMMs will be used to search sequences of unknown structure to provide predictions for heme sub structure and templates for the modelling of polypeptide structure, where available. Test cases, which were proteins that had unpublished structural data available, were used to test this structure prediction methodology; the results were compared with predictions produced using existing homology modelling servers.

2.2 Materials and methods

2.2.1 Selection of multiheme proteins and heme pairs

To analyse the heme packing motifs found in multiheme proteins, the Protein Data Bank (PDB) **[Berman *et al* 2000]** was interrogated to extract the coordinates of interacting pairs of *c*-hemes from multiheme proteins whose structures had been determined by X-ray crystallography at resolutions ≤ 3.0 Å **[Walsh *et al* 2009]** where the distance between the two heme groups was at most 14 Å. The distance between hemes refers to the minimum distance between the closest carbon atoms from each heme porphyrin ring. A maximum distance of 14 Å was used because this has been determined to be the maximum distance for efficient electron transfer between hemes **[Page *et al* 1999]**. Two databases of multiheme proteins and heme pairs were generated; one excluding sequences at the 40 % sequence identity level and the other at the 90 % level. The 40 % cut-off value was chosen as it is the point at which sequence identity and biochemical function begin to converge **[Brylinski & Skolnick 2008]** and the 90 % cut-off value was chosen to give a larger non-redundant dataset containing more structural data.

The 40 % set contained 37 multiheme proteins with 152 heme pairs, and the 90 % set contained 56 multiheme proteins with 282 heme pairs. Both datasets were then reduced by removing those heme pairs where one or both of the iron-ligating residues from either heme was not provided by the histidine residue. The resulting 40 % *bis*-His dataset contained 27 proteins with 125 heme pairs and the 90 % *bis*-His dataset contained 40 proteins and 245 heme pairs. Heme motif packing analysis was carried out using the 40% *bis*-His ligated dataset, to ensure the results were not distorted by the presence of highly homologous sequences. A breakdown of the proteins that make up this dataset can be seen in table 2.1.

Table 2.1 – Proteins in the 40 % *Bis*-His ligated dataset

SCOP Family	**SCOP Domain**	**PDB ID**	**Hemes**	**Heme pairs**	**Protein structure**
Cytochrome *c*$_3$-like	Cytochrome *c*$_3$	1AQE	8	7	Homodimeric (4 hemes per monomer)
		1GYO	8	7	
		1J0P	4	6	Monomeric
		1WAD	4	6	
		2BQ4	*4*	*6*	
		2CY3	4	6	
		3CAO	4	6	
	Cytochrome *c*$_7$	1RWJ	3	1	
		3BXU	*6*	*3*	Homodimeric (3 hemes per monomer)
	Nine-heme cytochrome *c*	1OFW	18	18	Homodimeric (9 hemes per monomer)
Di-heme elbow motif	Periplasmic nitrate reductase subunit NapB	1JNI	4	1	Dimeric (2 hemes per monomer)
		1OGY	4	1	
	Hydroxylamine oxidoreductase, HAO	1FGJ	24	8	Homotrimeric (8 hemes per monomer)
	Cytochrome *c*554	1FT5	4	2	Monomeric
	Dimeric di-heme split-soret cytochrome *c*	1H21	4	2	Homodimeric (2 hemes per monomer)
	Cytochrome c nitrite reductase	3BNJ	*10*	*5*	Homodimeric (5 hemes per monomer)
		1OAH	10	5	
	Flavocytochrome *c*$_3$ (respiratory fumarate reductase), N-terminal domain	1M1Q	4	4	Monomeric
		1Y0P	4	4	
	Putative Cytochrome *c*	1SP3	8	8	
	Diheme c-type Cytochrome DHC2	2CZS	*2*	*1*	
	Crystal structure of E.coli nrfB	2OZY	*5*	*5*	
	Hexameric multiheme cytochrome c nitrite reductase	2OT4	*48*	*9*	*Homohexameric (8 hemes per monomer)*
Formate dehydrogenase	Formate dehydrogenase N from *E.coli*	1KQF	6	1	Heterononameric (2 hemes in chains C,F&I)
Other (not in SCOP database)	Quniol:Fumarate reductase from *Wolinella Succinogenes*	2BS2	4	1	Heterohexameric (2 hemes in chains C&F)
	NarGHI mutant NarI-K86A	1Y5I	4	1	
	DHC purified from *Rhodobacter sphaeroides*	2FWT	2	1	Monomeric

***N.B** Colours of SCOP families refer to those seem in the raw output (see appendix I-IV)*

2.2.2 Clustering techniques

Similar packings of *c*-heme pairs in multiheme cytochromes were detected using a JAVA program written for this purpose. The program performs two least squares superpositions for each combination of heme pairs in the dataset based on the non-hydrogen atoms of the porphyrin rings, superimposing the hemes in each of the two possible permutations. RMSD values were calculated for each superposition, the smaller of the two values being taken to reflect the similarity of the two heme packings. The resulting distance matrix populated with RMSD values was clustered using the R package **[Ihaka and Gentleman 1996]**, a system for statistical computation and graphics. The clustering method used was the single-link (also known as nearest neighbour) hierarchal clustering method, where the distance between groups is defined as the distance between the closest pair of objects from each group.

Clustering with R resulted in cluster dendrograms which were then interrogated by another JAVA program, written to identify heme pair clusters at different RMSD cut-offs. The RMSD cut-off used in the following analyses is 1.5 Å, this value was chosen to compensate for inaccuracies in the crystal structures and subtle variations in the packings of heme pairs. The resulting clusters are called *heme pair clusters* or *heme pair motifs*.

Once the heme pair clusters had been determined, heme triplet clusters (similar packings of three neighbouring heme groups) were identified by a further JAVA program that analyses the pair clusters and identifies heme triplets when it finds two heme pairs from the same protein in different clusters that share a single common heme. These heme triplets are grouped according to the heme pairs that make up each triplet cluster, i.e. a triplet cluster that contained heme pairs from pair clusters n and m would be known as triplet cluster n-m. A similar method was used to obtain heme quartet (4 heme) and quintet (5 heme) clusters by comparing the list of pair clusters with the list of triplet clusters to identify the quartets, and with the quartet clusters to identify the quintets, respectively. An RMSD value is also calculated for the alignment of heme motifs within each triplet, quartet and quintet cluster to ensure that only one structural alignment of hemes is present, i.e. that each cluster contains only one heme triplet/quartet/quintet motif structure and that each element of the cluster falls within the 1.5 Å RMSD cutoff.

2.2.3 Extracting sequence data from the heme motif clusters

Amino acid sequence data was then extracted for each heme motif cluster identified in section 2.2.2. The sequences between the iron ligating histidine residues (Figure 2.1A) and heme binding CXXCH motifs (Figure 2.1B) were selected. These sequences will be referred to as PD and PP for the sequences between iron ligating histidine residues and heme biding CXXCH motifs, respectively. The letters P (proximal) and D (distal) refer to the identity of the ligating histidine residue at each end of the sequence, the PD sequences run from a proximal to a distal histidine and PP sequences run from proximal to proximal histidines.

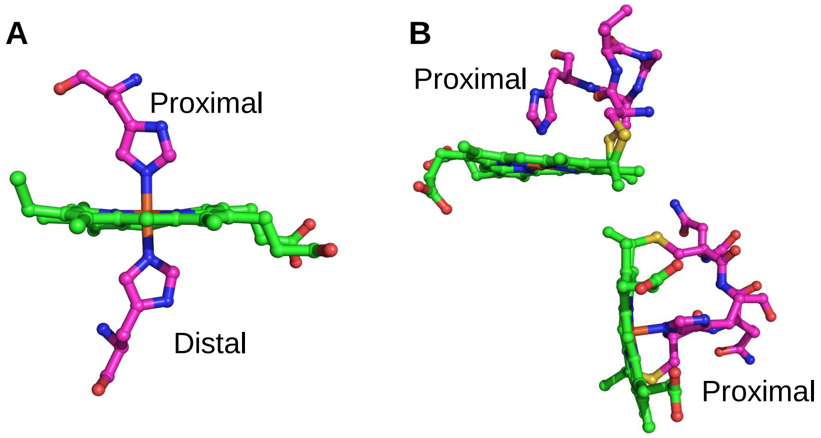


Figure 2.1 – Structures of **(A)** the iron ligating histidine residues and **(B)** the heme binding CXXCH motifs that the JAVA program searches for while collecting the PD and PP sequence data. The histidine ligands are labelled as proximal or distal.

To extract the PD sequences a JAVA program was written that identifies a set of user defined heme groups, then picks out these specific hemes from the original PDB files, identifies the iron ligating residues for each heme group by finding the closest two histidines to the heme iron and extracts the amino acid sequence between them. This program outputs; a file with sequences between the iron ligating histidines in FASTA format and a PDB file containing the coordinates of the aminoacids in this sequence.

The software written to extract the PP sequences works in much the same way. Also written in JAVA, this program; identifies the two hemes in the user defined list of heme pairs, finds them in their original PDB files, identifies the two cysteine residues that make up the CXXCH motifs binding each heme and extracts the amino acid sequence between them. As with the PD sequence program, files containing the amino acid sequence and coordinates are output. This process was performed on all pair, triplet, quartet and quintet motifs in the database.

2.2.4 Clustering of the extracted sequence data

To create subclusters of similar sequences within each pair, triplet and quartet heme clusters, three methods were employed. The first was a simple sequence length cut-off, grouping together sequences of similar lengths (typically 0-25, 26-50, 51-100, 100-150 and 150+ residues). The actual lengths were chosen based on the observed grouping of sequence lengths within each cluster. A second set of subclusters were created based on phylogenetic analysis, this was done by submitting all sequences in a given cluster to multiple sequence alignment via TCOFFEE **[Notredame *et al* 2000]** and interrogating the resulting phylogram. The third sub set of clusters were based on structural homology between the intervening polypeptide sequences. A JAVA program was written to perform a least squares superposition on the hemes from each cluster and calculate a C-α RMSD value for the intervening polypeptide, these RMSD values were output to a matrix that was clustered using the R package **[Haka *et al* 1669]** with the single-link hierarchal clustering method. The resulting dendrograms were interrogated to determine structural subclusters within each heme cluster. These analytical methods were performed on the PD sequences for each heme and PP sequences for each motif from each pair, triplet and quartet cluster. The quintet clusters were not analysed, as the majority only contained one sequence.

2.2.5 Generation of Hidden Markov Models

Hidden Markov Models (HMMs) were generated for each subcluster identified in section 2.2.4. To do this, multiple sequence alignments were generated for each subcluster using TCOFFEE **[Notredame *et al* 2000]** and output in MSF alignment format for HMM generation using the programs hmmbuild and hmmcalibrate from the HMMER package **[Eddy 1998]**. This resulted in sets of HMMs for each heme motif based on both sequence and structural homologies. These HMMs were used to search for the respective heme motif within the sequence of a protein of unknown structure.

2.2.6 Making predictive models using HMMs

Hmmsearch from the HMMER package **[Eddy 1998]** was used to search within the amino acid sequences of proteins of unknown structure with the HMMs generated for each heme motif. The region of the target sequence the HMM aligned with is referred to as a “hit”. Valid hits were identified as those that incorporated either a histidine residue at each end of the hit (with only one of the histidines as part of a CXXCH motif) in the case of the PD sequences or a CXXCH motif at each end of the hit in the case of the PP sequences. If two HMMs from different heme motifs produced significant hits in the same part of the sequence, that with the lowest E-value was taken as the correct result in that region of the sequence.

The first search pass was performed with the HMMs based on sequence data to predict the heme substructure, since these were found to have greater success in identifying the positions of heme motifs. A second pass was performed with the HMMs based on structural homologues and were used to indentify potential templates for the polypeptide structure between heme groups.

Once the make up of the heme substructure was predicted, a model was built using a JAVA program that had access to a representative structure for each motif and superimposes the first heme of the new motif onto the last heme of the previous as demonstrated in figure 3.2. This produces a PDB format file with a heme substructure for the protein of interest that is used as a “scaffold” for modelling the polypeptide structure.

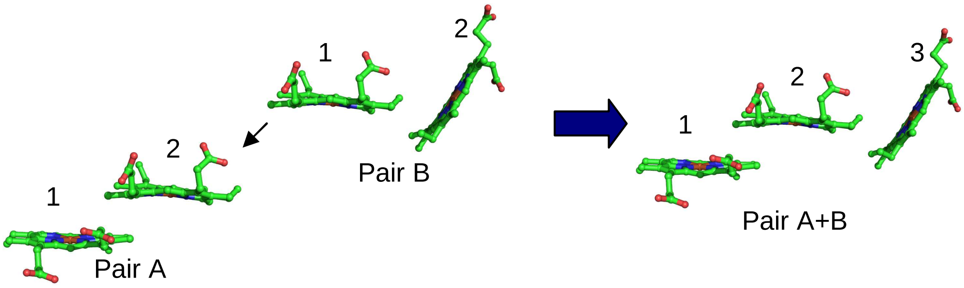


Figure 2.2 – A representation of how the JAVA program builds the heme substructure. Heme 1 of pair B is superimposed onto heme 2 of pair A, to create the triplet structure, the next pair will be superimposed onto heme 3 of this structure.

To add a polypeptide structure to the protein of interest, the hits from the structure-based HMMs are mapped onto the heme substructure. Where templates for polypeptide structure have been identified the hemes for each motif are superimposed to place the polypeptide template in the correct orientation and MODELLER **[Eswar *et al* 2006]** is used to predict the full 3D protein structure. Distance and angle restraints are placed on the heme ligating residues, as their positions have been found to be very highly conserved within hemoprotein structures, as will be shown in section 2.3.10.

The created models are then each given an overall score and E-value based on an average of the scores and E-values of the HMM hits that make up both the heme substructure and polypeptide template for the model, giving an indication of the quality of the prediction. For instances where there is no HMM coverage between heme motifs in the target sequence, from either the sequence based or structure based HMMs, penalty scores and E-values of -5 and 1 will be used.

2.2.7 Comparisons with existing protein structure prediction servers

In order to give a side by side comparison of the above methodologies with existing modelling techniques, models were also generated using the I-TASSER **[Zhang 2008]**, Phyre **[Kelley and Sternberg 2009]** and SwissModel **[Arnold *et al* 2006]** servers to ascertain the relative merits of each technique. These models were created by submitting the protein sequence of interest to each server and downloading the resulting PDB file.

2.2.8 Determination of the residues in van der Waals contact with each heme

To identify the residues in van der Waals contact with each heme a JAVA program was written to; read in a list of user defined heme groups, identify these hemes within their original PDB files and ascertain the residues within van der Waals contact. This was achieved by calculating the sum of the van der Waals radii plus an extra 50 % for each atom of each residue with each atom of each heme as a cut-off for the residues that will be identified as being in van der Waals contact with each heme, in short, if the radii overlap the residue and the heme are in van der Waals contact. The values for the van der Waals radii were obtained from the Cambridge Crystallographic Data Centre **[Allen 2002]** and the sum of van der Waals radii plus an extra 50 % value was used to compensate for inaccuracies within the crystal structures.

There were several outputs from this analysis, all taking the form of PDB files. For each run of the program, a PDB file was output containing the complete structures of all amino acids adjudged to be in contact with the heme, three more were also output where each residue was identified by a single atom as it's centroid and coloured by amino acid type, side chain charge (positive, negative or neutral) or side chain polarity (polar or non-polar).

2.3 Results

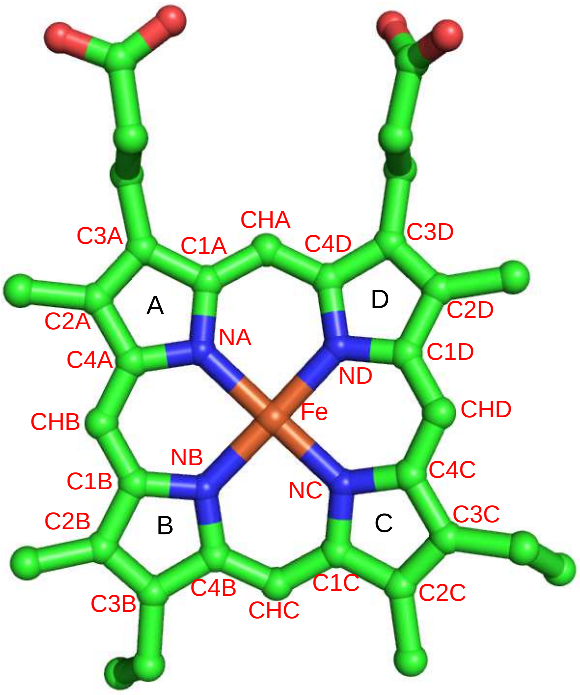


Figure 2.3 – The structure of a standard C-type heme, the four pyrrole rings A, B, C and D have been labelled.

Figure 2.3 shows the structure of a standard C-type heme with the four pyrrole rings labelled A, B, C and D, this figure will act as a reference for the following section when specific pyrrole rings are referred to.

2.3.1 Heme pair clusters in the 40 % *bis*-His ligated set

Table 2.2 – Breakdown of pair cluster sizes from the 40 % bis-His ligated set

Cluster Size	Number of clusters of this size	Number of heme pairs	Percentage of total heme pairs	Cumulative percentage of heme pairs
26+	1	28	22.4	22.4
20-25	1	20	16	38.4
10-19	2	20	16	54.4
5-9	4	34	27.2	81.6
1-4	19	23	18.4	100

In total, 27 pair clusters were identified, the two largest of these containing 28 and 20 heme pairs, corresponding to the di-heme elbow and parallel pair heme pair motifs respectively (Figure 2.4). As table 2.2 shows, these two clusters accounted for nearly 40 % of the total heme pairs. The nearest atoms between each heme in the pairs of the di-heme elbow motif cluster belonged to the pyrrole ring C, with an average heme-heme distance of 5.9 Å, this suggests electron transfer properties for this motif, since they are well within the minimum electron transfer distance of 14 Å **[Page *et al* 1999]** and it has been proposed that the exposed sulfur of the cysteine residue may act to facilitate electron transfer with the C ring of a heme **[Tollin *et al* 1986]**. The pyrrole A rings provide the nearest contacts between the hemes in the parallel pair cluster, with an average heme-heme distance of 4.1 Å, the alignment of the pair also brings the B rings together, with an average heme-heme distance of 5.1 Å, that are likely to facilitate electron transfer as they also contain bound sulfur from a cysteine residue.

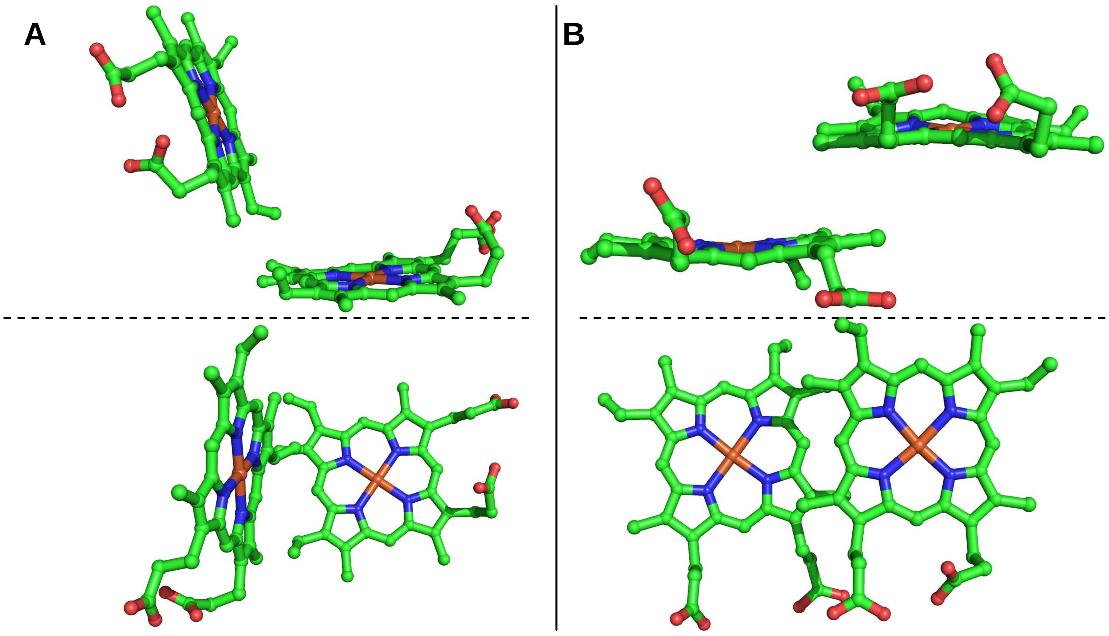
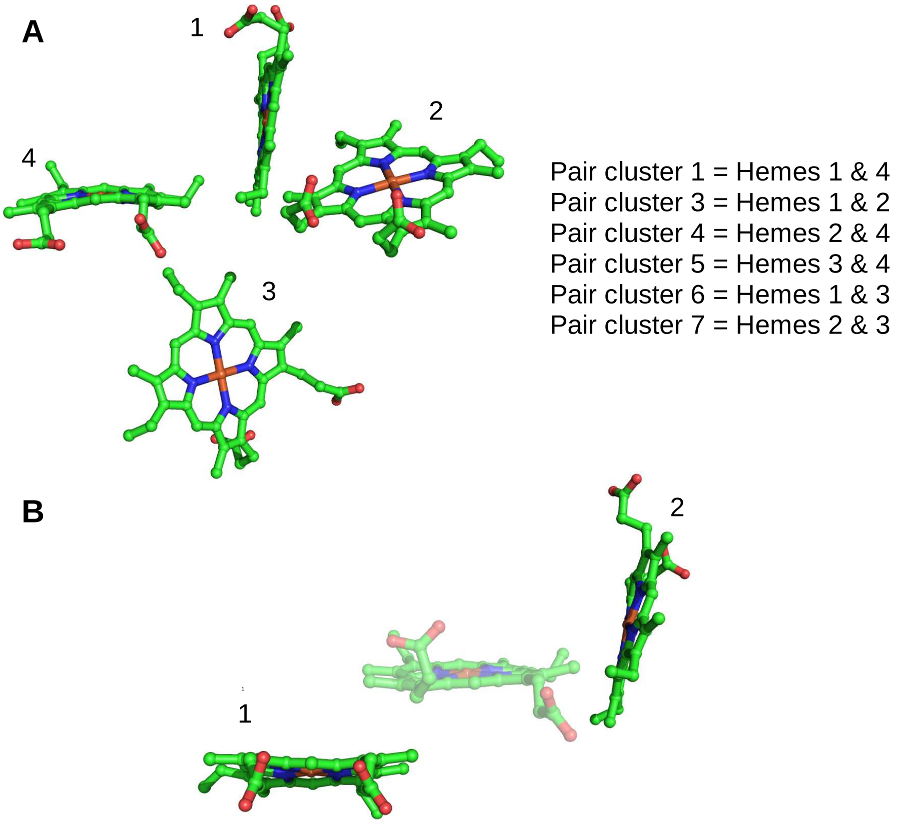


Figure 2.4 – Orthogonal views of the heme motifs for pair clusters 1 & 2, the two largest clusters of heme pairs. The di-heme elbow motif **(A)** contained 28 heme pairs and the parallel heme stacking motif **(B)** contained 20 pairs.

The next six largest clusters accounted for almost 40% of the total heme pairs, with two clusters containing 10 pairs, three 9 pairs and one 7 pairs. Clusters 3-7 form part of the cytochrome c_3-like motif and found in the tetraheme cytochrome c_3 domain (Figure 2.5A), the nearest-approach heme atoms in these clusters come from the D-D, B-B, B-D, B-B and D-B pyrrole rings, respectively. The minimum heme-heme distances are 9.4, 5.4, 6.5, 8.4 and 10.0 Å, respectively, potentially allowing electron transfer between the hemes in each of these pairs. Cluster 8 is part of the di-heme elbow motif SCOP family and looks similar to cluster 1 (the di-heme elbow motif), but differs from it due to the increased heme-heme distance between the two hemes: 5.9 Å for cluster 1 and 11.8 Å from cluster 8 (Figure 2.5B).



Pair cluster 1 = Hemes 1 & 4
Pair cluster 3 = Hemes 1 & 2
Pair cluster 4 = Hemes 2 & 4
Pair cluster 5 = Hemes 3 & 4
Pair cluster 6 = Hemes 1 & 3
Pair cluster 7 = Hemes 2 & 3

Figure 2.5 – A breakdown of the structures of clusters 3-8. **(A)** Clusters 3-7 (as well as cluster 1) can be found in the tetraheme cytochrome c3 domain. **(B)** Cluster 8, from the di-heme elbow like SCOP family is similar to cluster 1, as can be seen by the position of the transparent heme that indicates where the heme would lie relative to heme 2 in the cluster 1 pair motif.

The packing of subunit-spanning heme pairs (where each heme is coordinated by residues from a different polypeptide chain) are found in clusters 9, 10 and 13 (Figure 2.6). No intra-subunit heme pairs are found in these clusters suggesting that the subunit-spanning clusters represent unique heme pair packings. Cluster 9 contains two pairs from homodimeric cytochrome *c* nitrite reductases, cluster 10 contains two pairs from dimeric cytochrome c_3s and cluster 13 contains one pair from the hexameric cytochrome *c* nitrite reductase. The closest pyrrole rings between these pairs are A-A, B-B and C-D with minimum heme-heme distances of 5.2, 7.3 and 6.5 Å respectively. Representations of these pairs can be seen in figure 2.3. A complete breakdown of composition of the heme pair clusters can be seen in Appendix I.

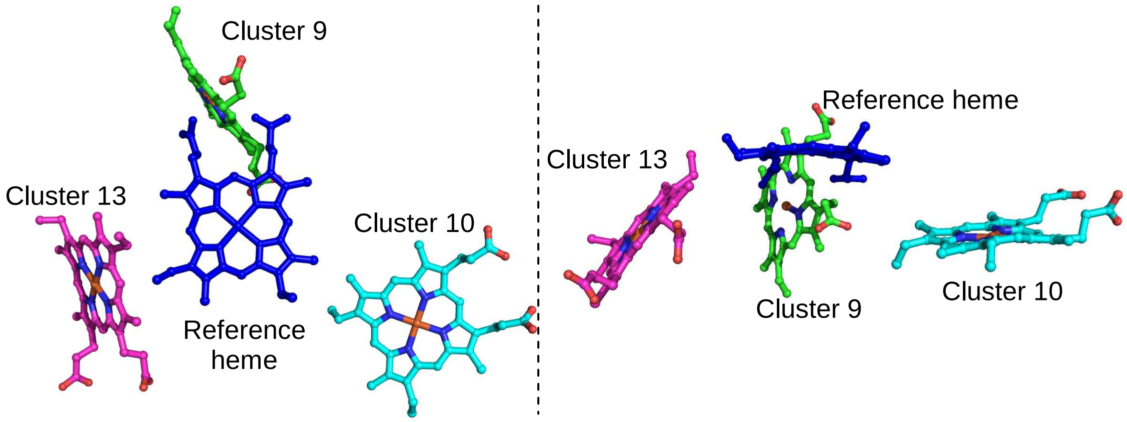


Figure 2.6 – Orthogonal views of the subunit-spanning pair motifs. One heme from each pair is aligned with a reference heme (Blue). The clusters shown are; cluster 9 (Green) from cytochrome *c* nitrite reductase, cluster 10 (Cyan) from dimeric cytochrome *c*$_3$ and cluster 13 (Magenta) from hexameric cytochrome *c* nitrite reductase.

2.3.2 Heme triplet clusters in the 40 % *bis*-His ligated set

Table 2.3 – Breakdown of triplet cluster sizes from the 40 % bis-His ligated set

Cluster Size	Number of clusters of this size	Number of heme triplets	Percentage of total heme triplets	Cumulative percentage of heme triplets
21+	1	23	20	20
11-20	0	0	0	20
5-10	5	42	36.5	56.5
1-4	40	50	43.5	100

Initial results identified 115 heme triplets, grouped into 46 clusters, the largest of which consisted of 23 heme triplets and corresponded to a triplet motif constructed from pair clusters 1 & 2 (the di-heme elbow and parallel pair motifs). However, the RMSD values for this cluster identified it as being made up of two separate clusters differing structurally according to the common heme shared by the di-heme elbow and parallel stacking pairs which combine to form the triplet. In other words, the clusters differ solely according to whether the first or second heme from the di-heme elbow pair constitutes the shared heme of the triplet. Examination of the amino acid sequence linking the hemes in each triplet has shown that they also differ according to the order in which the hemes from each pair motif are coordinated in the sequence, i.e. whether the motifs are ordered di-heme elbow → parallel pair or parallel pair → di-heme elbow in the sequence. Figure 2.7 shows the configuration of hemes in each of these clusters; 13 triplets fall into the parallel pair → di-heme elbow cluster, triplet cluster 1a (green), and 10 into the di-heme elbow → parallel pair cluster, triplet cluster 1b (cyan).

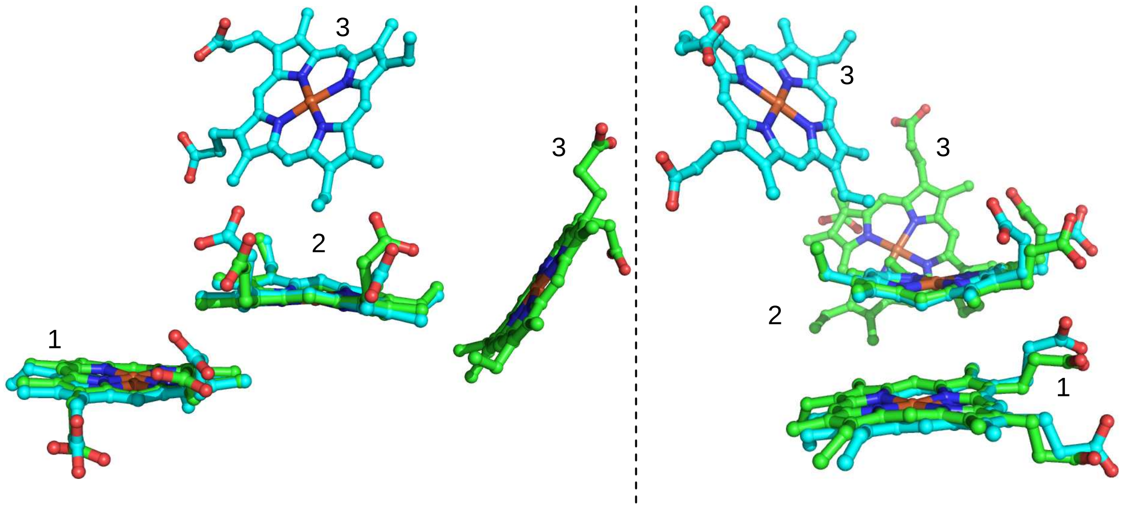
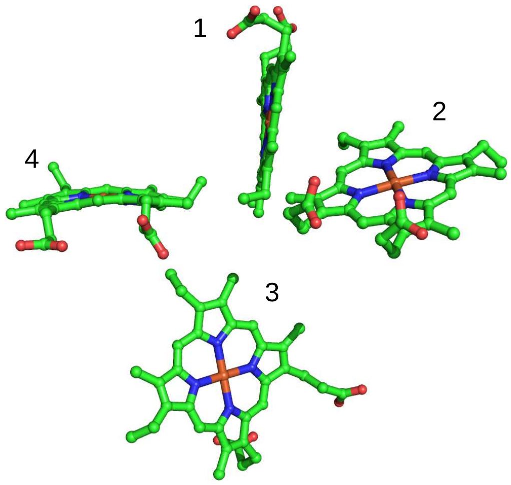


Figure 2.7 – Orthogonal views of the two subclusters found in triplet cluster 1 (heme numbers are arbitrary). These subclusters are defined by the position of heme 3 relative to heme 2 and the order the hemes from each pair are found in the sequence. In triplet cluster 1a (green) the order of the pairs is parallel pair → di-heme elbow and in triplet cluster 1b (cyan) the order of the pairs is di-heme elbow → parallel pair.

Of the next largest set of clusters (5-10 triplets in each), 4 of them contain cytochrome c_3-like proteins, more specifically they are made up of the different triplets found in the cytochrome c_3 tetraheme domain. Figure 2.8 shows this tetraheme cluster and the triplet clusters found within it.



Triplet cluster 2 = Hemes 1, 2 & 4
Triplet cluster 3 = Hemes 1, 3 & 4
Triplet cluster 4 = Hemes 1, 2 & 3
Triplet cluster 5 = Hemes 2, 3 & 4

Figure 2.8 – The cytochrome c_3 tetraheme domain, and the triplet clusters found within it (heme numbers are arbitrary).

The other cluster present in this set of clusters (those containing 5-10 triplets) is populated by proteins from the di-heme elbow SCOP family and refers to a triplet formed between the di-heme elbow motif and cluster 8 pair cluster motifs. A complete breakdown of the heme triplet clusters can be seen in Appendix II.

2.3.3 Heme quartet clusters in the 40 % bis-His ligated set

Table 2.4 – Breakdown of quartet cluster sizes form the bis-His ligated 40 % set

Cluster Size	Number of clusters of this size	Number of heme quartets	Percentage of total heme quartets	Cumulative percentage of heme quartets
5-10	3	23	23.2	23.2
2-5	6	13	13.1	36.3
1	63	63	63.7	100

In total 99 heme quartets were identified and grouped into 72 unique clusters. As with the triplet clusters, each quartet cluster contained proteins from only one SCOP family. The largest cytochrome c_3-like SCOP family containing cluster (cluster 2, containing 9 quartets), corresponded to the cytochrome c_3 tetraheme domain (Figure 2.5). The two other larger clusters (those containing more than 5 quartets) corresponded to conserved heme quartets in di-heme elbow motif family proteins (Figure 2.9). These quartet motifs are constructed from the sequential packing of di-heme elbow and parallel pair motifs, further emphasising the importance of these motifs in this family of proteins. The close proximity of pyrrole C & B rings along this quartet suggests these motifs would have electron transfer properties.

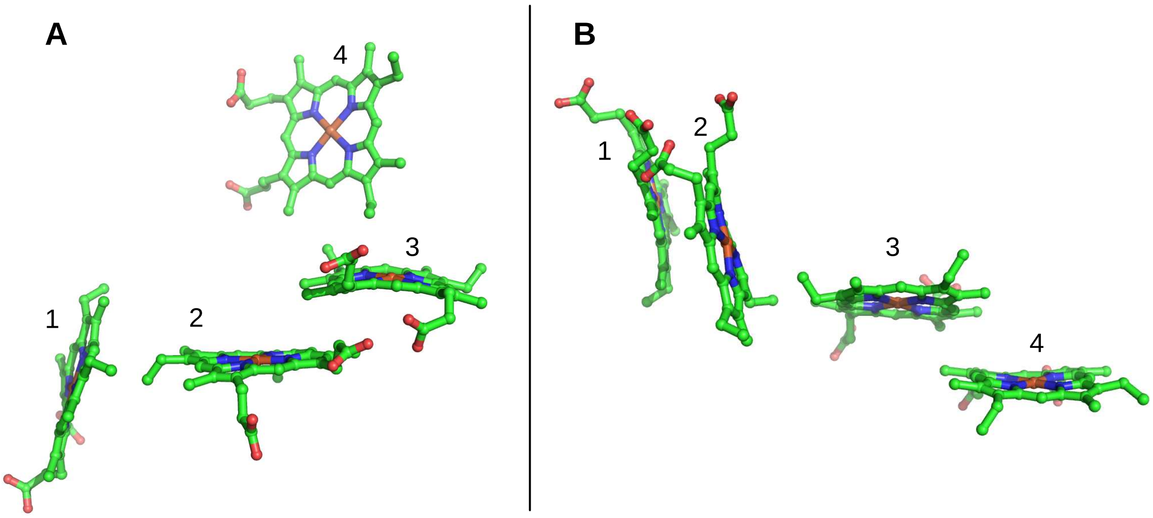


Figure 2.9 – The two largest SCOP di-heme elbow motif family quartet clusters (heme numbers are arbitrary). **(A)** Cluster 1 (8 quartets); constructed from one parallel pair (hemes 2-3) and two di-heme elbow (hemes 1-2 & 3-4) heme pair motifs. **(B)** Cluster 3 (6 quartets); constructed from one di-heme elbow (hemes 2-3) and two parallel pair (hemes 1-2 & 3-4) heme pair motifs.

The quartet clustering results emphasize the diversity adopted in the packing of hemes in multiheme proteins with larger heme substructures, since over 60 % of all heme quartets segregate into unique heme packing motif clusters. A complete breakdown of the heme quartet clusters can be seen in Appendix III.

2.3.4 Heme quintet clusters in the 40 % *bis*-His ligated set

Table 2.5 – Breakdown of quintet cluster sizes form the bis-His ligated 40 % set

Cluster Size	Number of clusters of this size	Number of heme quintets	Percentage of total heme quintets	Cumulative percentage of heme quintets
2	3	6	10.2	10.2
1	53	53	89.8	100

In total 59 heme quintets were found and separated into 56 unique clusters. As with the triplet and quartet clusters, each quintet cluster contained proteins from only one SCOP family. The number of quintets is significantly lower than the number of quartets as only 10 of the initial 28 proteins have more than 4 hemes. The main finding from the quintet analysis is the link between nrfB (PDB ID: 2OZY **[Clarke *et al* 2007]**) and the hexameric octaheme cytochrome c nitrite reductase (PDB ID: 2OT4 **[Polyakov *et al* 2009]**) that both contain instances of the largest quintet cluster (Figure 2.10).

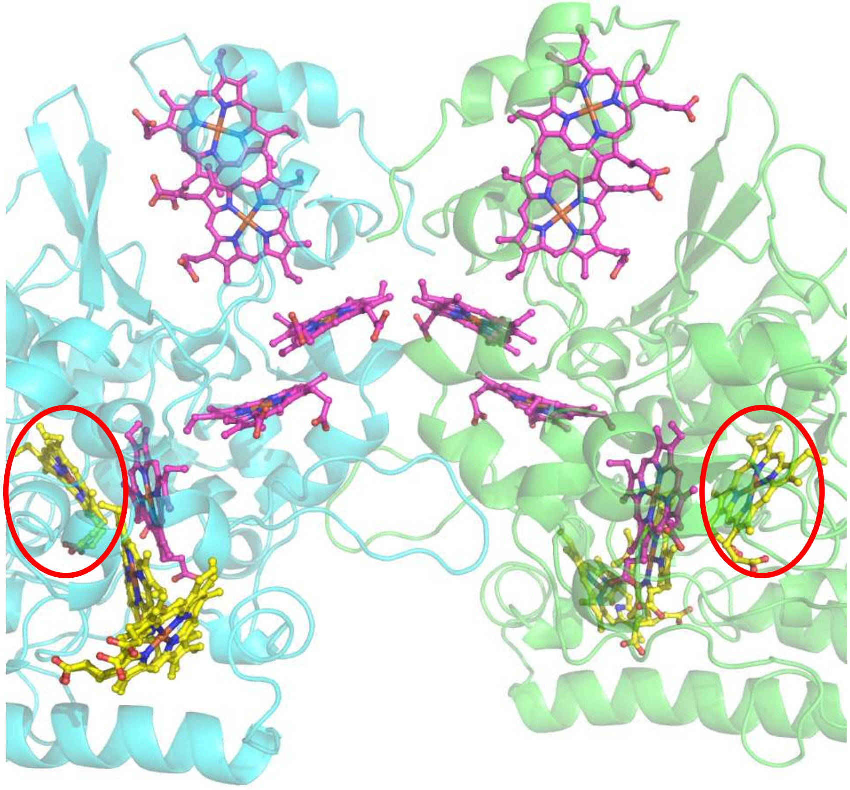


Figure 2.10 – The layout of hemes in the two chains pf the octaheme cytochrome *c* nitrite reductase (2OT4) **[Polyakov *et al* TBP]**. The magenta hemes refer to the quintet cluster of hemes also found in nrfB (2OZY) **[Clarke *et al* 2007]**, the yellow hemes refer to the other three hemes of the octaheme domain and the active site hemes of the nitrate reductase are circled red.

Approximately 90% of the quintets identified arise from clusters containing only one instance, with over half of these coming from the nine-heme cytochrome *c* (PDB ID:1OFW **[Bento *et al* 2003]**) due to the large number of unique quintets found in its densely packed heme substructure. A complete breakdown of the heme quintet clusters can be seen in Appendix IV.

No attempts were made to identify higher order clusters than quintets as there are insufficient multiheme cytochromes with six or more hemes. In addition, the majority of quintet clusters contain only one heme motif and more than half of these arise from the 9 heme cytochrome c. To a large extent, this makes the quintet clustering results redundant.

2.3.5 Distribution of heme motifs

Figure 2.11 shows a breakdown of heme motif clusters and the distribution of heme motifs within these clusters for the pairs, triplets, quartets and quintets. It shows the change in motif distribution from a smaller number of highly-populated clusters in the heme pair analysis to large numbers of lowly-populated clusters in the heme quintet analysis.

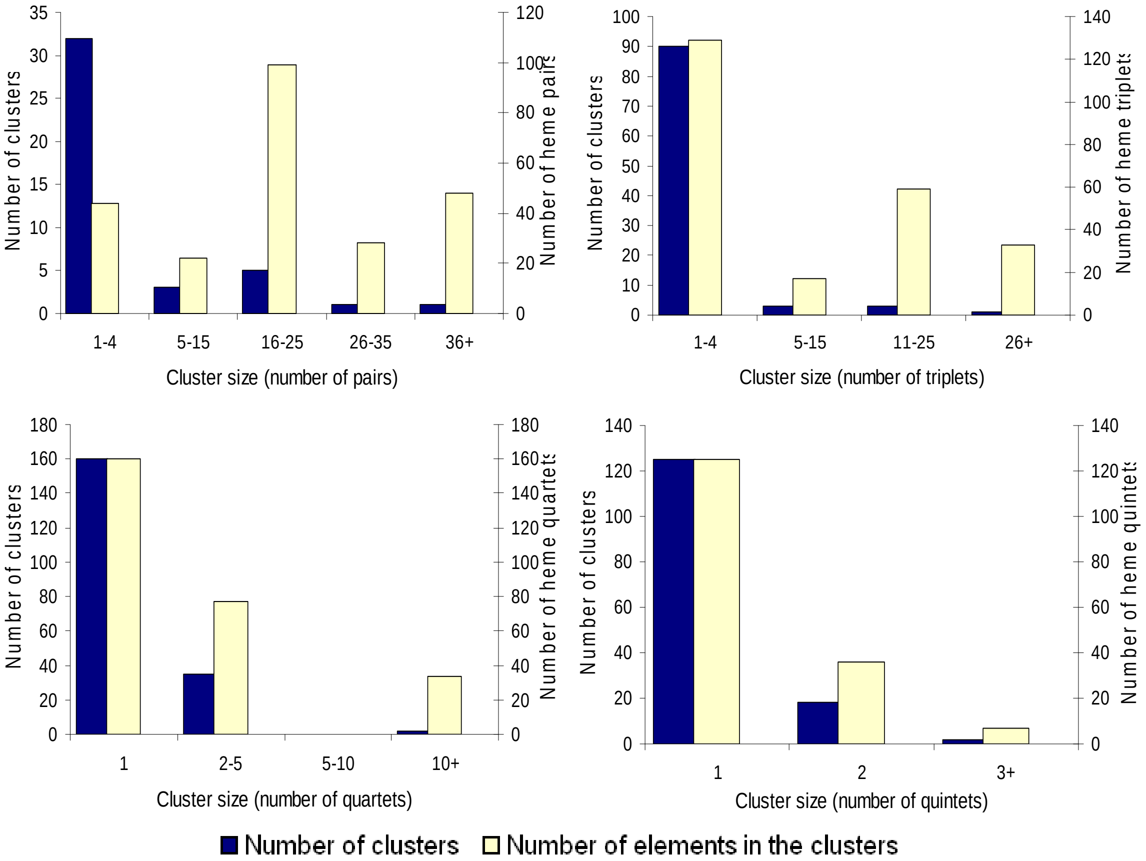


Figure 2.11 – Histograms showing an analysis of cluster size distribution of the heme pair, triplet, quartet and quintet motifs. These graphs demonstrate the shift from highly populated heme pair and triplet clusters, to predominantly single occupancy quintet clusters.

2.3.6 The effects of inclusion of non *bis*-His ligated heme pairs

The effects of inclusion of non *bis*-His ligated heme pairs on the preceding analysis can now be discussed. Of the 27 heme pairs added, only two were incorporated into existing pair clusters; the heme pair between hemes 90 and 92 of the c_7-type cytochrome from *Geobacter sulffereducins* (PDB ID: 1RWJ **[Pokkuluri *et al* 2004]**) was added to cluster 3 and the heme pair between hemes 91 and 92 from the same protein was added to cluster 4. These pairs from 1RWJ are structurally homologous to pairs from the dimeric cytochrome c_7 (PDB ID: 3BXU **[Morgado *et al* 2008]**), the 3BXU pairs were in the initial *bis*-His analysis while the 1RWJ pairs were not, due to heme 92 of 1RWJ having histidine - methionine ligation. This change in ligation has been proposed to give heme a higher reduction potential, with a midpoint reduction potential 50 mV higher than an equivalent domain where all three hemes are bis-His ligated **[Pokkuluri *et al* 2004]**.

All other non bis-His ligated heme pairs were grouped into unique clusters, with the two largest containing 6 and 7 pairs, respectively, both originating from proteins in the di-heme elbow motif SCOP family and both containing the active site heme for their respective protein structures.

The additional heme triplet, quartet and quintet motifs created by the incorporation of the non bis-His ligated heme pairs also fall into unique clusters, suggesting the motifs may be specific to the function of the hemes within their individual proteins and have no close evolutionarily links to other heme motifs.

2.3.7 Breakdown of PD sequence from heme pair subclusters

The extracted PD sequences for the most populous pair clusters were separated into subclusters. The sequences from the di-heme elbow (cluster 1) and parallel pair (cluster 2) motif clusters ranged from 12-195 and 13-215 residues in length and could be separated into 34 and 26 subclusters respectively. The sequences from clusters 3 to 7, that come exclusively from cytochrome c_3 like family proteins, have less diversity as can be seen by the lower number of subclusters. Table 2.6 shows a more detailed breakdown of these subclusters.

Table 2.6 – Breakdown of PD subclusters

Cluster ID	Number of sequence based subclusters		Number of structure based subclusters		Number of phylogram based subclusters		Total number of subclusters
Cluster 1	4	3	4	9	6	8	34
Cluster 2	4	5	4	3	5	5	26
Cluster 3	2	2	2	3	4	4	17
Cluster 4	2	2	4	3	4	4	19
Cluster 5	2	1	2	5	3	3	16
Cluster 6	1	2	5	5	3	3	19
Cluster 7	1	2	5	3	3	3	17

*NB the numbers in blue refer to the first heme in each pair motif and the red to the second

2.3.8 Breakdown of PP sequence from heme pair subclusters

The extracted PP sequences for the most populous pair clusters were separated into subclusters. The sequences from the di-heme elbow and parallel stacking pair motif clusters ranged from 27-73 and 24-145 residues in length respectively and could be separated into 20 and 15 subclusters respectively. The homology in the sequences and structures for clusters 3-7, that conform to the pairs found in the tetraheme c_3 domain, resulted in very few subclusters for these pair clusters. Table 2.7 shows a more detailed breakdown of these subclusters.

Table 2.7 – Breakdown of PP heme pair subclusters

Cluster ID	Number of sequence based subclusters	Number of structure based subclusters	Number of phylogram based subclusters	Total number of subclusters
Cluster 1	3	7	10	20
Cluster 2	4	3	8	15
Cluster 3	3	3	3	9
Cluster 4	3	3	3	9
Cluster 5	2	2	2	6
Cluster 6	2	2	2	6
Cluster 7	2	2	2	6

2.3.9 Breakdown of PP sequences from heme triplet subclusters

The extracted PP sequences for the most populous triplet clusters were separated into subclusters. The sequences from the most populous triplet cluster, based on packings of di-heme elbows and parallel pairs, ranged from 55-234 residues in length and could be separated into 17 subclusters. The first of which corresponds to the two heme substructures found in this cluster (Figure 2.7), with triplet cluster 1a referring to a parallel pair → di-heme elbow and triplet 1b referring to a di-heme elbow → parallel pair packing of pair motifs. The homology in the sequences and structures for clusters 2-5, that refer to the triplets found in the tetraheme cytochrome c_3 domain, resulted in very few subclusters for these triplet clusters. Table 2.8 shows a more detailed breakdown of these subclusters.

Table 2.8 – Breakdown of PP heme triplet subclusters

Cluster ID	Number of sequence based subclusters	Number of structure based subclusters	Number of phylogram based subclusters	Total number of subclusters
Cluster 1a	2	2	4	8
Cluster 1b	2	3	4	9
Cluster 2	3	3	3	9
Cluster 3	2	2	2	6
Cluster 4	2	2	2	6
Cluster 5	2	2	2	6

2.3.10 Distribution of residues in Van der Waals contact with the hemes in the most populous heme clusters

Analysis of the distributions of residues in van der Waal's contact with each heme in each pair cluster did not identify any specific residues, beyond the iron ligating histidines and heme binding cysteines, that are consistently found in a specific position around the heme, although it did identify very specific geometries for the heme binding histidines and cysteines (Figure 2.12).

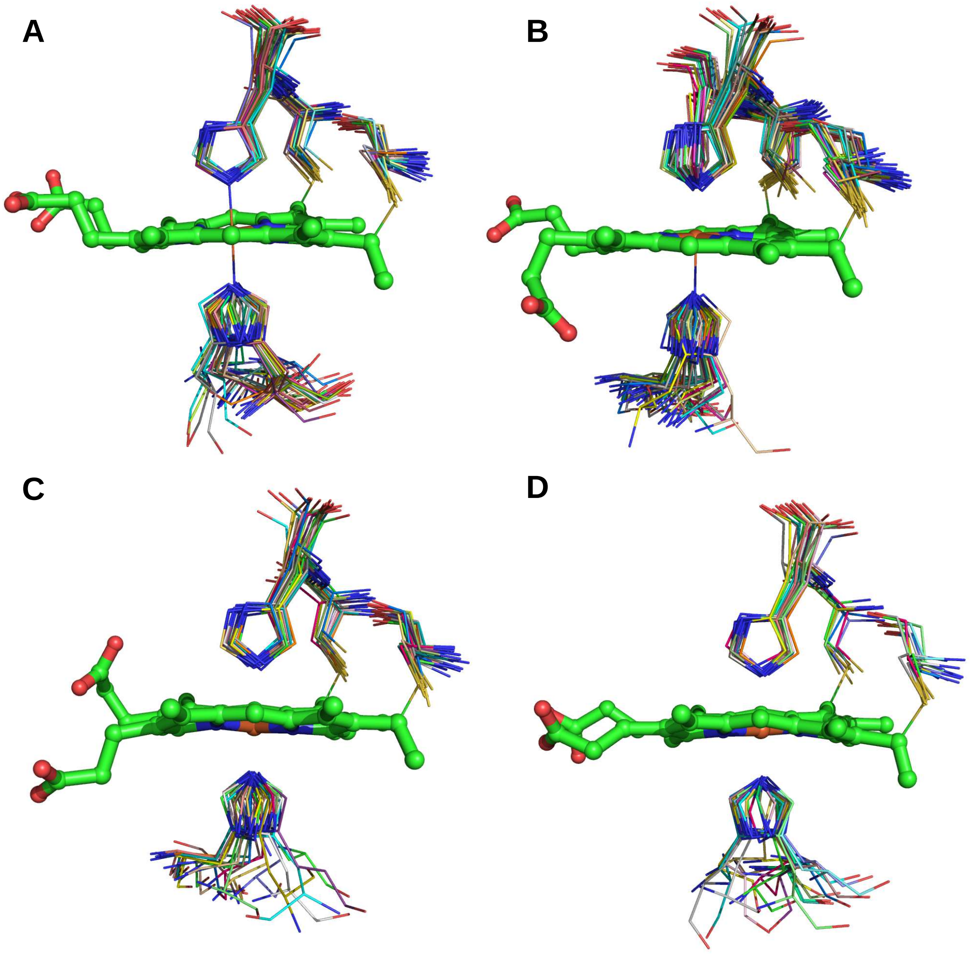


Figure 2.12 – Positions of the heme binding histidine and cysteine residues from the di-heme elbow motif first heme **(A)** and second heme **(B)**, and the parallel pair motif first heme **(C)** and second heme **(D)**

However, if the residues are grouped by their charge (i.e. whether they are positive, negative or neutral), it can be seen that the heme environment is predominantly neutral for both hemes of the parallel pair and di-heme elbow pair motifs (Figure 2.13).

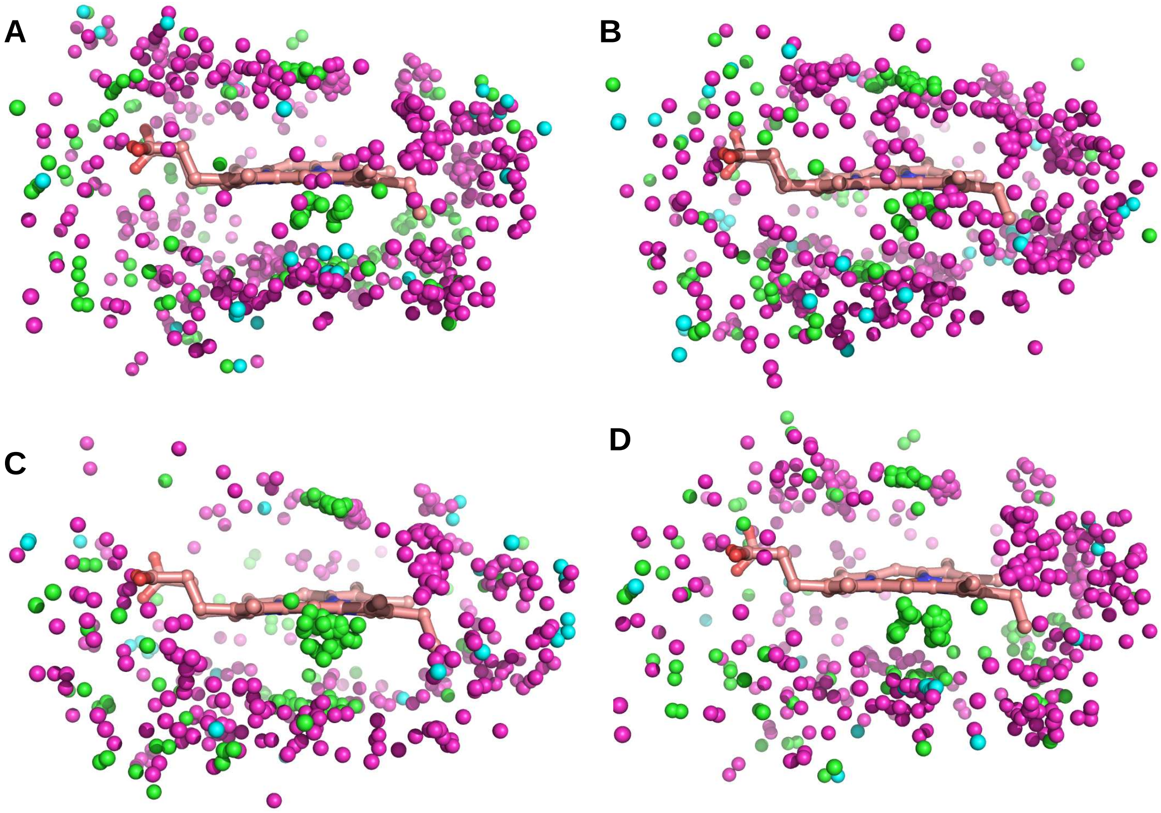


Figure 2.13 – Centroid positions of the residues in van der Waal's contact with the di-heme elbow motif first heme **(A)** and second heme **(B)**, and the parallel pair motif first heme **(C)** and second heme **(D)**. Residues with positive side chains are coloured green, negative side chain are cyan and neutral side chains magenta

The results of analysis of residues in van der Waal's contact, grouped by side chain polarity, do not show trends in the distribution of polar and non-polar residues around the heme groups (Figure 2.14).

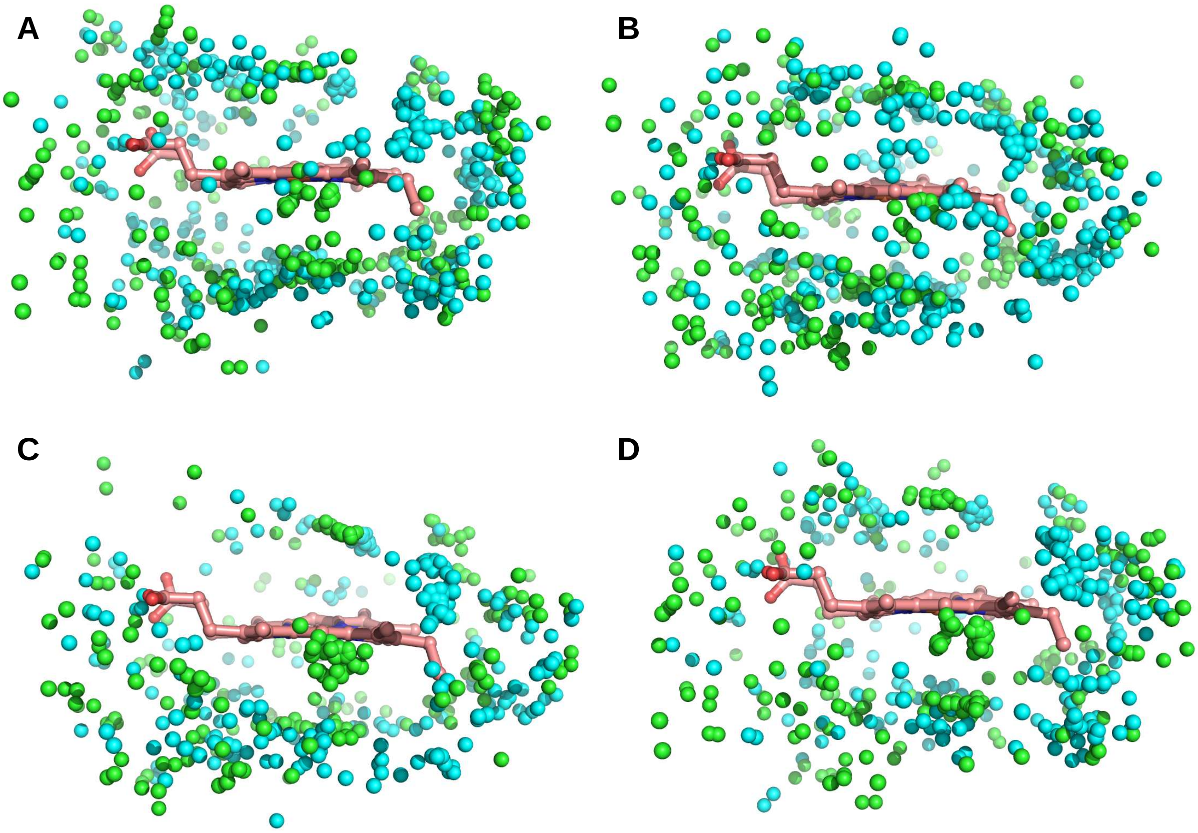


Figure 2.14 – Centroid positions of the residues in van der Waal's contact with the di-heme elbow motif first heme **(A)** and second heme **(B)**, and the parallel pair motif first heme **(C)** and second heme **(D)**. Residues with polar side chains are green and non-polar side chains are cyan.

2.3.11 Testing the HMM prediction methodology

To test the HMM based protein structure prediction methodology, two test cases were used, one from each of the major SCOP families. The small tetraheme cytochrome from *Shewanella frigidimarina* and the 12 heme cytochrome GSU_1996 from *Geobacter sulfurreducens*, both of which have unpublished structural data available from within the research group, in the case of the STC, and from an external collaboration with P.R.Pokuluri (Biosciences Division, Argonne National Laboratory), in the case of GSU_1996.

2.3.11.1 *Shewanella frigidimarina* small tetraheme cytochrome (STC)

The closest homologue to the STC from the non-redundant protein database was with 1M1Q **[Leys *et al* 2002]**, an STC from *Shewanella oneidensis* that had 69 % sequence identity, to ensure this highly homologous protein didn't bias the predictions, a new set of HMMs were generated with the 1M1Q sequence omitted from the multiple sequence alignment used to build the HMMs.

The hits found with the PD based HMMs did not produce particularly useful results, primarily due to many of the HMM hits not beginning and ending with a histidine residue, or if they did, either both or neither of the histidines occurred in a heme coordinating CXXCH motif, making the result meaningless in the context of iron ligating histidine prediction.

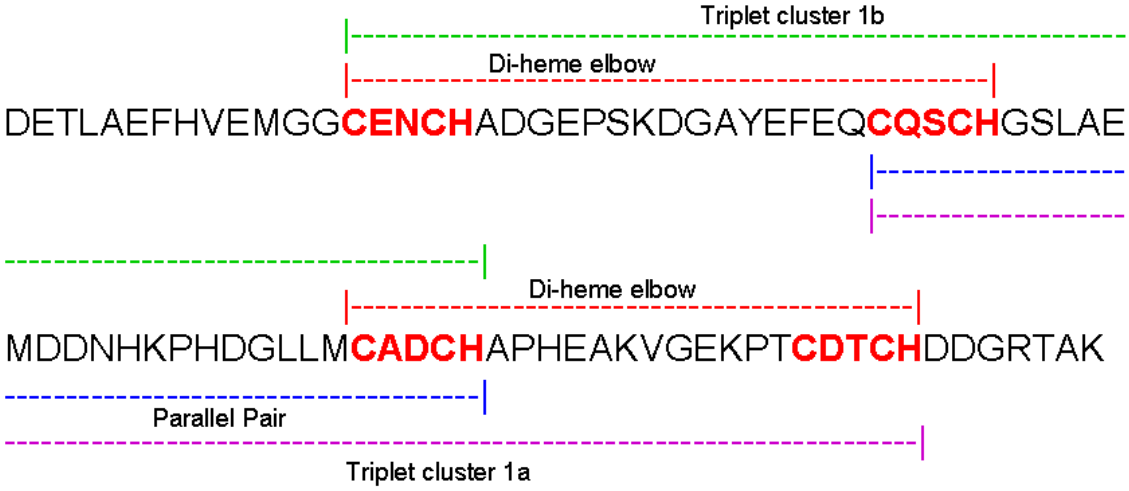

Triplet cluster 1b
Di-heme elbow
DETLAEFHVEMGGCENCHADGEPSKDGAYEFEQCQSCHGSLAE
Di-heme elbow
MDDNHKPHDGLLMCADCHAPHEAKVGEKPTCDTCHDDGRTAK
Parallel Pair
Triplet cluster 1a

Figure 2.15 – The distribution of the HMM hits on the STC sequence, the dotted lines refer to the regions covered by HMMs from di-heme elbows (red), parallel pairs (blue), triplet cluster 1a (purple) and triplet cluster 1b (green).

The hits found with the PP based HMMs gave a consensus heme substructure of two di-heme elbow motifs and a parallel pair motif in a sequential composition (Figure 2.15). These predictions were based on hits from pair, triplet and quartet sequence based HMMs. Searches carried out using the structure-based HMMs led to prediction of 3D structure for the polypeptide sequence between the first and last CXXCH motifs, that is, between residues 15 and 79 (comprising 75 % of the total sequence). This predication had an overall score of 34.18 and E-value of 2.0×10^{-5}.

The heme substructure and templates for the polypeptide structure were fed into MODELLER **[Eswar *et al* 2006]**, along with spatial restraints on the geometries of heme-coordinating residues i.e. the iron ligating histidine residues and cysteine residues of the CXXCH motif. The modelled structure was superimposed with the crystal structure (Figure 2.16) resulting in an RMSD of 1.7 Å. The close correspondence of theoretical model and crystal structure provides support for the validity of the method.

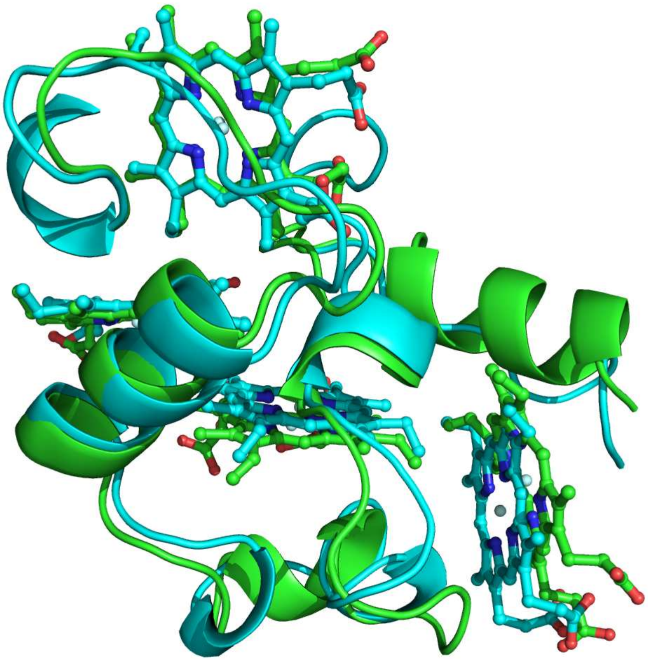


Figure 2.16 – A superposition of the modelled STC structure (Green) with the STC crystal structure (Cyan), showing the conservation in structure between the two models.

Both Phyre and Swissmodel predictions for the STC were built using the STC from *Shewanella oneidensis* (PDB ID: 1M1Q **[Leys *et al* 2002]**) as a template; as a result the structures of the predicted models are very close to the crystal structure, despite the lack of heme groups, with RMSDs of 0.46 Å and 0.45 Å, respectively. The I-TASSER prediction was built using the oxidised and reduced structures of the STC from *Shewanella oneidensis* (PDB IDs: 1M1Q and 1M1R respectively **[Leys *et al* 2002]**) as well as the NMR structure for the *Shewanella Frigidimarina* structure (PDB ID: 2K3V **[Paixao et al 2008]**), resulting in a model with an RMSD of 0.70 Å to the crystal structure.

2.3.11.2 *Geobacter sulfurreducens* GSU_1996

GSU_1996 is a multiheme cytochrome containing 12 heme groups. Searches performed using the HMMs derived from the cytochrome c_3-like SCOP family of clusters suggested the protein to be composed of a tandemly repeating array of four cytochrome c_7 domains (identified by means of hits with a subcluster of pair cluster 4) but was unable to predict structures for the heme pairs linking each domain. Searches with the HMMs derived from the di-heme elbow family of clusters suggested this inter-domain packing of hemes fell into the parallel pair heme motif substructure and was also able to provide a template for modelling of the structure of the connecting polypeptide chain (Figure 2.17). This predication had an overall score of 39.05 and E-value of 0.15.

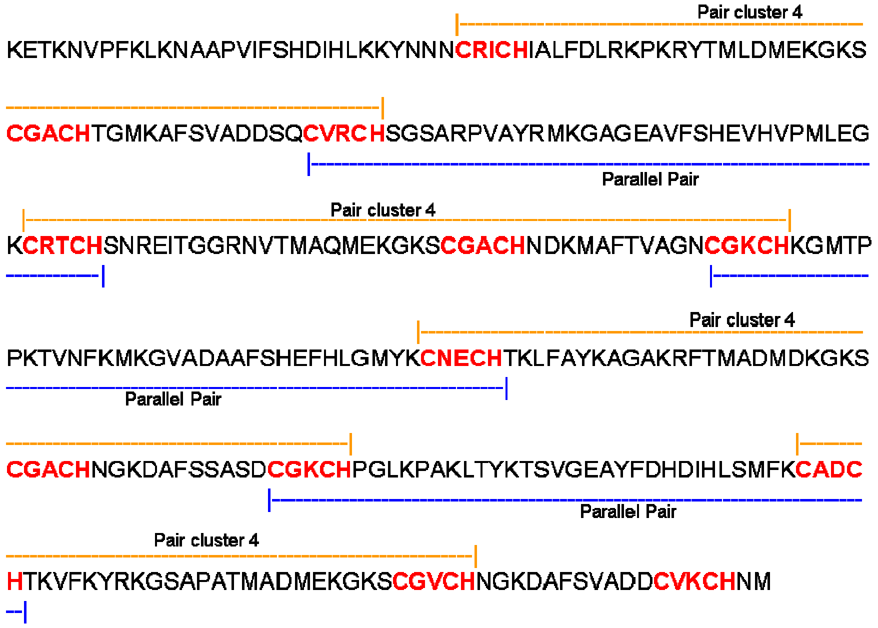

Pair cluster 4
KETKNVPFKLKNAAPVIFSHDIHLKKYNNNCRICHIALFDLRKPKRYTMLDMEKGKS
CGACHTGMKAFSVADDSQCVRCHSGSARPVAYRMKGAGEAVFSHEVHVPMLEG
Parallel Pair
Pair cluster 4
KCRTCHSNREITGGRNVTMAQMEKGKSCGACHNDKMAFTVAGNCGKCHKGMTP
Pair cluster 4
PKTVNFKMKGVADAAFSHEFHLGMYKCNECHTKLFAYKAGAKRFTMADMDKGKS
Parallel Pair
CGACHNGKDAFSSASDCGKCHPGLKPAKLTYKTSVGEAYFDHDIHLSMFKCADC
Parallel Pair
Pair cluster 4
HTKVFKYRKGSAPATMADMEKGKSCGVCHNGKDAFSVADDCVKCHNM

Figure 2.17 – The distribution of the HMM hits on the GSU_1996 sequence, the dotted lines refer to the regions covered by HMMs from parallel pairs (blue) and pair cluster 4 (orange).

This information was fed to MODELLER and a first generation model created. From this model, the identity of distal heme ligating residues could be ascertained by using a program written in JAVA to find the closest available ligand to the heme iron. It was found that 8 of the 12 hemes were *bis*-His ligated, while every 3rd heme was His-Met ligated. Spatial restraints were placed on the heme coordinating residues and a second generation model produced for the protein (Figure 2.18).

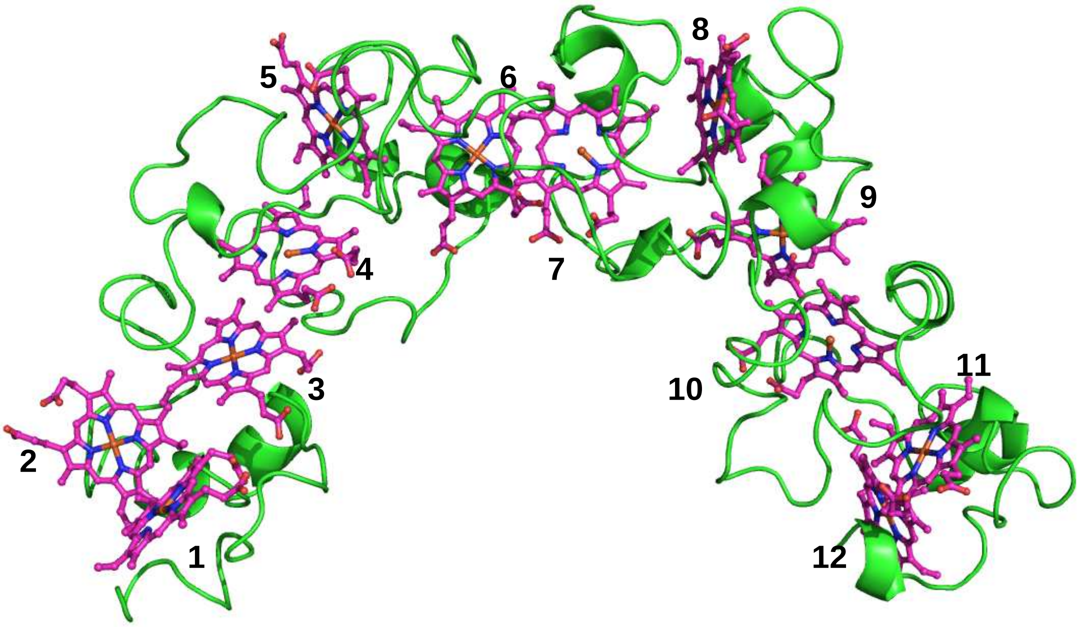


Figure 2.18 – A predicted model for the protein GSU_1996 as created using the prediction methodology described in this chapter. The hemes are displayed as magenta sticks and the amino acid backbone as a green cartoon.

Comparisons between the modelled structure and the available experimental data provided subsequently by P.R.Pokuluri (Biosciences Division, Argonne National

Laboratory) has shown that the predicted structure has a good fit to the crystal structure if split up into two regions. The first region, covering hemes 1-6, has an RMSD of 2.7 Å, while the second region, covering hemes 7-12, has an RMSD of 5.0 Å. However, the predicted and crystal structures diverge at the point linking hemes 6 and 7 (Figure 2.19), with an RMSD of 15.4 Å for the complete structure.

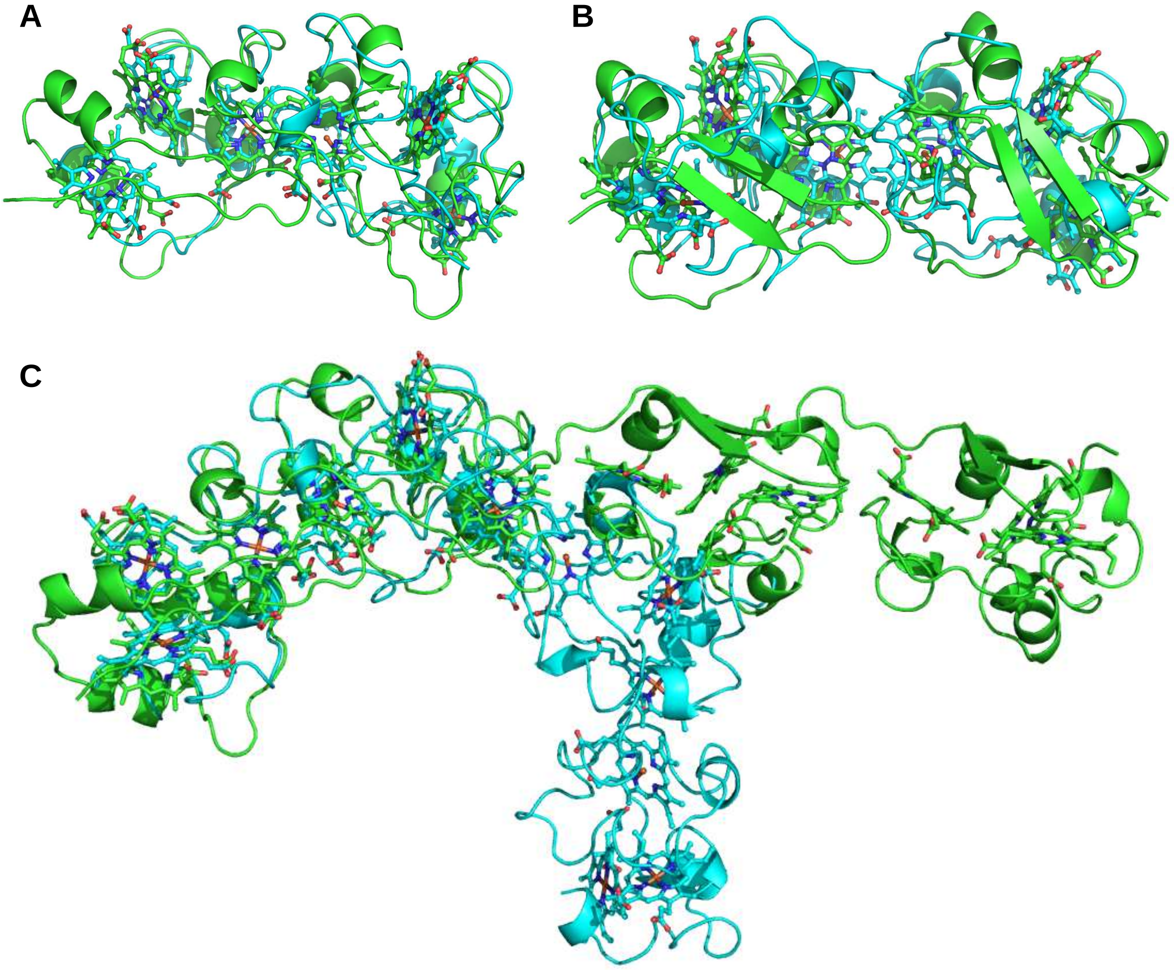


Figure 2.19 – Alignments of the GSU_1996 predicted structure (Cyan) with the crystal structure (Green). **(A)** The alignment between the region of the protein covering hemes 1-6 (RMSD 2.7 Å). **(B)** The alignment between the region of the protein covering hemes 7-12 (RMSD 5.0 Å). **(C)** The alignment between the region of the protein covering hemes 1-6 with the rest of the predicted model (RMSD 15.4 Å) shown to demonstrate the diversion in the two structures.

The heme pair linking hemes 6 and 7 from the GSU_1996 crystal structure was compared with the existing heme pairs in the database by adding it to the complete list of heme pairs and recalculating the pair clusters to observe which, if any, cluster this pair was located in. It was found that this domain linking pair motif had a novel heme packing geometry that was unlike any previously observed structures in the heme pair database. This demonstrates a limitation in the methodology as it is unable to handle structures with hitherto unseen heme packings.

The Phyre and Swissmodel predictions for GSU_1996 were both built using hexadecaheme cytochromes as templates, with the Phyre template coming from *Desulfovibrio vulgaris Hildenborough* (PDB ID: 1GWS **[Czjzek *et al* 2002]**) and the Swissmodel template coming from *Desulfovibrio gigas* (PDB ID: 1Z1N **[Santos-Silva *et al* 2007]**), these structure had a 14 % and 17 % sequence identity to GSU_1996 respectively. The I-TASSER prediction was built using a nonaheme cytochrome c structure (PDB IDs: 1DUW **[Umhau *et al* 2001]**) two different hexadecaheme cytochrome structures (PDB IDs: 1Z1N and 2E84 **[Shibata *et al* 2004]**) and the structure of a zinc finger DNA binding protein (PDB ID: 2I13 **[Segal *et al* 2006]**)

The predicted GSU_1995 structures contained no heme groups, which make up a large proportion of the final structure, and as a result the predicted structures were a very poor match to the crystal structure, with RMSDs of 25.2 Å for the Phyre prediction, 21.3 Å for the Swissmodel prediction and 22.07 Å for the I-TASSER prediction (Figure 2.20).

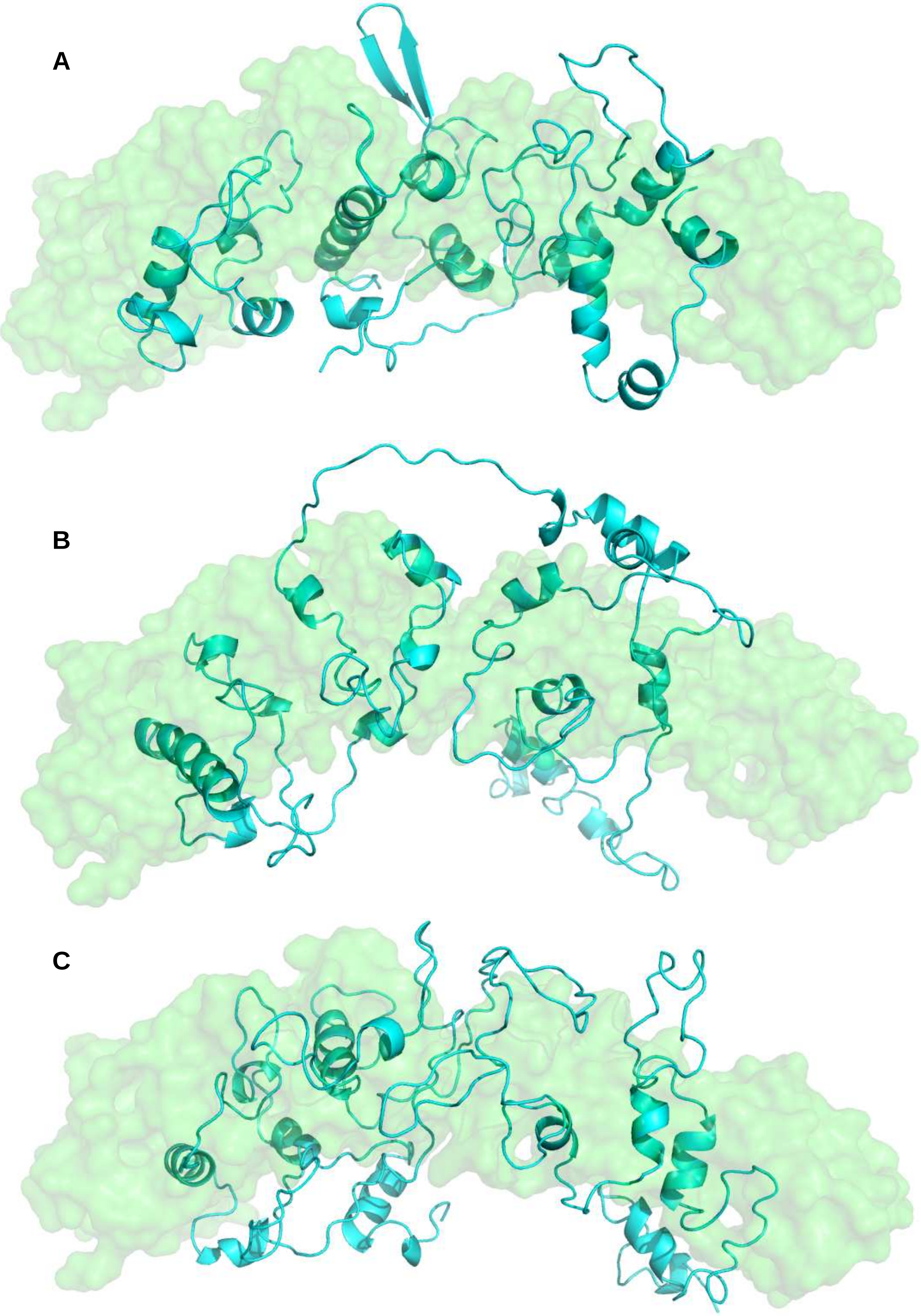


Figure 2.20 – Superposition of the GSU_1996 structure as predicted by **(A)** Phyre, **(B)** Swissmodel and **(C)** I-TASSER (all cyan), with the crystal structure (Green), due to the major differences between the structures, for clarity, only the surface of the crystal structure is shown.

2.4 Discussion

2.4.1 The distribution of heme motifs in multiheme proteins

The results of clustering the heme pair motifs used in this analysis showed that the majority of pair clusters contain proteins from only one SCOP family **[Murzin *et al* 1995]**. The exception to this was the most populous cluster (corresponding to the di-heme elbow motif), which contained a mixture of heme pairs from the cytochrome c_3-like (11 pairs) and di-heme elbow motif (19 pairs) SCOP family proteins, identifying an evolutionary link between the two families. This also highlights an issue with the naming of the families in the SCOP database, as it seems counterintuitive that the di-heme elbow motif family would not contain all the hemes that contain the di-heme elbow heme pair motif. The addition of non *bis*-His ligated pairs had little effect on the composition of the clusters with 25 of the 27 extra heme pairs falling into unique clusters, suggesting these hemes have evolved unique packing arrangements to facilitate their functions as active site hemes.

The clusters identified for the larger heme motifs (triplets, quartets and quintets) all contained proteins from only one SCOP family. This suggests that the packing of the hemes is specific to the function of the protein since the proteins in the cytochrome c_3-like family function in electron transport and the proteins of the di-heme elbow motif family are predominantly enzymes.

The clustering results have also shown that, for the dataset used in this analysis, all members of the di-heme elbow SCOP family contain both di-heme elbow pair and interacting parallel pair motifs, suggesting a functional relationship between the two. In fact, almost all of the heme substructures observed in SCOP di-heme elbow family proteins can be built using sequential packing of these two motifs, with the addition of an active site motif if required to add enzymatic functionality.

A recent investigation into the distribution of multiheme C-type cytochromes in prokaryotic organisms **[Sharma *et al* 2010]** that clustered a range of multiheme cytochrome sequences, identified several conserved multiheme cytochrome structures for six of the fifteen most populous clusters identified from their analysis. These corresponded to; the Nitrite reductase NrfA, NrfB, hydroxylamine oxidase and tetrathionate reductase, that all contain sequential packings of di-heme elbows, parallel pairs and an active site pair (where required), NapB, that contains a single parallel pair and the di-heme cytochrome c that contains a single unique heme pair found in one structure in a single cluster in this chapter. Their conclusions, with regards to the structural properties of multiheme cytochromes, were that the lack of structural templates for may of the clusters of multiheme cytochromes identified in their analysis would limit the structural characterisation of these multiheme cytochromes due to the inability to predict key factors such as the position and orientation of heme cofactors and their ligands **[Sharma *et al* 2010]**. Work in this chapter has shown that it is possible to break these larger structures down into their component parts and use them to construct models based on the assembly of these component parts. This method not only aids in the prediction of the polypeptide structure, but it crucially also predicts the position and orientation of the heme cofactors that are integral to the function of all multiheme cytochromes.

2.4.2 Sequence-derived subclusters found in heme motif clusters

The results of the subclustering analyses performed on the most populous heme clusters created subclusters based on both sequence and structural information for heme coordinating polypeptide. They identified a much greater homology between the sequences and structures of the cytochrome c_3-like SCOP family of proteins, with the majority of clusters based on hemes from this family having very few polypeptide derived subclusters. In contrast, the clusters based on di-heme elbow SCOP family proteins demonstrated much more variation in the linking polypeptide sequences, with many more polypeptide derived subclusters (See Tables 2.6-2.8). This pattern fits with the observation of the more diverse functions of proteins from the di-heme elbow SCOP family.

When the HMMs built from these subclusters were used to search sequences of unknown structure, it was found that the HMMs built using the heme coordinating PP sequences identified more valid hits than those built using the iron ligating PD sequences. This was due to the PD-based HMMs often giving a hit where neither or both of the histidine residues at each end of the hit were part of a CXXCH motif, these hits cannot be correct as each *bis*-His ligated iron must be ligated by one histidine from a CXXCH motif and one from another histidine found in any other region of the sequence that is not part of a CXXCH motif. The reason for this problem is likely to be that a single histidine residue at each end of the HMM is not a sufficient enough sequence motif to insure the HMM hits in a valid section of the search sequence. However, it appears that the CXXCH motif is sufficient enough to get consistent hits in desirable regions, as there were very few instances during this research of PP based HMM hits where the hit to the target sequence did not begin and end with a CXXCH motif.

2.4.3 Assessment of the MHC structure prediction method

The comparisons of the predicted STC and GSU_1996 models with the crystal structures for each protein have shown that the HMM based methodology developed during this work has the ability to predict at least partial protein structures with an acceptable degree of accuracy. The predicted model for STC was close to that of the crystal structure over 75 % of the polypeptide chain (that between the first and last CXXCH motifs), with a final RMSD of 1.7 Å for the alignment of the two structures.

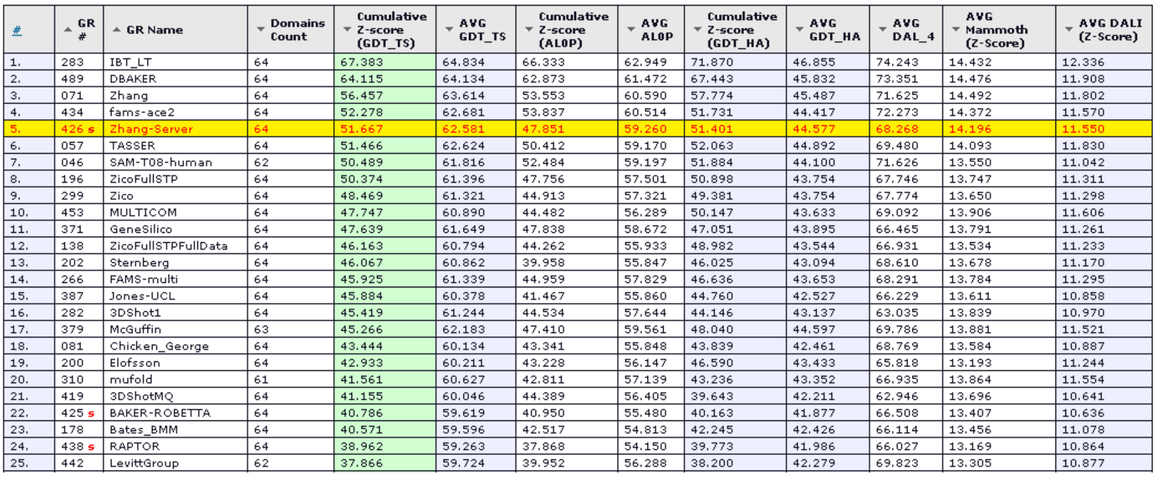
If separated into two regions the predicted model for GSU_1996 was close to the crystal structure over the first region covering hemes 1-6, as an RMSD of 2.7 Å would suggest, and the second region covering hemes 7-12 was also quite close to the crystal structure, with an RMSD of 5.0 Å. However, taking the protein as a whole it was identified that the linking heme pair between these two regions of the protein in the crystal structure was very different to that in the predicted structure, with an RMSD of 15.4 Å for the whole structure. An analysis of this linking heme pair has shown that it is unlike any previously observed during this work. This demonstrates a limitation of the HMM based prediction methodology, as it is unable to factor in unique heme packing into its predictions.

The overall scores calculated for these models (34.18 for the STC and 39.05 for GSU_1996) suggest that they are both accurate predictions, with the GSU_1996 prediction seemingly being more accurate due to its higher score. However, the overall E-values for each model (2.0×10^{-5} for the STC and 0.15 for GSU_1996) point towards the STC being the more accurately predicted structure, a finding which is confirmed by the RMSDs calculated from the alignments between the predictions and the crystal structures. This shows how both parameters are needed to make a judgement on the accuracy of any predictions and also provides some example overall scores and E-values for predictions of proven accuracy.

2.4.4 HMM methodology vs existing prediction servers

The comparisons of the models predicted using the HMM based methodologies developed in this chapter and existing techniques available at the I-TASSER, Phyre and Swissmodel servers, identified specific advantages to the HMM based method. The I-TASSER, Phyre and Swissmodel servers may have made a more accurate prediction of the STC structure, as proved by RMSD's of 0.701, 0.459 and 0.452 Å respectively, but the poorer performance of the HMM based method, which had an RMSD of 1.7 Å for its superposition to the crystal structure, is likely to be due to the removal of the close homologue that was in fact used as the template for the two structure prediction servers. However, these models were still missing the heme groups that the HMM based methods are able to generate, a feature that is lacking in all online prediction servers. It should also be noted that the inclusion of the close homologue to the HMM prediction methodology gave a final model with an RMSD fit of 0.389 Å to the crystal structure, a more accurate result than those generated by the online servers.

The inability of I-TASSER, Phyre and Swissmodel to incorporate heme into their predictions becomes even more of an issue during the prediction of GSU_1996. The absence of any close homologue covering the whole of the protein caused severe problems for the Phyre and Swissmodel servers, with their final models having RMSDs of 22.07, 25.2 and 21.3 Å when compared to the crystal structure. The RMSD of 15.4 Å for the HMM based model may not seem significantly better, but the HMM model was shown to be a close match to the crystal structure over the two separate halves of the protein, and was only let down in its ability to predict the novel heme packing between hemes 6 and 7. In contrast, the models generated by I-TASSER, Phyre and Swissmodel had no such regions of close homology, this is likely to be due to their lack of heme incorporation, since in heme rich proteins such as GSU_1996, the heme substructure is likely to be the primary driving force in the overall protein structure.



#	GR #	GR Name	Domains Count	Cumulative Z-score (GDT_TS)	AVG GDT_TS	Cumulative Z-score (AL0P)	AVG AL0P	Cumulative Z-score (GDT_HA)	AVG GDT_HA	AVG DAL_4	AVG Mammoth (Z-Score)	AVG DALI (Z-Score)
1.	283	IBT_LT	64	67.383	64.834	66.333	62.949	71.870	46.855	74.243	14.432	12.336
2.	489	DBAKER	64	64.115	64.134	62.873	61.472	67.443	45.832	73.351	14.476	11.908
3.	071	Zhang	64	56.457	63.614	53.553	60.590	57.774	45.487	71.625	14.492	11.802
4.	434	fams-ace2	64	52.278	62.681	53.837	60.514	51.731	44.417	72.273	14.372	11.570
5.	426 s	Zhang-Server	64	51.667	62.581	47.851	59.260	51.401	44.577	68.268	14.196	11.550
6.	057	TASSER	64	51.466	62.624	50.412	59.170	52.063	44.892	69.480	14.093	11.830
7.	046	SAM-T08-human	62	50.489	61.816	52.484	59.197	51.884	44.100	71.626	13.550	11.042
8.	196	ZicoFullSTP	64	50.374	61.396	47.756	57.501	50.898	43.754	67.746	13.747	11.311
9.	299	Zico	64	48.469	61.321	44.913	57.321	49.381	43.754	67.774	13.650	11.298
10.	453	MULTICOM	64	47.747	60.890	44.482	56.289	50.147	43.633	69.092	13.906	11.606
11.	371	GeneSilico	64	47.639	61.649	47.838	58.672	47.051	43.895	66.465	13.791	11.261
12.	138	ZicoFullSTPFullData	64	46.163	60.794	44.262	55.933	48.982	43.544	66.931	13.534	11.233
13.	202	Sternberg	64	46.067	60.862	39.958	55.847	46.025	43.094	68.610	13.678	11.170
14.	266	FAMS-multi	64	45.925	61.339	44.959	57.829	46.636	43.653	68.291	13.784	11.295
15.	387	Jones-UCL	64	45.884	60.378	41.467	55.860	44.760	42.527	66.229	13.611	10.858
16.	282	3DShot1	64	45.419	61.244	44.534	57.644	44.146	43.137	63.035	13.839	10.970
17.	379	McGuffin	63	45.266	62.183	47.410	59.561	48.040	44.597	69.786	13.881	11.521
18.	081	Chicken_George	64	43.444	60.134	43.341	55.848	43.839	42.461	68.769	13.584	10.887
19.	200	Elofsson	64	42.933	60.211	43.228	56.147	46.590	43.433	65.818	13.193	11.244
20.	310	mufold	61	41.561	60.627	42.811	57.139	43.236	43.352	66.935	13.864	11.554
21.	419	3DShotMQ	64	41.155	60.046	44.389	56.405	39.643	42.211	62.946	13.696	10.641
22.	425 s	BAKER-ROBETTA	64	40.786	59.619	40.950	55.480	40.163	41.877	66.508	13.407	10.636
23.	178	Bates_BMM	64	40.571	59.596	42.517	54.813	42.245	42.426	66.114	13.456	11.078
24.	438 s	RAPTOR	64	38.962	59.263	37.868	54.150	39.773	41.986	66.027	13.169	10.864
25.	442	LevittGroup	62	37.866	59.724	39.952	56.288	38.200	42.279	69.823	13.305	10.877

Figure 2.21 – An excerpt from the Group performance table from the CASP website (http://predictioncenter.org/casp8/groups_analysis.cgi), with the position of the Zhang-Server (I-TASSER) highlighted in yellow. GR numbers with an **S** next to them refer to groups taking part in the server-CASP experiment, where the results are produced by automated servers.

At the most recent CASP (Critical Assessment of Techniques for Protein Structure Prediction) event (CASP 8) **[Moult *et al* 2009]** the I-TASSER server was determined to be the most accurate of the structural prediction servers (Figure2.21). The methods described in this chapter appear to provide a novel and useful first solution to the problem of predicting the three dimensional structures of multiheme cytochromes – a problem beyond the limitations of current structure prediction servers.

Chapter 3 - A study of multiheme cytochromes from the cytochrome rich bacterial species *Shewanella oneidensis* and *Geobacter sulfurreducens*

3.1 Introduction

The *Geobacter* and *Shewanella* species of proteobacteria are of interest to the scientific community due to their novel electron transfer capabilities, their ability to generate electricity from waste organic matter and their role in bioremediation of contaminated environments **[Giometti 2006]**.

Geobacter sulfurreducens is a species of the *Geobacteraceae* family; they are comma shaped, gram negative, anaerobic bacteria that have been found as a predominant microbial component of diverse subsurface environments, including aquatic sediments, pristine deep water aquifers and petroleum-contaminated shallow aquifers **[Coates *et al* 1996]**. The complete genome sequence of *G.sulfurreducens* is 3.8 mega-bases in size and encodes a predicted 3466 proteins.

Shewanella oneidensis is a species of the *Shewanellaceae* family; they are gram negative bacteria that have the ability to grow both aerobically and anaerobically. First isolated in Lake Oneida, *S.oneidensis* is found predominantly in aquatic environments, thriving equally well near the water surface using oxygen for respiration or at the bottom of the water using iron or manganese oxides as electron acceptors **[Myers and Nealson 1988]**. A multi-component branched electron transport system utilizing a variety of c-type cytochromes, reductases, iron-sulfur proteins and quinines is thought to be the reason for this respiratory versatility **[Richardson 2000]**. The complete genome sequence of *S.oneidensis* is 5.0 mega-bases in size and encodes a predicted 4758 proteins **[Heidelberg *et al* 2002]**.

This chapter analyses the proteomes of *S.oneidensis* and *G.sulfurreducens*, identifying the multiheme cytochromes in each, predicting the structures of a subset of these cytochromes and using these structures to infer putative functional properties of these proteins.

3.2 Materials and Methods

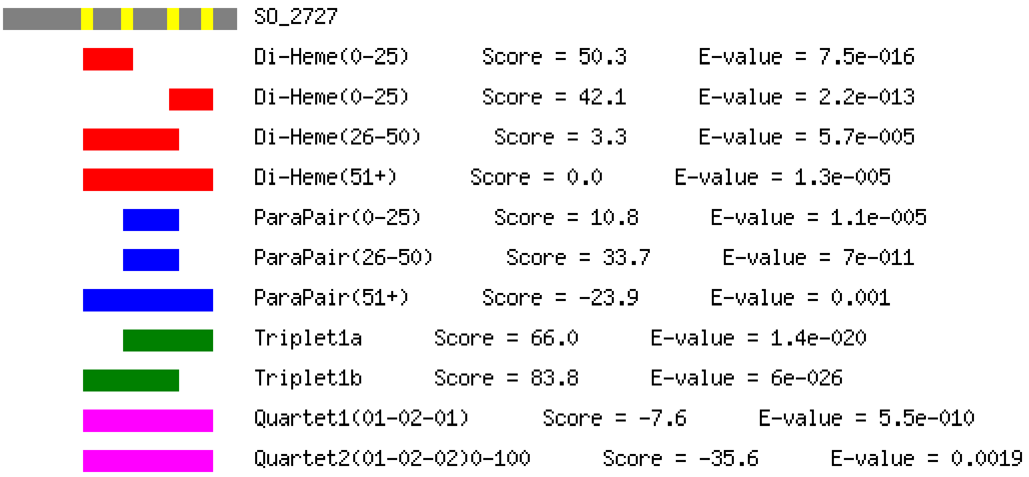
3.2.1 Identification of multiheme cytochromes

Identification of multiheme cytochromes was performed using the Comprehensive Microbial Resource (CMR) server **[Peterson *et al* 2001]** to search for specific sequence motifs in the *Shewanella oneidensis* and *Geobacter sulfurreducens* genomes. Multiheme cytochromes were identified using the search term [C].{2}[C][H], which corresponds to the known heme binding CXXCH motif. This process was repeated for both the CXXXCH ([C].{3}[C][H]) and CXXXXCH ([C].{4}[C][H]) motifs, both of which have been observed as heme binding motifs **[Aubert *et al* 1998, Pattarkine *et al* 2006]**. The resulting lists of sequences were interrogated to remove all cytochromes containing only a single heme, leaving only the multiheme cytochromes.

3.2.2 Automation of prediction methodology

In order to streamline the structural prediction methodologies developed in Chapter 2 it was necessary to automate the prediction process, as the manual searching of each sequence with each Hidden Markov Model (HMM) would be very time consuming. This automation was achieved by writing a PERL script to read in a sequence file containing a single or multiple sequences, execute hmmsearch (part of the HMMER package **[Eddy 1998]**) and use it to search against the sequence(s) in this file with all the available HMMs or a subset if preferred (e.g. just the HMMs derived from the di-heme elbow family or cytochrome c3-like SCOP families). This program outputs a single file with a summary of the hits for each of the HMMs against the sequence(s) of interest containing the identity of the HMM responsible for each hit, the start and end residue numbers of the hit in the protein sequence, the score and the E-value.

A second PERL script takes this information and plots the position of the HMM hits against the protein sequence in a graphical format, to give an “at-a-glance” overview of the prediction for each sequence (Figure 3.1). These outputs can be used to identify target protein sequences that are likely to be good candidates for prediction, i.e. a protein with significant HMM coverage will give a more complete structural prediction than a protein where there are very few or contradicting HMM hits.



```
SO_2727
Di-Heme(0-25)      Score = 50.3     E-value = 7.5e-016
Di-Heme(0-25)      Score = 42.1     E-value = 2.2e-013
Di-Heme(26-50)      Score = 3.3     E-value = 5.7e-005
Di-Heme(51+)      Score = 0.0     E-value = 1.3e-005
ParaPair(0-25)      Score = 10.8     E-value = 1.1e-005
ParaPair(26-50)      Score = 33.7     E-value = 7e-011
ParaPair(51+)      Score = -23.9     E-value = 0.001
Triplet1a      Score = 66.0     E-value = 1.4e-020
Triplet1b      Score = 83.8     E-value = 6e-026
Quartet1(01-02-01)      Score = -7.6     E-value = 5.5e-010
Quartet2(01-02-02)0-100      Score = -35.6     E-value = 0.0019
```

Figure 3.1 – An example output of the automated HMM searching software. A schematic representation of the primary sequence is displayed as a grey bar, with the position of the hemes highlighted in yellow. The hits found with the di-heme elbow (red), parallel pair (blue), triplet (green) and quartet (magenta) based HMMs are shown as coloured bars, the scores and E-values are also displayed.

3.2.3 Creation of 3D structural models from selected sequence targets

Once likely targets had been identified by the method set out in section 3.2.1, the same methodologies set out in Chapter 2 were used to create a 3D model, using the JAVA code to build a heme substructure and MODELLER to incorporate the identified polypeptide templates.

3.2.4 Structure validation

All structural validation was carried out using the PROCHECK **[Laskowski *et al* 1993]** software at the Joint Center for Structural Genomics (JCSG) server (http://www.jcsg.org/prod/scripts/validation/sv2.cgi).

3.3 Results

3.3.1 Breakdown of multiheme cytochromes found in *Shewanella oneidensis*

In total 34 putative multiheme cytochromes with either CX_2CH, CX_3CH or CX_4CH heme binding motifs were identified in the *Shewanella oneidensis* genome, ranging from 2-10 hemes in size. A graphical breakdown of the distribution of the identified protein sequences can be seen in figure 3.2.

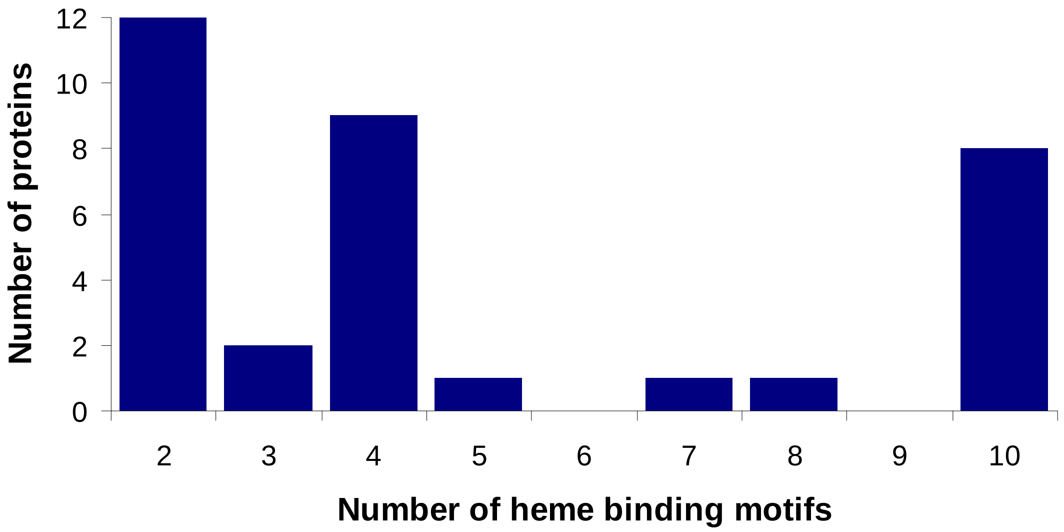


Figure 3.2 – The distribution of putative multiheme cytochromes in the *Shewanella oneidensis* genome, this graph shows the number of heme binding motifs against the number of proteins with that specific numbers of motifs.

3.3.2 Breakdown of multiheme cytochromes found in *Geobacter sulfurreducens*

In total, 85 multiheme cytochromes were identified in the *Geobacter sulfurreducens* genome, ranging from 2-35 hemes in size. A graphical breakdown of the distribution of the identified protein sequences can be seen in figure 3.3.

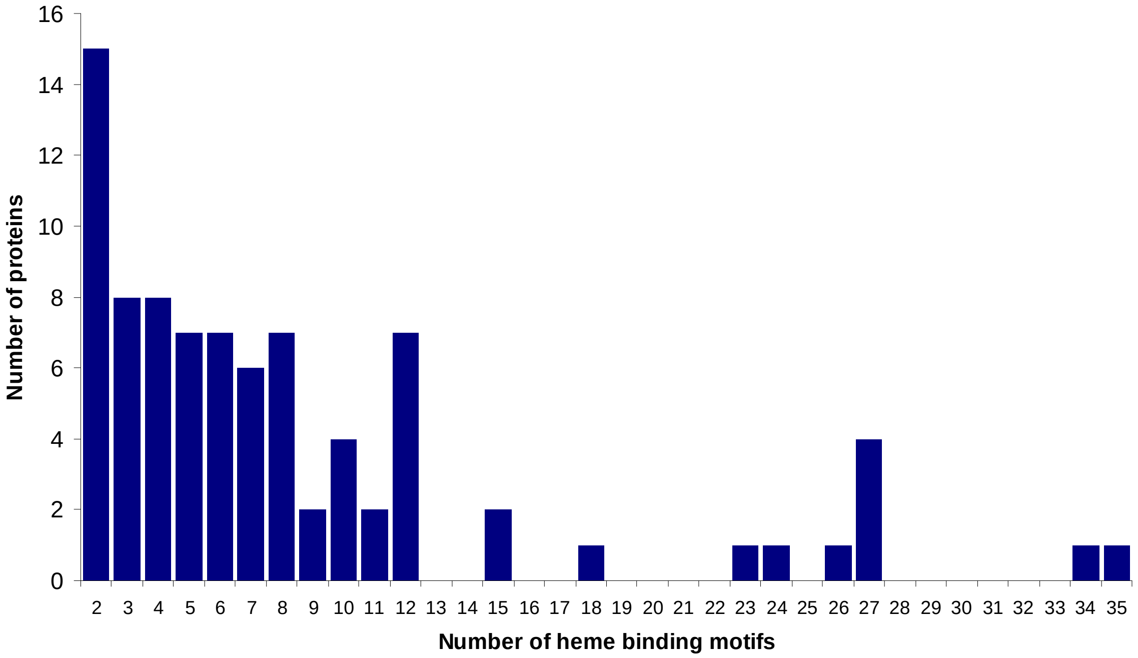


Figure 3.3 – The distribution of multiheme cytochromes in the *Geobacter sulfurreducens* genome, this graph shows the number of heme binding motifs against the number of proteins with that specific numbers of motifs.

3.3.3 Structure predictions for multiheme cytochromes from *Shewanella oneidensis* and *Geobacter sulfurreducens*

Heme substructure and polypeptide template predictions were made for each multiheme cytochrome identified in *Shewanella oneidensis* and *Geobacter sulfurreducens* using the automated HMM software. From these predictions, a subset were chosen for a more thorough examination. These proteins were then selected either for their relevance to other research groups at UEA (in the case of MtrA, MtrC and MtrF), the quality of their initial prediction (in the case of GSU_0357) or their unusual nature (in the case of GSU_2210).

3.3.3.1 Structure prediction for the *Shewanella oneidensis* decaheme cytochrome MtrA

The results of the HMM searches on the MtrA sequence resulted in a prediction for the heme substructure encompassing all ten hemes. This substructure consisted of four di-heme elbow motifs and five parallel pair motifs arranged alternately in series along the length of the protein sequence. Templates for polypeptide structures were identified between residues 67 and 317, accounting for 84 % of the total protein sequence (Figure 3.4A). This prediction had an overall score of 9.42 and E-value of 9.75x10^{-5}. Validation of the final predicted 3D-structure showed that only 1.4 % of non-glycine and non-proline residues fall into the disallowed regions of the Ramachandran plot (Figure 3.4B).

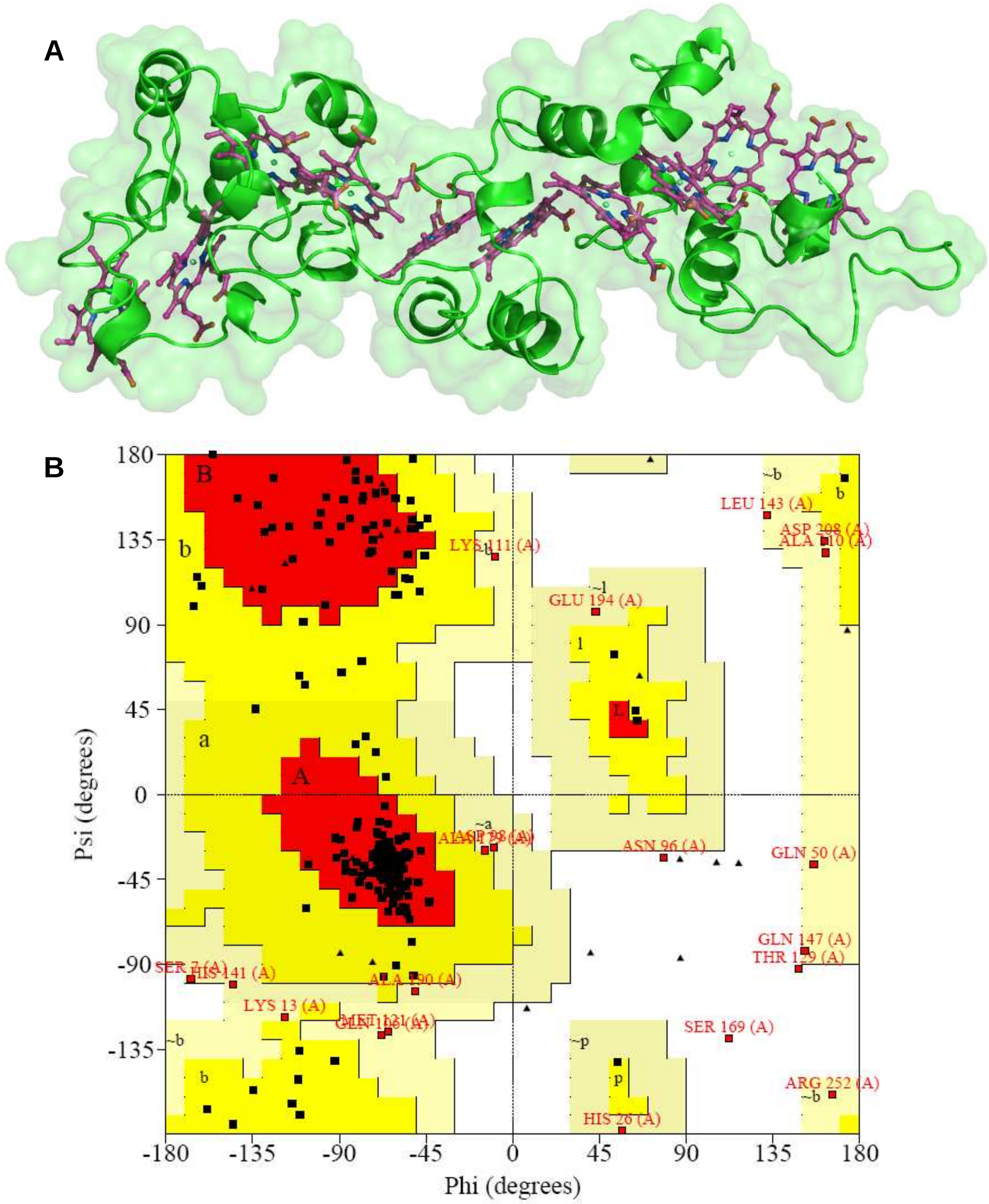


Figure 3.4 – (A) A model for the 3D-structure of MtrA, covering all ten hemes and residues 67-317 (84% of the total structure). The hemes are displayed as magenta sticks, the polypeptide structure in cartoon format and a transparent protein surface is also shown. **(B)** A Ramachandran plot for the MtrA structure. 74.3% of residues are in the most favoured regions, 16.7% in additionally allowed regions, 7.7% in generously allowed regions and 1.4% in disallowed regions.

3.3.3.2 Structure predictions for the *Shewanella oneidensis* decaheme cytochromes MtrC and MtrF

MtrC and MtrF are homologous decaheme cytochromes (27.2 % identity, 39 % similarity), both of which have their heme binding CXXCH motifs separated into two groups of five separated by 197 amino acids in MtrC and 170 in MtrF. MtrC is of particular interest as it has been shown to form a complex with MtrA and MtrB **[Hartshorne *et al* 2009]**. The results of the HMM searches on the MtrC sequence predicted a heme substructure for all ten hemes, with two di-heme elbows and two parallel stacking pair motifs arranged sequentially in series for each group of five hemes. The first heme domain begins with a di-heme elbow motif and the second with a parallel pair motif (Figure 3.5).

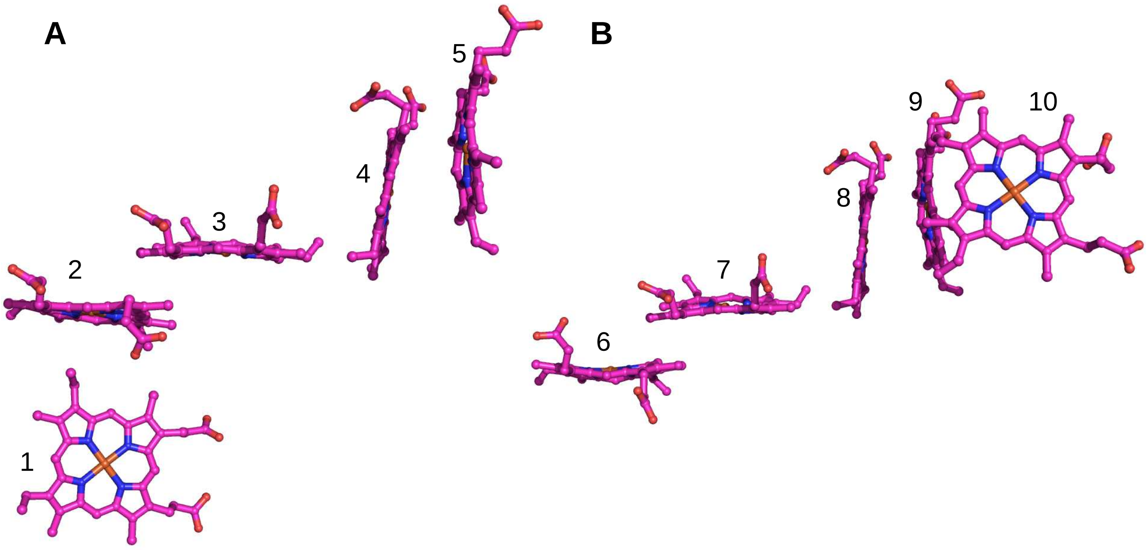


Figure 3.5 – The heme substructure for the first **(A)** and second **(B)** heme domains of MtrC. The heme numbers refer to the order in which the heme binding CXXCH motifs appear in the sequence.

Assignment of templates for the MtrC polypeptide structure prediction was of limited success, with templates identified for the regions covering hemes 1-2, 3-4 and 6-7 only (Figure 3.6), equating to only 11 % of the complete structure.

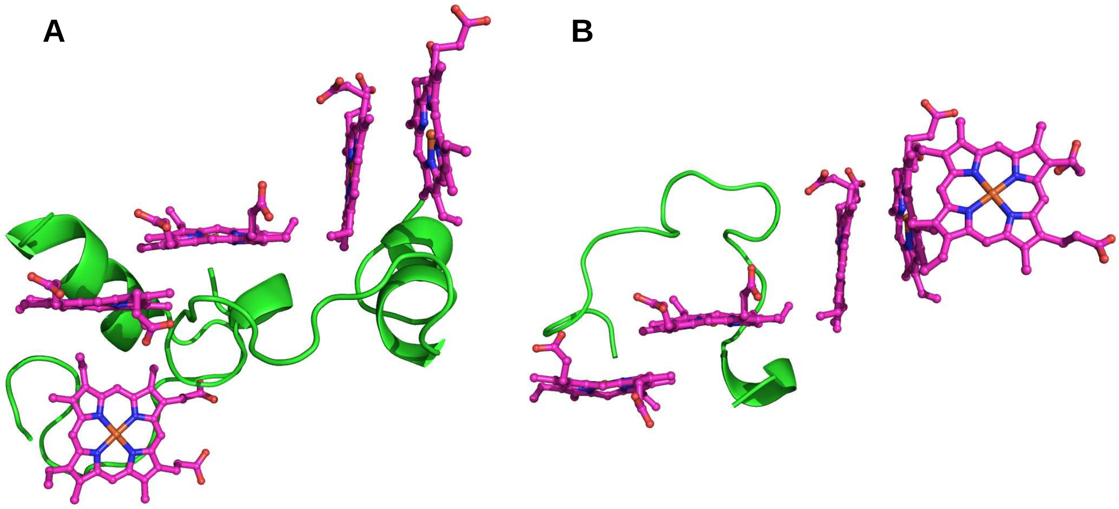


Figure 3.6 – The predicted polypeptide structure, displayed in cartoon format, for; **(A)** MtrC heme domain 1 and **(B)** MtrC heme domain 2, covering 11% of the total sequence.

In an attempt to get a more complete structural prediction for this type of decaheme cytochrome, a model was built for the homologous protein MtrF. The HMM searches once again predicted a heme substructure for all ten hemes, this substructure was found to be the same as that predicted for MtrC (Figure 3.5). Templates for polypeptide structure were again limited, although there were more templates for polypeptide structure identified for the first heme domain, with hemes 1-3 and 4-5 covered, although only hemes 6-7 where covered from the second domain (Figure 3.7), equating to 16 % of the complete structure in total.

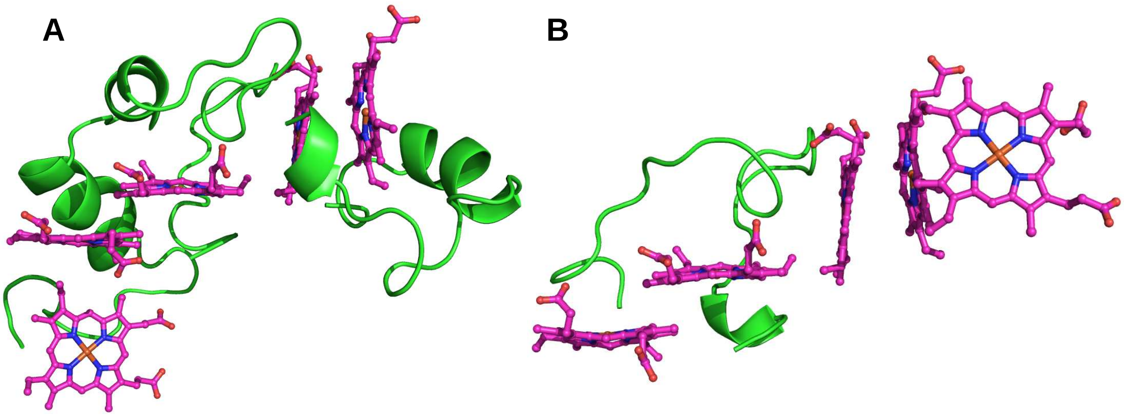


Figure 3.7 – The predicted polypeptide structure, displayed in cartoon format, for; **(A)** MtrF heme domain 1 and **(B)** MtrF heme domain 2, covering 16% of the total sequence.

The overall scores for the MtrC and MtrF predictions were 0.11 and 4.44 respectively, while the overall E-values were 0.17 and 0.12 respectively. This would suggest that these are not highly accurate predictions

3.3.3.3 Structure prediction for the *Geobacter sulfurreducens* 27-heme cyochrome, GSU_2210

The results of the HMM searches against the GSU_2210 sequence resulted in a prediction for the heme substructure of all 27 hemes. This substructure consisted of nine repeats of the three heme cytochrome c_7 domain, linked by eight parallel stacking pair motifs, templates for polypeptide were also identified between residues 30 and 684, accounting for 95 % of the total protein sequence (Figure 3.8). The prediction had an overall score of 21.36 and E-value of 0.014. The structure had a helix-like superstructure, with six c_7 domains per turn of the helix, with a diameter of 67.3 Å.

Validation of this predicted structure identified only 4.8 % of non-glycine and non-proline residues as falling into the disallowed regions of the Ramachandran plot (Figure 3.8C), most of which fall in the regions of the structure that link every other c7 domain, i.e. the 2nd, 4th, 6th and 8th parallel pairs of the structure.

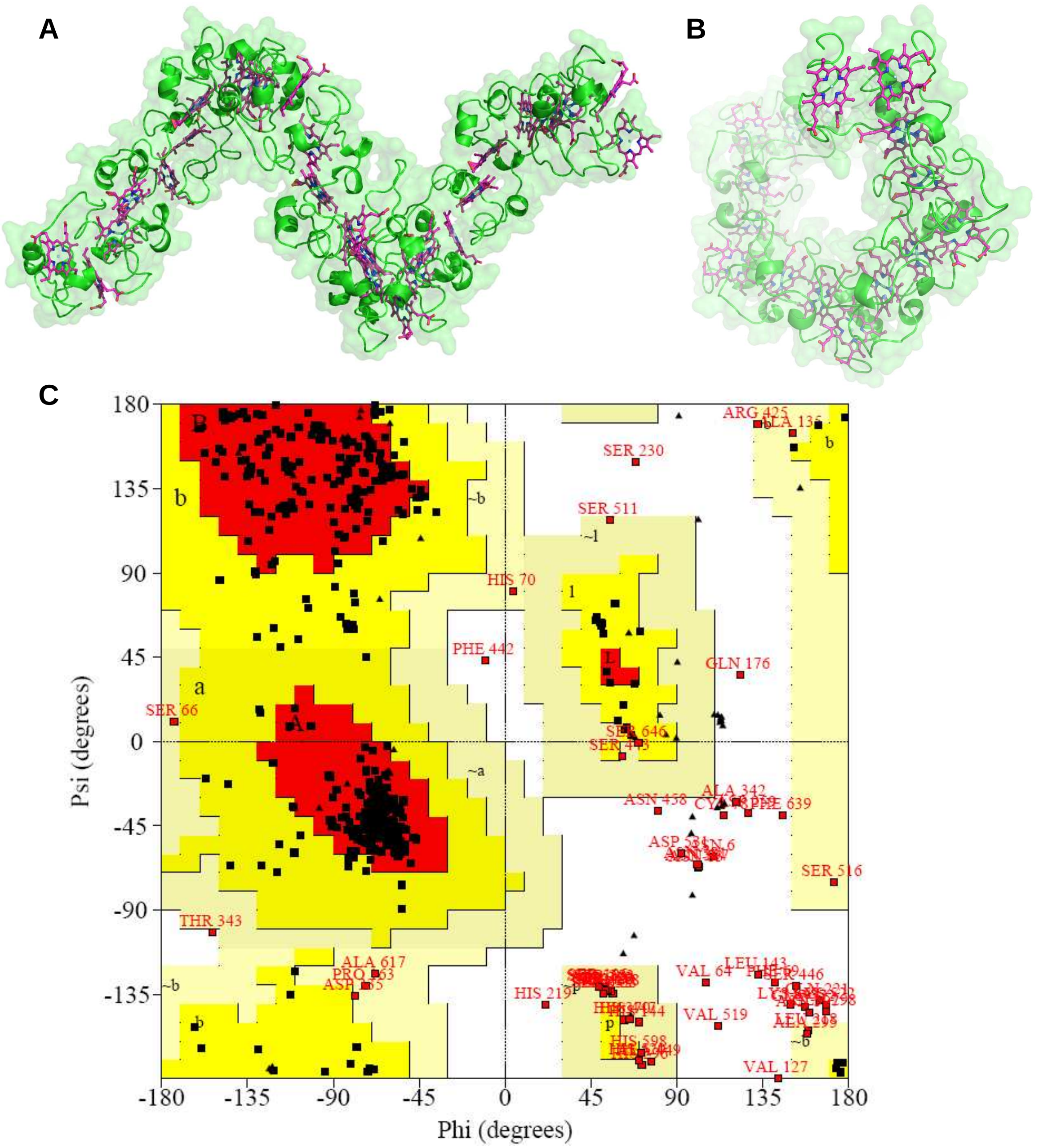


Figure 3.8 – The first predicted model for GSU_2210, displaying **(A)** the side-on view and **(B)** the view down the "super-helix" created by the global fold of the protein, covering all 27 hemes and residues 30-684 (95% of the total structure). The hemes are displayed as magenta sticks, the polypeptide structure in cartoon format and a transparent protein surface is also shown. **(C)** A Ramachandran plot for the GSU_2210 structure. 75.6% of residues are in the most favoured regions, 14.6% in additionally allowed regions, 5.0% in generously allowed regions and 4.8% in disallowed regions.

Data provided subsequently by P.R.Pokuluri (Biosciences Division, Argonne National Laboratory) concerning the structure of the homologous protein GSU_1996 that contains six repeats of the cytochrome c_7 domain (see chapter 2 section 2.3.10.2), suggested a discrepancy between the predicted structure of GSU_2210 and the true structure, this change centred around a difference in heme packing between hemes 6 and 7 of the homologous GSU_1996 structure (the heme pair linking the 2nd and 3rd c_7 domains. With this in mind, a second model for GSU_2210 was constructed using the alternate heme packing motif identified in GSU_1996 to replace the original prediction of parallel heme pairs between every other cytochrome c7 domain repeat in the structure (Figure 3.9). This resulted in a more ‘open’ helix-like superstructure than the previous model, with the nine c7 domains not being enough to complete a turn of the helical superstructure. In fact 10 would be needed for one complete turn of the helical superstructure, which would have an increased diameter of 101.8 Å.

Validation of this predicted structure identified only 2.3 % of non-glycine and non-proline residues fall into the disallowed regions of the Ramachandran plot (Figure 3.9C).

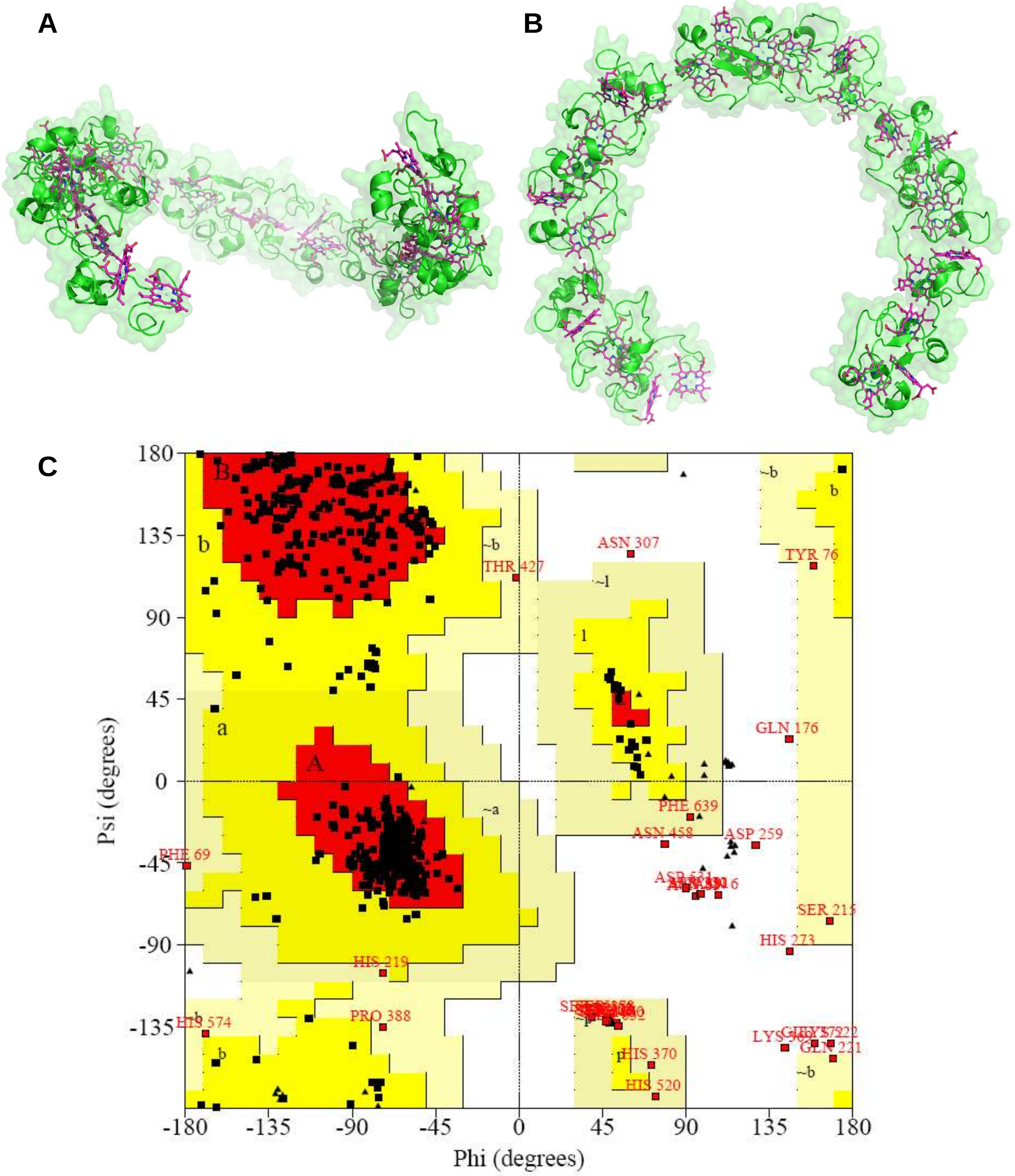


Figure 3.9 – The second predicted model for GSU_2210 built using the template for the novel heme pair, identified in GSU_1996, to link every third c7 domain. Displaying; **(A)** the side on and **(B)** the top down views, covering all 27 hemes and residues 30-684 (95% of the total structure). The hemes are displayed as magenta sticks, the polypeptide structure in cartoon format and a transparent protein surface is also shown. **(C)** A Ramachandran plot of the alternative GSU_2210 structure. 81.5% of residues are in the most favoured regions, 12.8% in additionally allowed regions, 3.4% in generously allowed regions and 2.3% in disallowed regions.

3.3.3.4 Structure prediction for the *Geobacter sulfurreducens* octaheme cytochrome GSU_0357

The results of the HMM searches on the GSU_0357 sequence predicted a heme substructure for all eight hemes. This substructure consisted of three di-heme elbow motifs, three parallel stacking pair motifs and an active site heme pair. Templates for polypeptide structure were identified between residues 18 and 399, accounting for 75 % of the total protein sequence (Figure 3.10A). The prediction overall score of 52.88 and E-value of 2.33x10^{-5}.Validation of this predicted structure identified only 0.6 % of non-glycine and non-proline residues as falling into the disallowed regions of the Ramachandran plot (Figure 3.10B).

A search of the Dali database using DaliLite V3 **[Holm and Rosenström 2010]** with the predicted GSU_0357 structure found it to be most similar to the various eight heme cytochrome c nitrite reductase structures present in the database, with the highest similarity to chain A of the *Thiolkalivibrio nitratireducens* nitrite reductase structure (PDB ID: 2OT4 **[Polyakov *et al* 2009]**) that has 49 % sequence identity and an RMSD of 0.6 Å to the GSU_0357 model.

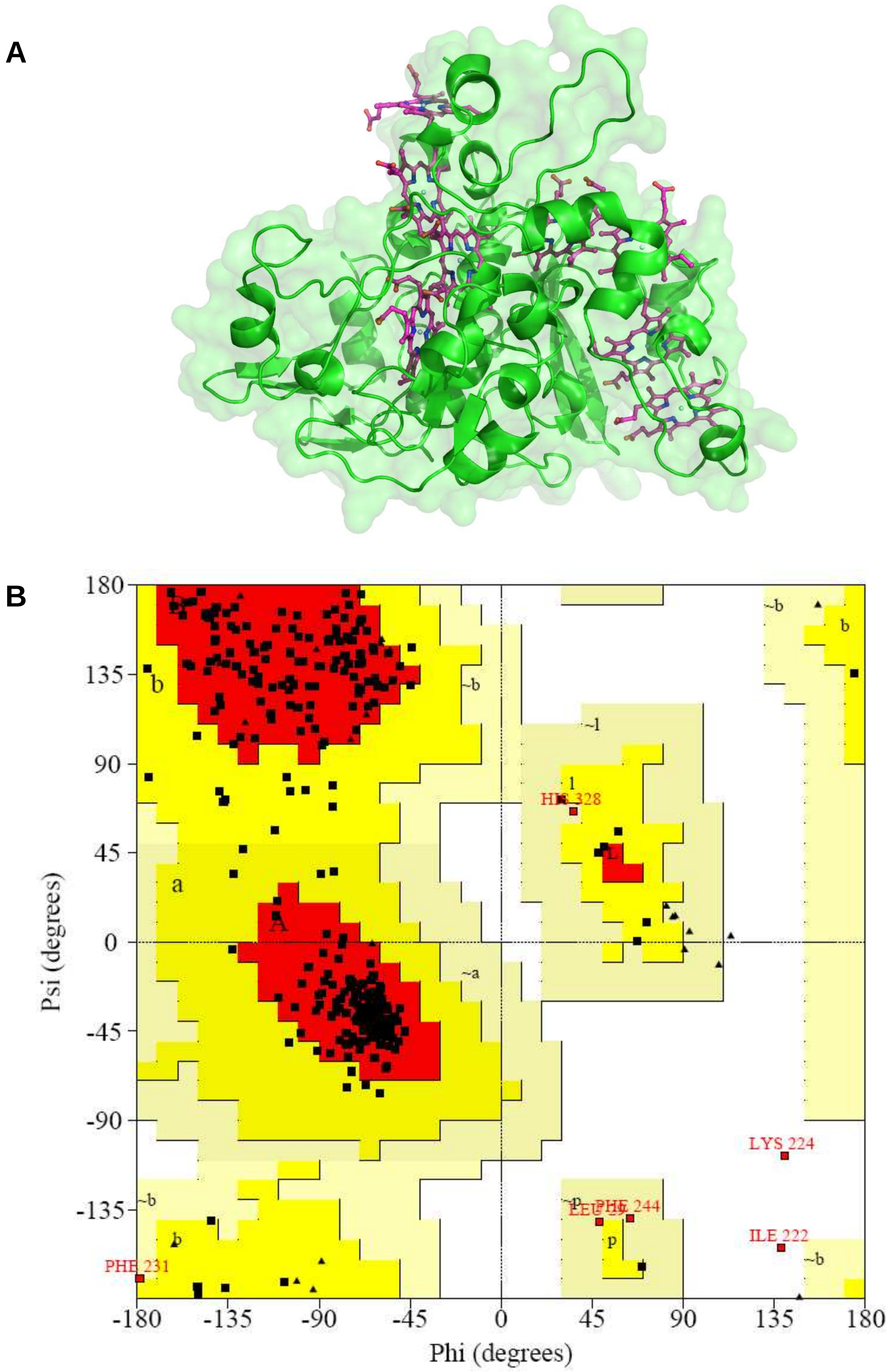


Figure 3.10 – (A) The predicted structure for GSU_0357, covering all eight hemes and residues 18-399 (75% of the total structure). The hemes are displayed as magenta sticks, the polypeptide structure in cartoon format and a transparent protein surface is also shown. **(B)** A Ramachandran plot of the MtrA structure. 86.0% of residues are in the most favoured regions, 12.2% in additionally allowed regions, 1.2% in generously allowed regions and 0.6% in disallowed regions.

3.3.3.5 Structural prediction for the *Geobacter sulfurreducens* five heme cytochrome GSU_3223

The results of the HMM searches on the GSU_3223 sequence predicted a heme substructure for all five hemes. This substructure consisted of two di-heme elbow motifs, two parallel stacking pair motifs arranged sequentially in series. Templates for polypeptide structure were identified between residues 88 and 183, accounting for 49 % of the total protein sequence (Figure 3.11A). The prediction overall score of 4.69 and E-value of 0.09. Validation of this predicted structure identified only 5.1 % of non-glycine and non-proline residues fall into the disallowed regions of the Ramachandran plot (Figure 3.11B).

A search of the Dali database using DaliLite V3 **[Holm and Rosenström 2010]** with the predicted GSU_3223 structure found it to be most similar to the various cytochrome c nitrite reductase NrfHA complex structures in the database, with the highest similarity to chain L of the Desulfovibrio vulgaris NrfHA complex (PDB ID: 2J7A **[Rodrigues *et al* 2006]**) that has 16 % sequence identity and an RMSD of 4.8 Å to the GSU_3223 structure.

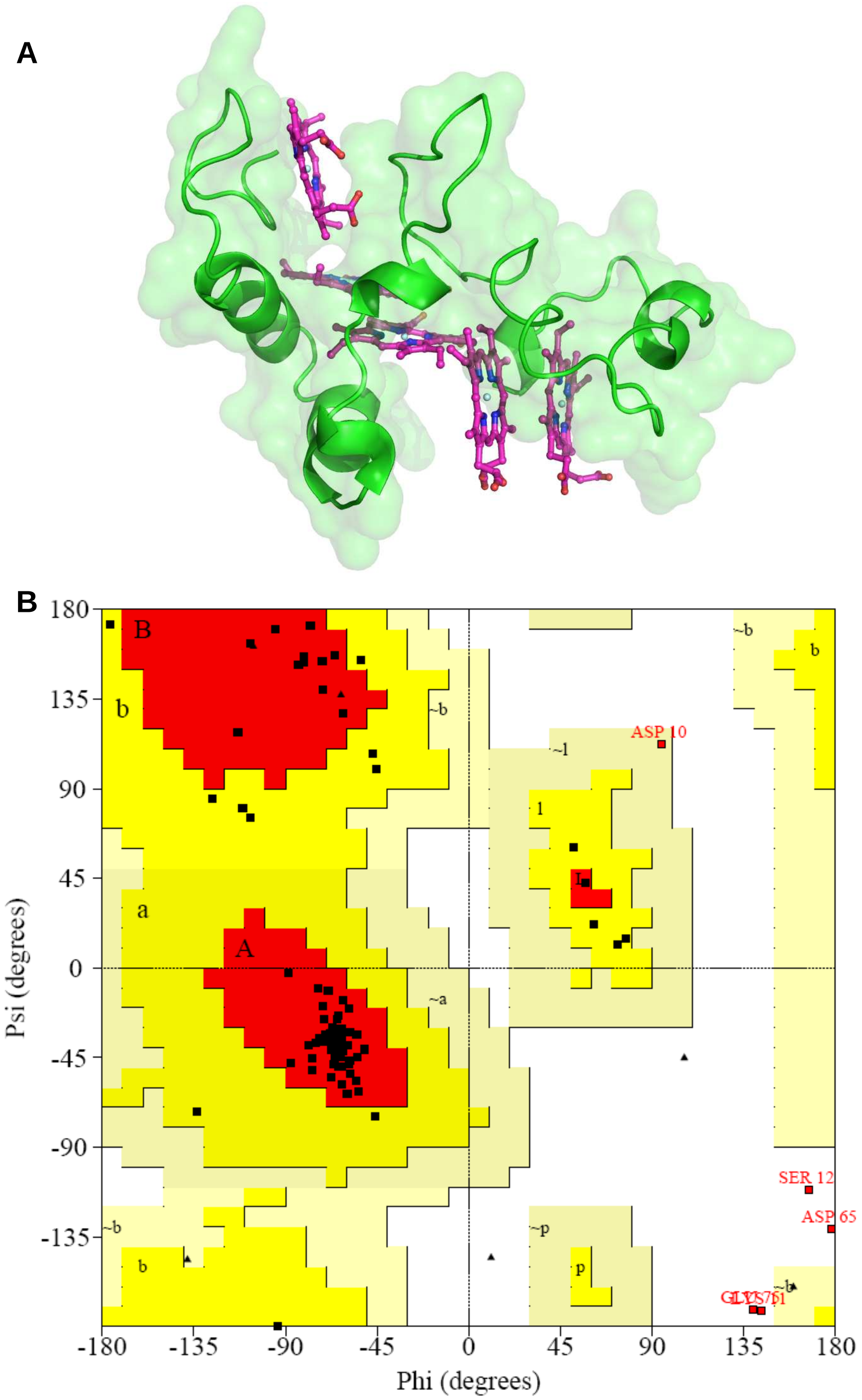


Figure 3.11 – (A) The predicted structure for GSU_3223, covering all 5 hemes and residues 88-183 (49% of the total structure). The hemes are displayed as magenta sticks, the polypeptide structure in cartoon format and a transparent protein surface is also shown. **(B)** A Ramachandran plot of the MtrA structure. 76.9% of residues are in the most favoured regions, 16.7% in additionally allowed regions, 1.3% in generously allowed regions and 5.1% in disallowed regions.

A BLAST search performed using the GSU_3223 sequence identified a region to the N-terminal side of the heme binding region that was homologous to the copper binding region of the copper specific repressor CsoR from Mycobacterium tuberculosis (PDB ID: 2HH7 **[Liu *et al* 2007]**) (34 % identity over a 37 residue region). In the CsoR structure the copper is bound to the protein by a cysteine residue found at the N-terminal end of the second α-helix of a three α-helix domain (Figure 3.12A).

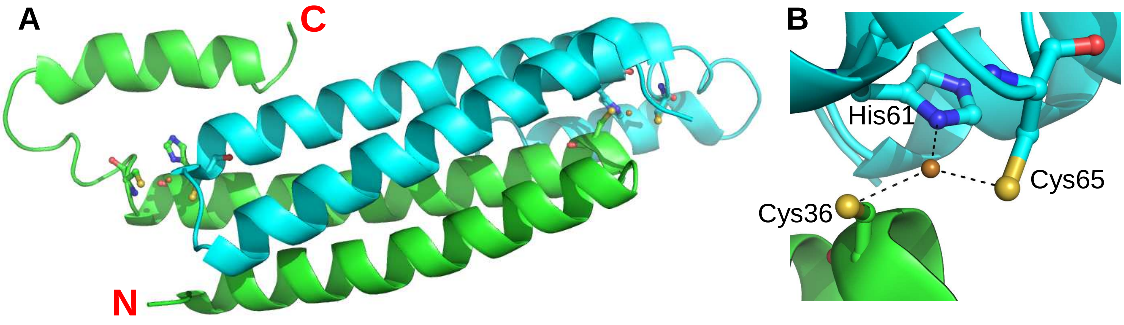


Figure 3.12 – (A) The structure of a monomer of copper bound CsoR showing the position of copper binding at the N-terminal end of the second α-helix and the residues involved in copper coordination in the homodimer. **(B)** The residues involved in Cu(I) coordination at the dimer interface. The polypeptide backbone is displayed in cartoon format, the copper binding residues in stick format, the Cu(I) ion as a sphere and the N and C termini are also labelled for one of the monomers.

A secondary structure prediction for GSU_3223 (calculated using PSI-PRED **[Jones 1999, Bryson *et al* 2005]**) identified three N-terminal α-helices in positions equivalent to those in CsoR, although there is a large 25 residue insertion between the first and second α-helix. Also conserved between GSU_3223 and CsoR are the copper binding cysteine (Cys36), the preceeeding tyrosine and proceeding valine and aspartic acid residues, suggesting GSU_3223 could potentially have copper binding properties. Although in order to bind copper in a fashion similar to that observed in CsoR, GSU_3223 would need to dimerise forming an anti-parallel four helix bundle with a trigonally coordinated copper complex stabilised by two cysteine and one histidine residues (Figure 3.12B) **[Liu *et al* 2007]**, which would be unlikely given the lack of residues homologous to His61 and Cys65 from CosR in the GSU_3223 sequence. However, there are two cysteine residues at the C-terminal end of the first predicted α-helix of GSU_3223 that could potentially provide sufficient ligands to bind the copper if the helical packing is similar to that observed in CosR, using these residues would also not require GSU_3223 to for a homodimer to bind the copper.

3.4 Discussion

3.4.1 What can be learnt from the predicted structure of MtrA?

The decaheme cytochrome MtrA is known to interact with the membrane bound protein MtrB **[Ross et al 2007, Hartshorne *et al* 2009]** and it has also been shown to interact directly with the extracellular decaheme cytochrome MtrC **[Hartshorne *et al* 2009]**. The current schematic for the MtrCAB complex (Figure 3.12), based on the observations of Hartshorne *et al* **[2009]**, is that the complex receives electrons from the quinol pool via an electron transfer protein, such as CymA, these electrons are transferred thorough MtrA to MtrC using MtrB to hold the complex together. The overall score (9.42) and in particular the overall E-value ($9.75x10^{-5}$) for the predicted model of MtrA suggests it is a reasonably accurate prediction and thus can be used to help shed light on the nature of this complex.

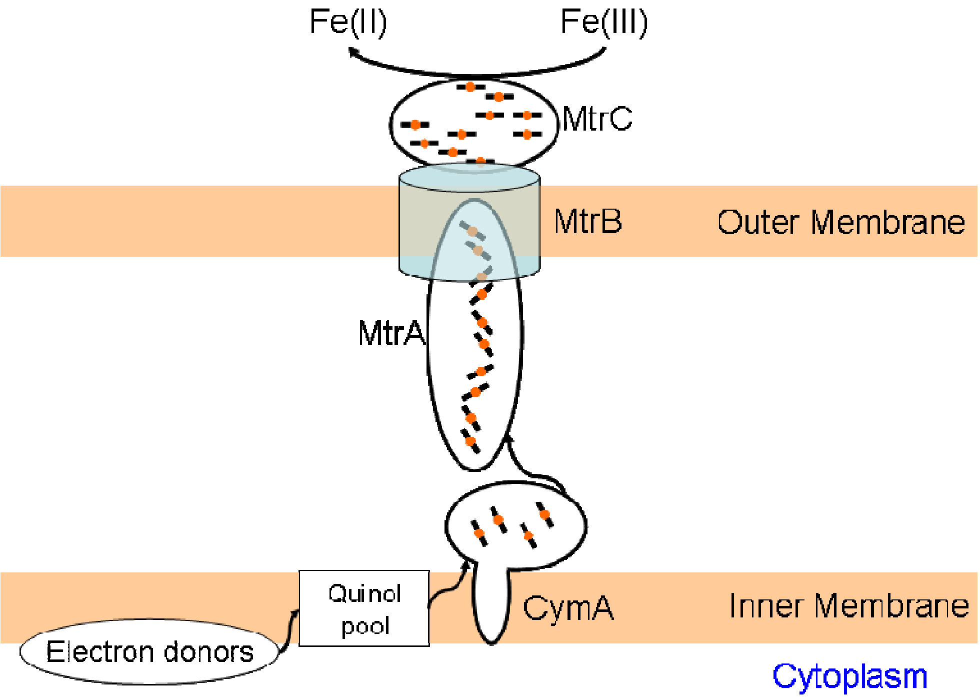


Figure 3.12 – A schematic for the electron transfer complex MtrCAB and the role it plays in the transfer of across the outer membrane, showing MtrA inserting into MtrC, allowing it to transfer electrons to MtrC for iron reduction.

FepA **[Buchanan *et al* 1999]** is a 22 strand β-barrel membrane protein with a cavity diameter of 30 Å. If we divide the diameter by the number of β strands we get a contribution of 1.36 Å from each strand. MtrB is a 28 strand porin, so by multiplying the contribution from each β strand by the number of β strands, we get an approximate cavity diameter of 38 Å for MtrB. The predicted structure for MtrA has a diameter of ~30 Ǻ for ~68 Å of the protein from the C-terminal end and ~40 Å for the remaining 29 Å of the length of the protein at the N-terminal end (Figure 3.13). This suggests that MtrA may be able to insert into the β-barrel of MtrB, but only for an approximate 68 Å region incorporating the C-terminus, the remaining protein covering the N-terminal end of MtrA appears too wide to insert into MtrB.

Although it should be noted that 34 residues from the N-terminus and 12 residues from the C-terminus are not included in the model for MtrA as there was no template for the polypeptide in these regions. These missing residues are likely to have some influence on the dimensions of the model, although the small number of residues missing at the C-terminal shouldn't affect the dimensions of the protein to the point that it no longer fits into the proposed MtrB β-barrel.

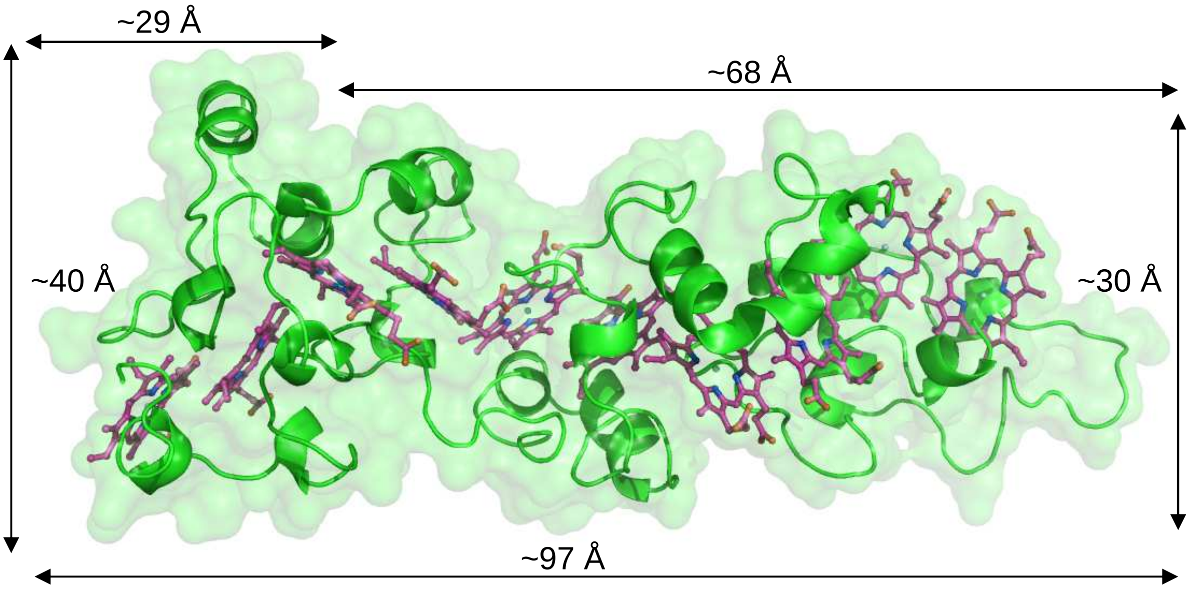


Figure 3.13 – The approximate dimensions of the predicted structure for MtrA, showing the ~68 Å portion of the protein at the C-terminal end of the protein with a sufficiently small diameter to fit into the lumen of the MtrB channel. The hemes are displayed as magenta sticks, the polypeptide structure in cartoon format and a transparent protein surface is also shown.

This hypothetical structure for an MtrAB complex fits in with the known role of the MtrCAB complex as a complex responsible for the transport of electrons across the outer membrane **[Ross *et al* 2007]**. The close packing of hemes in the MtrA structure would allow electron transfer along the length of the protein via the heme groups. The dimensions of MtrA relative to the predicted dimensions of MtrB would allow MtrA to insert into MtrB to reduce the electron transfer distance between MtrA and MtrC, resulting in electron transfer between MtrA and MtrC, using MtrB as a sheath (Figure 3.12).

3.4.2 What can be learnt form the predicted structures for MtrC and MtrF

The predicted structures for MtrC and MtrF suggest that both proteins incorporate close packings of hemes, providing them with the potential for electron transport properties, which would be expected given the role presumed of MtrC as a reducing agent for external electron acceptors, such as Fe(III) **[Shi *et al* 2007]**. The protein sequence would suggest these hemes are grouped into two domains of five hemes due to the 185 residue gap between the 5th and 6th hemes, but as there was no prediction for this interlinking sequence, it was unclear from the analysis whether these two heme domains are in contact with each other (i.e. within the minimum electron transfer distance of 14 Å). The overall scores and E-values for the MtrC and MrtF structures suggests the predictions are not that close to their actual structures, this is confirmed by an analysis of early crystallographic data for MtrC.

A novel heme substructure has been proposed for MtrF using crystallographic data collected and refined to 3.5 Å by Tom Clarke (UEA) (Figure 3.14A). The heme substructure identified in this medium-low resolution dataset (Figure 3.14B) does contain some previously identified heme pair packing motifs. For example, the packing between heme pairs 1-2 and 4-5 appear to be parallel pair-like, while the packing between hemes 2 and 3 appear to be di-heme elbow-like. However, the heme packing between hemes 3 and 4 is novel, thus making it impossible to predict by the methods set out in this thesis.

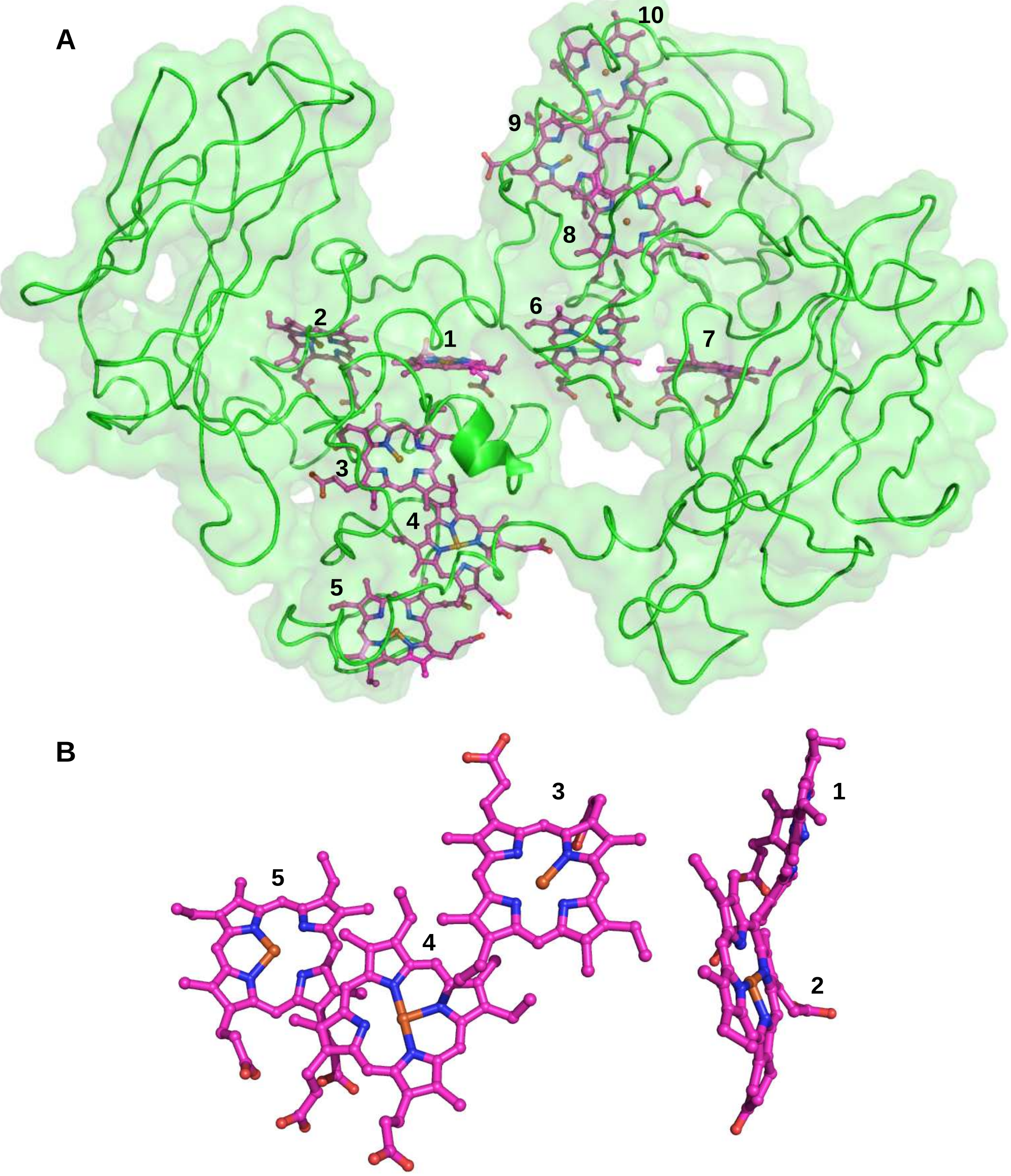


Figure 3.14 – Structures of **(A)** the polyalanine model for MtrF, provided by Tom Clarke (UEA) and **(B)** the heme substructure for the 1st heme domain. The hemes are numbered in the order in which their CXXCH motifs appear in the protein sequence. The hemes are displayed as magenta sticks, the polypeptide structure in cartoon format and a transparent protein surface is also shown.

3.4.3 What can be learnt from the GSU_2210 structural model?

The predicted models for GSU_2210 form what could be described as “molecular wires”, close packing of hemes that allow electron transfer over long distances. The complete structure for proteins homologous to GSU_2210 that are built from repeating linked cytochrome c_7 domains have been difficult to crystallise, with currently only the structure of individual domains having been published **[Pokkuluri *et al* 2004]**. This difficulty in crystallising the complete proteins may point towards flexibility between the individual c_7 domains, a flexibility that has been suggested by the predicted models for GSU_2210. Both of these models appear to be plausible solutions for the GSU_2210 structure, with the first prediction based solely on HMM hits having and overall score and E-value that suggest it is a reasonably accurate prediction and both this prediction and the second that incorporated the novel packing identified in GSU_1996 by Pokkuluri *et al* performing well in structural validation checks. This would suggest that the final structure for GSU_2210 may have a certain amount of flexibility, allowing for subtle changes in the packing of the c_7 domains and thus a flexible pathway for electron transport, although it is unlikely the degree of flexibility will be to the extent shown in the two predicted structures.

It has also been proposed that cytochromes with high heme contents, such as the 27 heme protein GSU_2210 and dodecaheme protein GSU_1996, can act as capacitors that enhance the electron storage capacity of the bacterial periplasm **[Morgado *et al* 2009]**. It is thought this capacitance can permit continued electron flow from the inner membrane to the periplasm, generating the energy that could be used to create a proton motive force to power the flagella motors, moving the organism to locate new external terminal electron acceptors once the current supply becomes exhausted **[Esteve-Nunez *et al* 2008]**. The predicted periplasmic location (based on the location of the protein in separated cell fractions **[Ding *et al* 2006]**), predicted extended helical structure and apparent flexibility of GSU_2210 are likely to aid in the distribution of electrons about the organism, fitting in with Morgado *et al’s* hypotheses.

3.4.4 What can be learnt from the GSU_0357 structural model?

The structure for GSU_0357 is an example of an application of the HMM methodology on an active site containing protein, in this case a cytochrome c nitrite reductase. The high overall score, low overall E-value and an RMSD value of 0.13 Å from a superposition of the predicted model of GSU_0357 with the crystal structure of the homologous *Thiolkalivibrio nitratireducens* cytochrome c nitrite reductase suggest the prediction is very accurate and highlights the structural conservation between this family of enzymes.

3.4.5 What can be learnt from the GSU_3223 structure?

The structure for GSU_3223 is novel example of a protein that appears to contain both a chain of C-type hemes, as suggested by the presence of the five CXXCH motifs and by positive hits against heme packing HMMs, as well as a CsoR-like copper binding domain N-terminal to the heme binding domain.

The presence of the CsoR-like copper binding domain and the predicted location of the three N-terminal α-helices, also found in the CosR structure, in the GSU_3223 sequence suggest it could potentially have copper-binding properties, although the absence of the other copper binding cysteine and histidine residues at the C-terminal end of the second α-helix would dispute this. However, the GSU_3223 sequence does contain two cysteine residues at the C-terminal end of the first α-helix, which would be in close proximity to the copper binding site if the helical packing is homologous to that observed in CsoR, which could potentially provide the necessary ligands for the copper.

There are currently no structures for proteins homologous to GSU_3223 available in the PDB, therefore, despite it’s relatively low overall score (4.69) and high E-value (0.09), this prediction does shed some light on the basic structure of this novel cytochrome, which would certainly be worth further investigation.

Chapter 4 - The structures of two stoichiometries of the copper chaperone CopZ and mechanistic insights into Cu(I) transfer between CopZ and its cognate Cu(I)-transporting P-type ATPase, CopA

4.1 Introduction

A range of distinct Cu(I)-binding forms of CopZ from *B.subtilis* have been determined in solution by Kihlken *et al*. Using UV visible absorbance spectroscopy and analytical ultracentrifugation (AUC) they identified three distinct dimeric forms of CopZ containing; 1, 2, and 3 Cu(I) ions respectively **[Kihlken *et al* 2002]**. These findings were based on absorbance changes at 265nm during addition of Cu(I) (in the form of CuCl) to apo-CopZ (in aliquots of ~0.07 Cu(I) per CopZ monomer) that identified three distinct phases of binding at the 0.5, 1.0 and 1.5 Cu(I) per monomer levels, and AUC experiments carried out with each of the three distinct copper loaded species that indicated the formation of dimers in solution **[Kihlken *et al* 2002]**.

This chapter reports the results of crystal structure analyses on two of the Cu(I)-binding forms, the 1 Cu(I) dimer and the 2 Cu(I) dimer, although interestingly these were not the forms found in the crystal structures. The 2 Cu(I) dimer gave a crystal structure with a homodimer containing a tetranuclear copper cluster (hereafter referred to as $Cu_4(CopZ)_2$) and the 1 Cu(I) dimer gave a crystal structure with a homotrimer containing a trinuclear cluster (hereafter referred to as $Cu_3(CopZ)_3$). These structures are analysed and compared with other copper transport proteins.

An additional experiment was also performed based on work by Einsle *et al* **[Einsle *et al* 2007]** to ascertain the oxidation state of the four coppers in the tetranuclear cluster of the $Cu_4(CopZ)_2$ structure. This involved performing a fluorescence scan on a $Cu_4(CopZ)_2$ crystal and collecting multiple datasets across the copper X-ray absorption edge and collectively refine the anomalous scattering factors of these datasets.

After unsuccessful attempts to acquire a crystal structure from CopA, a homology model was created to analyse potential methods for CopA's interaction with CopZ. These analyses were based on substituting CopZ monomers from the $Cu_4(CopZ)_2$ and $Cu_3(CopZ)_3$ structures with the homology model for CopA and examining the properties of the protein-protein interface.

4.2 Materials and Methods

4.2.1 The structure of the P1 crystal form of $Cu_4(CopZ)_2$

4.2.1.1 Crystallisation

A protein solution of CopZ, determined by luminescence spectroscopy to contain a dimeric form of CopZ binding two Cu(I) ions (data not shown), was provided by Liang Zhou (UEA). Previous work with this form of the protein had shown that crystals would grow in a solution containing; 0.1M sodium acetate pH 4.6, 0.2 M $CaCl_2$ and 30 % (v/v) propan-2-ol, at a temperature of 4 ℃.

Crystals were grown using the hanging drop vapour diffusion technique, utilising drops containing 2 µl of the concentrated CopZ solution and 2 µl of the crystallisation solution, equilibrated against a 1000 µl reservoir of the crystallisation solution at a temperature of 4℃. Crystals of typical dimensions 250-450 µm grew from these experiments within 1-3 days. However, it was found that crystals grown under these conditions rapidly degraded on exposure to air, due to evaporation of isopropanol.

In an attempt to reduce exposure to air, a vapour batch method for crystallisation adapted from that used by Mortuza *et al* **[Mortuza *et al* 2004]** was implemented (Figure 4.1). A Terasaki plate (Molecular Dimensions Ltd) was glued into a square Petri dish; 8 ml of silicone oil was poured over the wells, 4 µl of a solution containing a 1:1 mixture of concentrated CopZ solution (14.5 mg ml^{-1}) and crystallisation solution was pipetted into the wells, under the oil layer. 30 ml of a 30 % (v/v) propan-2-ol solution was poured into the Petri dish, the plate was sealed with Parafilm and incubated at 4℃. Plate crystals with maximum dimensions of 80-400 µm grew within 5-7 days.

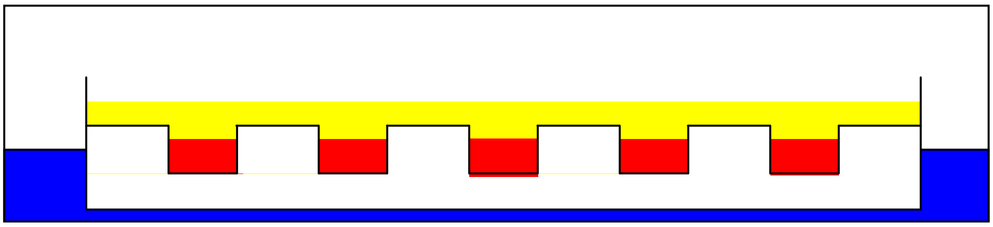


Figure 4.1 – A representation of the method used for crystallisation of $Cu_4(CopZ)_2$. A Terasaki plate was glued into a square Petri dish; 8 ml of silicone oil (yellow) was poured over the wells, 4 µl of a 1:1 protein solution/crystallisation solution (red) was pipetted under the oil, 30 ml of a 30 % (v/v) propan-2-ol solution (blue) was poured into the dish around the plate, the dish was sealed with Parafilm and incubated at 4 ℃.

Two different crystals forms were produced by the different crystallisation techniques, with the vapour batch diffusion technique producing plate crystals and the standard vapour diffusion technique producing more three dimensional crystals.

4.2.1.2 Crystal harvest optimisation and data collection

When harvesting the crystals grown by vapour batch diffusion, several methods were experimented with to decide which produced the best quality crystals most suitable for data collection that would reduce the exposure to air and sufficiently cryoprotect the crystals (all harvesting was undertaken at 4℃). T he best results were obtained from crystals harvested by injecting a cryoprotecting solution of 20 % (v/v) ethylene glycol directly into the drop, under the oil layer, through which the crystals were extracted by mounting in a free standing film using a LithoLoop (Protein Wave Corp, Japan) and immediate cryocooling via rapid immersion into liquid nitrogen. Thus, keeping the crystals under the protective oil layer for as long as possible to reduce exposure to air.

Several methods were also tested for harvesting the hanging drop vapour diffusion grown crystals. The best method found was to remove a small amount of the mother liquor from the drop and replace it with an excess of a cryoprotecting solution (typically 4-16 µl of the mother liquor plus 20-30 % (v/v) ethylene glycol), thus removing the need to transfer crystals to a separate drop containing the cryoprotecting solution, which would have increased their exposure to air, leading to degradation via isopropanol evaporation. Crystals were quickly mounted in a free standing film using a LithoLoop (Protein Wave Corp, Japan) and immediate cryocooling via rapid immersion into liquid nitrogen.

X-ray diffraction datasets were collected at the SRS (Daresbury Laboratory, UK) on station 10.1 using a MAR225 CCD detector. From crystals grown by the vapour batch diffusion technique, a SAD dataset was collected at the copper K-edge (λ = 1.379 Å, 8.99 keV), with a detector distance of 135mm and an exposure time of 8 seconds per image. 360 1° oscillations about the goniometer Φ axis were recorded. As well as these full datasets a Cu fluorescence scan was performed on the crystal before the data collection. A SAD dataset was also collected from a crystal grown by the vapour diffusion technique at the copper K-edge (λ = 1.382 Å, 8.97 keV), with a detector distance of 120 mm and an exposure time of 8 seconds. 180x1° oscillations about the goniometer Φ axis were recorded. A Cu-K fluorescence scan was also performed on this crystal.

4.2.1.3 Structure determination and refinement

Analysis of a SAD dataset collected from a crystal grown using the vapour batch diffusion technique using MOSFLM **[CCP4 1994]** suggested the space group was of a triclinic crystal system. XPREP **[Sheldrick 1991]** was used for preliminary space group determination, suggesting a P1 space group. The data was scaled using SCALA **[CCP4 1994, Kabsch 1988]** and the space group confirmed to be P1. Molecular replacement was carried out with MOLREP **[CCP4 1994]** using the existing CopZ NMR structure as a search model. Initial structure refinement was carried out using REFMAC5 **[CCP4 1994, Murshudov *et al* 1997]**. COOT **[Emsley and Cowtan 2004]** was used for map interpretation and remodelling of the structure. Changes were made to several side chains before further refinement with REFMAC5 and addition of water molecules with ARPwaters **[Perrakis *et al* 1997]**. As was found with previous work on this form of CopZ, a $Cu_4(CopZ)_2$ crystal structure was identified rather than the $Cu_2(CopZ)_2$ structure found in solution. The final structure from this refinement, carried out over the full resolution range (27.6-2.0 Å), at a resolution of 2.00 Å had an R-factor of 19.1%, and an Rfree of 24.4%. For full data collection and refinement parameters for this dataset see tables 4.1 and 4.2.

An analysis of a SAD dataset taken from a crystal grown with the hanging drop vapour diffusion technique was performed using the same techniques. The space group was determined to be $P2_1$, as with the P1 form a $Cu_4(CopZ)_2$ structure was found rather than the expected $Cu_2(CopZ)_2$ structure. The final structure from this refinement, carried out over the full resolution range (40-1.70 Å), had an R-factor of 18.4 %, and an Rfree of 17.7 %. For full data collection and refinement parameters see Tables 4.1 and 4.2.

4.2.2 The crystal structure of $Cu_3(CopZ)_3$

4.2.2.1 Crystallisation

A concentrated protein solution of CopZ (58.2 mg ml^{-1}), determined by luminescence spectroscopy to contain a dimeric form of the protein binding a single Cu(I) ion (data not shown), was provided by Liang Zhou (UEA). Initial crystallisation screening experiments were carried out under anaerobic conditions at 4℃ using a Belle Technology glove box, with an oxygen concentration of 0.2 ppm. Hanging drop vapour diffusion experiments were set utilising the screens of Jancarik & Kim **[Jancarik and Kim 1991]** and Cudney *et al*. **[Cudney *et al* 1994]**. Clusters of rod shaped crystals appeared after 4 days from a crystal growth solution containing; 0.2 M ammonium acetate, 0.1 M sodium acetate pH 4.6 and 30 % (w/v) PEG 4000. Optimisation around this condition revealed that crystals were able to grow at a lower ammonium acetate concentration (0.05 M) and lower PEG 4000 concentration (26 % v/w) but grew best in the original screen conditions.

4.2.2.2 Crystal harvest and X-ray diffraction data collection

Crystal harvesting was carried out anaerobically at 16℃. Crystals were transferred to a cryoprotecting solution (0.2 M ammonium acetate, 0.1 M sodium acetate pH 4.6, 30 % (w/v) PEG 4000 and 25 % (v/v) ethylene glycol) and allowed to equilibrate for one minute. The crystals used for data collection were rod-shaped with dimensions ranging from 75-200 µm x 5-10 µm and were mounted in a free standing film using a cryo-loop (Hampton Research) and cryocooled by immediate immersion into liquid nitrogen.

X-Ray data was collected at the SRS (Daresbury Laboratory, UK) on station 10.1 using a MAR225 CCD detector to a maximum resolution of 1.9 Å. A SAD dataset was collected at the high energy side of the copper K-edge (λ = 1.38 Å), with a detector distance of 115 mm and an exposure time of 10 seconds per image. 340×1° oscillations about the goniometer Φ axis were recorded, producing 242 usable diffraction images.

4.2.2.3 Structure determination and refinement

Analysis of the SAD dataset using MOSFLM **[CCP4 1994]** suggested Laue group 6/m. The data was scaled with SCALA **[CCP4 1994, Kabsch 1988]** and molecular replacement was carried out using MOLREP **[CCP4 1994]** with a monomer from the $Cu_4(CopZ)_2$ crystal structure **[140]** used as the search model. From this analysis the space group was unambiguously determined to be $P6_3$. Initial refinement was carried out using REFMAC5 **[CCP4 1994, Murshudov *et al* 1997]**, and COOT **[Emsley and Cowtan 2004]** was used for model building. Changes were made to several side chains with COOT, before further refinement with REFMAC5. Interestingly, the protein was found to exist in a trimeric $Cu_3(CopZ)_3$ form (3 CopZ monomers and 3 Cu(I) ions), rather than the expected dimeric $Cu_1(CopZ)_2$ form. The final structure from this refinement, using data from the full resolution range (50-1.9 Å), had an R-factor of 27.3 %, and an Rfree of 33.5 %. Due to the size of these R-values, SFCHECK **[CCP4 1994]** was used to look for twinning that may not have been picket up by SCALA. SFCHECK suggested a twinning fraction of 4%, so SHELX **[Sheldrick and Schneider 1997]** was subsequently used to refine the structure as it can incorporate the effects of twinning into its analysis. The results from this refinement were more promising. After the addition of waters using COOT (20 in total) the final R-factors were R = 19.1 % and Rfree = 26.4 %. Figure 4.2 shows one of the residues (Glu9) as an example where the map was improved using SHELXL refinement. For full data collection and refinement parameters see Tables 4.1 and 4.2.

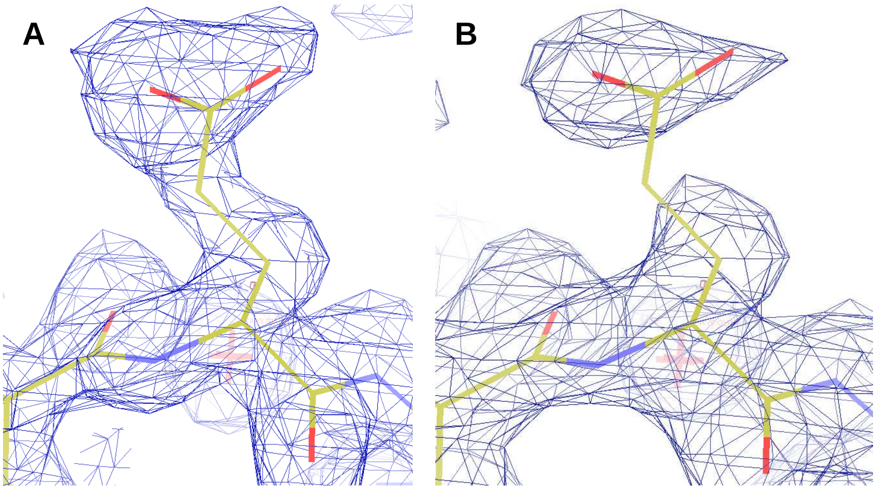


Figure 4.2 – Residue Glu9 from $Cu_3(CopZ)_3$, shown with the double difference Fourier maps generated by **(A)** SHELXL and **(B)** REFMAC5 refinement, both at a contour level of 1.0 sigma. This shows one of the areas of improvement between the two electron density predictions. This figure was created using COOT **[Emsley and Cowtan 2004]**

To ensure the results were not being influenced by the template models used in the analyses, SAD-phased maps were created using the copper anomalous signal. SHELXD & SHELXE **[Sheldrick and Schneider 1997]** were used to determine the positions of the anomalous scatterers and for solvent flattening. This analysis successfully identified the positions of the coppers in the trimer.

4.2.3 $Cu_4(CopZ)_2$ Copper oxidation state refinement

Given the sensitivity of Cu(I) to oxidation and the solvent exposure of the outer sites in the cluster, verifying that the cluster contained solely Cu(I) ions is no simple task. The shoulder apparent in X-ray fluorescence spectra of $Cu_4(CopZ)_2$ crystals at ~8984 eV (Figure 4.6D) arises from the 1s→4p transition. However, this alone is insufficient to unambiguously determine the composition of each copper ion in the cluster. Work done by Einsle *et al* **[Einsle *et al* 2007]** has suggested it is possible to determine the oxidation state of metal ions within a protein structure by taking multiple datasets from a single crystal around the X-ray absorption edge of the expected metal ion and collectively refine the anomalous scattering factors of these datasets. Datasets were therefore collected at five different wavelengths (8987, 8990, 8993, 8985 & 8998 eV), these wavelengths were chosen after performing a fluorescence scan on a CopZ crystal (Figure 4.6D) and picking wavelengths across the copper edge, statistics for these datasets can be seen in Table 4.5. A dataset was also collected at the low energy side of the Cu-K edge (8960 eV) from which a structural model was created and refined using data to a resolution of 1.79 Ǻ, to give an accurate structure for the oxidation state refinement (R values for this model with the copper edge datasets can be seen in Table 4.3). X-ray energy-dependent anomalous scattering factors ($\Delta f''$ and $\Delta f'$) were refined for each metal ion for each dataset using the phenix program suite **[Adams *et al* 2010].**

4.2.4 Molecular modelling and protein structure analysis

Protein structures were superimposed using SUPERPOSE from the CCP4 programme suite **[CCP4 1994, Krissinel and Henrick 2004]** or PyMOL **[DeLano 2002]**. Analysis of subunit interfaces was performed with PROTORP **[Reynolds *et al* 2009]**.

4.2.5 Structure predictions for polynuclear copper cluster proteins

4.2.5.1 Selection of existing poly-nuclear copper cluster proteins with solved structures

A search was performed on the PDB for proteins containing copper (I) ions by selecting the “Chemical ID” option in the advanced search parameters and searching for “CU1”. Relevant proteins were selected from the resultant list by identifying proteins containing at least two Cu(I) ions, where both were in van der Waals contact, and therefore likely to form a poly-nuclear copper cluster.

4.2.5.2 Creation of Hidden Markov Models (HMMs)

Homologous copper cluster packings were identified in the selected protein structures, the sequences between the first and last residues of the copper coordinating sequence were extracted, multiple sequence alignments performed with TCOFFEE **[Notredame *et al* 2000]** and HMMs built using HMMER **[Eddy 1998]**.

4.2.5.3 Building of predictive models

The newly created HMMs were used to search sequences of poly-nuclear copper cluster proteins with unknown structures. When a valid hit (one where all the copper ligating residues from the HMM and the target sequence line up) was identified, the structures of the sequences that made up the HMM were used as templates for model building with MODELLER **[Eswar *et al* 2006]**. As well as the HMM based predictions, models were also built using the tetra-nuclear copper cluster from the $Cu_4(CopZ)_2$ as a template.

4.3 Results

4.3.1 Statistics from data collections and structural refinement

Tables 4.1 and 4.2 summarise the data collection and refinement statistics for the P1 structure of $Cu_4(CopZ)_2$, the $P2_1$ structure of $Cu_4(CopZ)_2$ (collected for use in the oxidation state refinement experiment) and structure of $Cu_3(CopZ)_3$.

Table 4.1 - Data collection statistics for each CopZ dataset

Dataset	$Cu_4(CopZ)_2$ (SAD data collection)	$Cu_4(CopZ)_2$ (SAD data collection)	$Cu_3(CopZ)_3$ (SAD data collection)
Beamline	SRS 10.1	ESRF BM14	SRS 10.1
Space group	P1	$P2_1$	$P6_3$
Cell Parameters a , b , c (Å) α , β , γ (°)	 31.53 , 43.30 , 54.30 78.65 , 86.29 , 84.52	 23.39 , 74.83 , 40.88 90 , 101.59 , 90	 63.96 , 63.96 , 27.30 90 , 90 , 120
Wavelength (Å)	1.379	1.38	1.38
Resolution (Å)	30 – 2.0(2.11 – 2.0)	27.34 – 1.79 (1.89-1.79)	50 – 1.9 (2.0 – 1.9)
Completeness (%)	94.4 (86.2)	97.6 (95.3)	97.2 (84.2)
R_{sym} (%)	2.2 (3.8)	4.8 (19.6)	6.6 (23.8)
R_{anom} (%)	4.6 (19.4)	3.8 (14.3)	7.1 (26.3)
<I/σI>	39.6 (24.3)	18.9 (6.1)	29.1 (8.6)
Independent reflections	17865 (2402)	12659 (1762)	5023 (608)
Multiplicity	4.0 (3.9)	3.6 (3.4)	14.1 (10.7)
Overall temperature factor (Å^2)	20.1	11.8	19.2
Anomalous completeness (%)	93.9 (85.0)	93.4 (87.1)	96.7 (86.3)
Anomalous multiplicity	2.0 (2.0)	1.8 (1.7)	7.3 (5.2)

Numbers in brackets represent data in the high resolution shell

Table 4.2 - Refinement statistics for each CopZ dataset

Dataset	$Cu_4(CopZ)_2$ (P1)	$Cu_4(CopZ)_2$ ($P2_1$)	$Cu_3(CopZ)_3$
CopZ monomers per AU*	4	2	1
Cu ions per AU*	8	4	1
Refined structure Total atoms Water molecules	 2225 199	 1159 137	 527 20
R_{cryst} (%)	19.1	17.6	19.1
R_{free} (%)	24.2	22.3	26.4
Ramachandran Analysis (%) Most favoured Additional allowed Generously allowed	 95.9 4.1 0	 96.7 3.3 0	 85.5 14.5 0
RMS deviations Bonds (Å) Angles (°) Planes (Å)	 0.02 1.58 0.11	 0.01 1.17 0.08	 0.01 2.00 0.02
Mean Atomic B-value (Å^2)	21.6	12.0	17.6

*AU = Asymmetric unit

4.3.2 Structure of CopZ proteins – $Cu_4(CopZ)_2$ (P1 crystal structure)

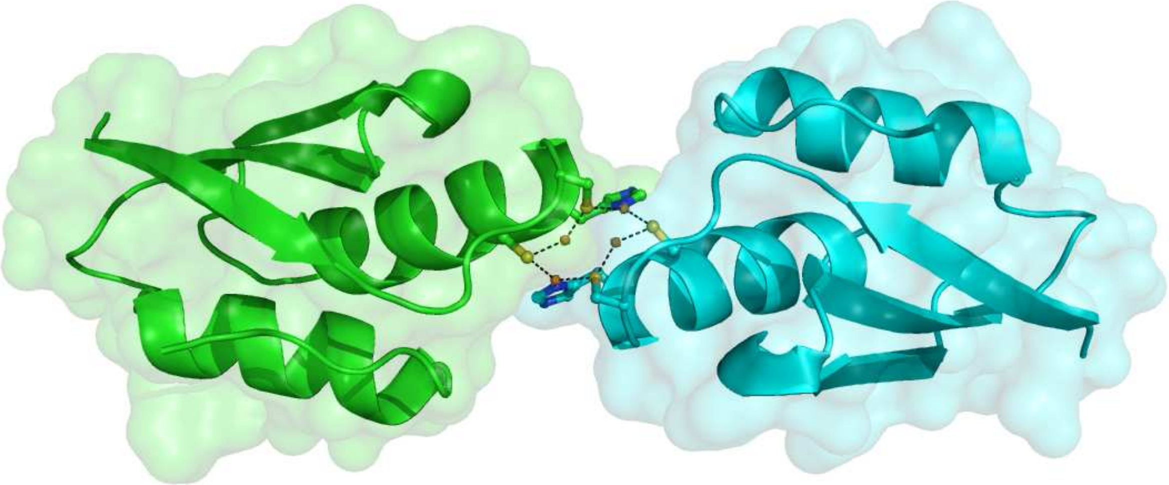


Figure 4.3 – The structure of the $Cu_4(CopZ)_2$ dimer, as seen from above the copper binding site. The side chains of the copper binding residues are displayed in stick format and copper ions as orange spheres.

The final model of the two $Cu_4(CopZ)_2$ dimers found in the asymmetric unit (Figure 4.3) contains all the 276 amino acid residues from the primary sequence (69 per monomer), eight copper ions and 199 water molecules. The principal secondary structure elements for each monomer as determined by the program STRIDE **[Frishman and Argos 1995]** are: two α helices (Gln14–Glu26 and Val53–Gln63), a 3_{10} helix (Leu37-Ala39) and three β strands (Glu2–Glu9, Val30–Val36 and Lys41–Phe46). The (Φ, Ψ) torsion angles of all residues fall within the allowed regions of the Ramachandran plot.

This CopZ structure was found to have a novel tetranuclear Cu(I) cluster with two subsets of Cu(I) ions in different coordination environments (Figure 4.4). The outer Cu(I) ions (labelled 3 and 4) exhibit distorted trigonal coordination, while the inner Cu(I) ions (labelled 1 and 2) exhibit distorted digional coordination. Four cysteine residues (Cys13 and Cys16 from each monomer) are central to the formation of the cluster, whereby each acts as a ligand to an inner (digonal) and outer (trigonal) copper ion. Two histidine residues (His15 from each monomer) provide the remaining ligands to the trigonal Cu(I) ion sites. In addition, two water molecules (W1 and W2) move to points within 2.53 and 2.61Ǻ of the trigonal copper ions sites 3 and 4, respectively, imparting a partial tetrahedral character. The Cu(I) ions in adjacent trigonal and diagonal sites lie at a distance of 2.57 Ǻ , while the distance between the digonal sites is 2.74 Ǻ. These distances, particularly the former, are shorter than the sum of the van der Waals radii of the ions, suggesting the presence of a true metal cluster.

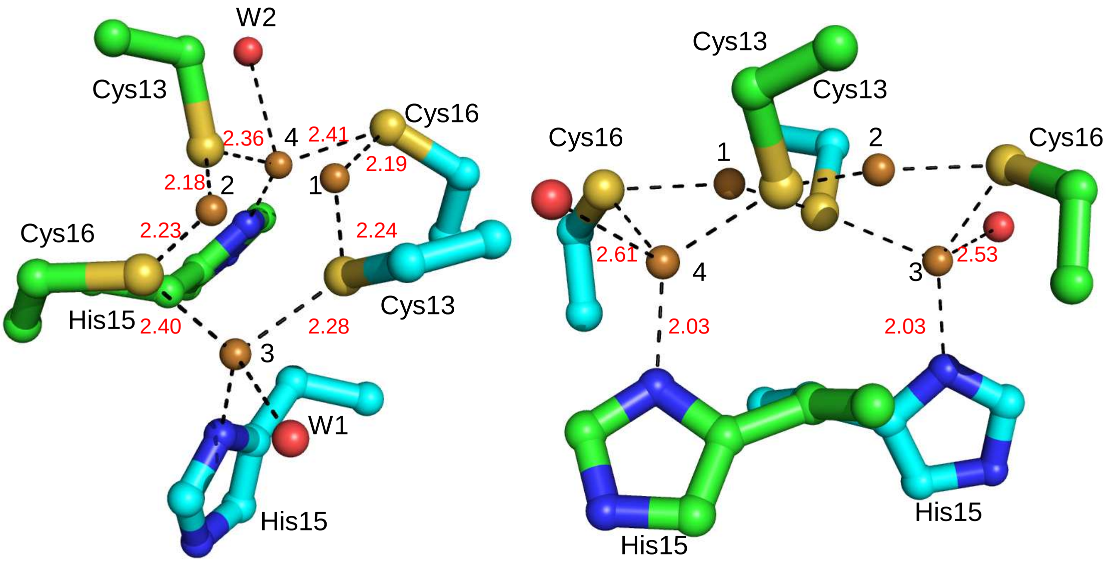


Figure 4.4 - The $Cu_4(CopZ)_2$ copper binding motif. Bonds between the copper coordinating residues Cys13, His15 & Cys16 and waters (W1 & W2) are marked with dashed lines. Bond lengths are an averaged value over the two dimers in the asymmetric unit (+/- 0.1 Ǻ).

The Ser12 residues from each subunit form part of a second coordination sphere of the inner Cu(I) sites. The serine hydroxyl oxygen-Cu(I) distances are 2.92 and 3.04 Ǻ for sites 1 and 2, respectively. In the same way, Tyr65 and Tyr650 form part of a second coordination sphere to the outer Cu(I) sites with phenolic hydroxyl-Cu(I) distances of 3.39 and 3.46 Ǻ for sites 3 and 4, respectively. The side chains of the methionine residue of each MXCXXC motif (Met11) point away from the cluster and insert into the core of the protein, making van der Waals contact with other hydrophobic residues, including the side chain of Tyr65. The methionine residue appears to contribute to local protein structural integrity and will therefore play an indirect role in copper binding.

The inner Cu(I) ions of the tetranuclear cluster are buried at the CopZ dimer interface and shielded from interaction with solvent. The sulfur atoms of the four cysteine residues acting as ligands to the inner and outer sites are also buried. Luminescence in the ~600 nm region is often observed for protein-bound copper clusters and is indicative of the cluster being in a solvent shielded environment **[Stillman 1995, Srinivasan *et al* 1998]**. The observed solvent exposure of the cluster here is consistent with the lack of a luminescence signal associated with this complex.

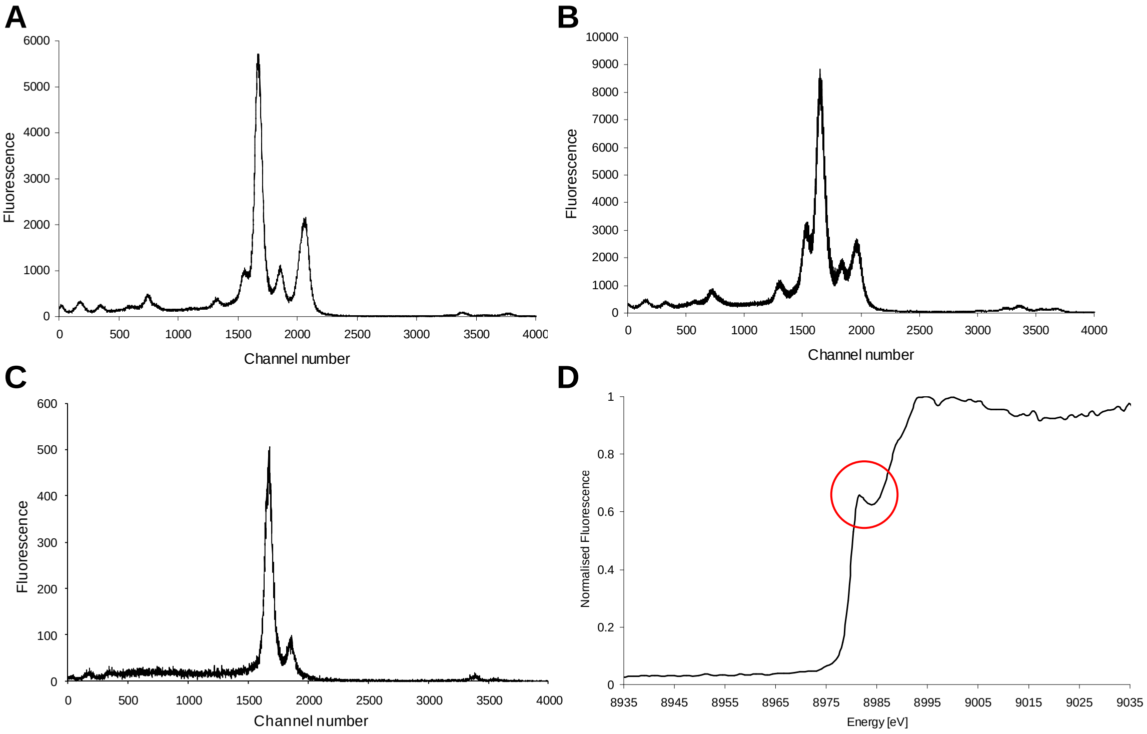


Figure 4.5 – Energy dispersive X-ray fluorescence spectra of $Cu_4(CopZ)_2$ taken at **(A)** 1.25 Å (below the zinc edge) and **(B)** 1.295 Å (above the zinc edge). Also shown are **(C)** a fluorescence scan of copper foil and **(D)** an EXAFS fluorescence scan that contains a typical Cu(I) feature (circled).

To prove the coppers were indeed coppers and not zinc, energy dispersive X-ray fluorescence (EXF) spectra were taken from the crystal and from copper foil (Figure 4.5). The fluorescence scans (Figures 4.6 A&B) taken either side of the zinc K-edge suggest zinc is not present in the crystal as the number and relative size of the peaks does not change. If zinc was present a large peak at a higher channel number than the copper peak would be expected in the EXF spectra taken on the high energy side of the zinc absorption edge (Figure 4.6B). The fluorescence scan on copper foil (Figure 4.6C) proves the peak in the scans on the crystal corresponds to copper, since they appear in the same place as the copper foil peak. The EXAFS fluorescence trace (Figure 4.6D) suggests that at least some of the coppers are in the Cu(I) form, as it contains the classic Cu(I) feature at 8983 eV corresponding to the 1s→4s transition of Cu(I) **[Hu *et al* 1997]**, highlighted by a circle in Figure 4.5D.

4.3.3 Copper cluster oxidation state refinement

Table 4.3 summarises the data collection statistics for the different wavelength datasets collected for oxidation state refinement. Table 4.4 shows the R values for these datasets calculated using the model created from the low energy dataset (8960 eV).

Table 4.3 – Statistics for the datasets collected from a single CopZ crystal

Dataset	**Rmerge (%)**			**Completeness (%)**		
	Overall	**Inner**	**Outer**	**Overall**	**Inner**	**Outer**
8960	48.0	24.0	19.6	97.6	99.2	95.3
8987	47.0	26.0	18.3	97.6	99.2	95.3
8990	48.0	25.0	18.2	97.4	99.2	94.4
8993	50.0	25.0	19.1	97.4	99.2	94.8
8995	49.0	27.0	19.3	97.5	99.2	94.9
8998	49.0	24.0	19.4	97.6	99.2	95.1
	Anomalous completeness (%)			***Anomalous multiplicity***		
	Overall	**Inner**	**Outer**	**Overall**	**Inner**	**Outer**
8960	99.3	99.5	87.1	1.8	1.9	1.7
8987	93.4	99.5	87.2	1.8	1.9	1.7
8990	92.6	98.9	85.4	1.7	1.9	1.6
8993	92.5	98.9	86.2	1.7	1.9	1.7
8995	93.3	98.2	87.1	1.8	1.9	1.7
8998	93.9	98.4	87.8	1.8	1.9	1.7

Table 4.4 – R and FreeR values for copper edge datasets against the model generated with the 8960 eV dataset

Dataset	**8987**	**8990**	**8993**	**8995**	**8998**
R (%)	17.84	17.77	17.71	17.71	17.62
FreeR (%)	21.1	21.1	21.1	21.1	21.1

Refinement of the X-ray energy-dependent anomalous scattering factors ($\Delta f''$ and $\Delta f'$) for each metal ion gave results consistent with scattering from copper ions alone, and furthermore, suggests that all four copper sites are in the Cu(I) oxidation state, as no pre-edge features arising from tetrahedral Cu(II) are detected in the region 8988-8990 eV (Figure 4.6).

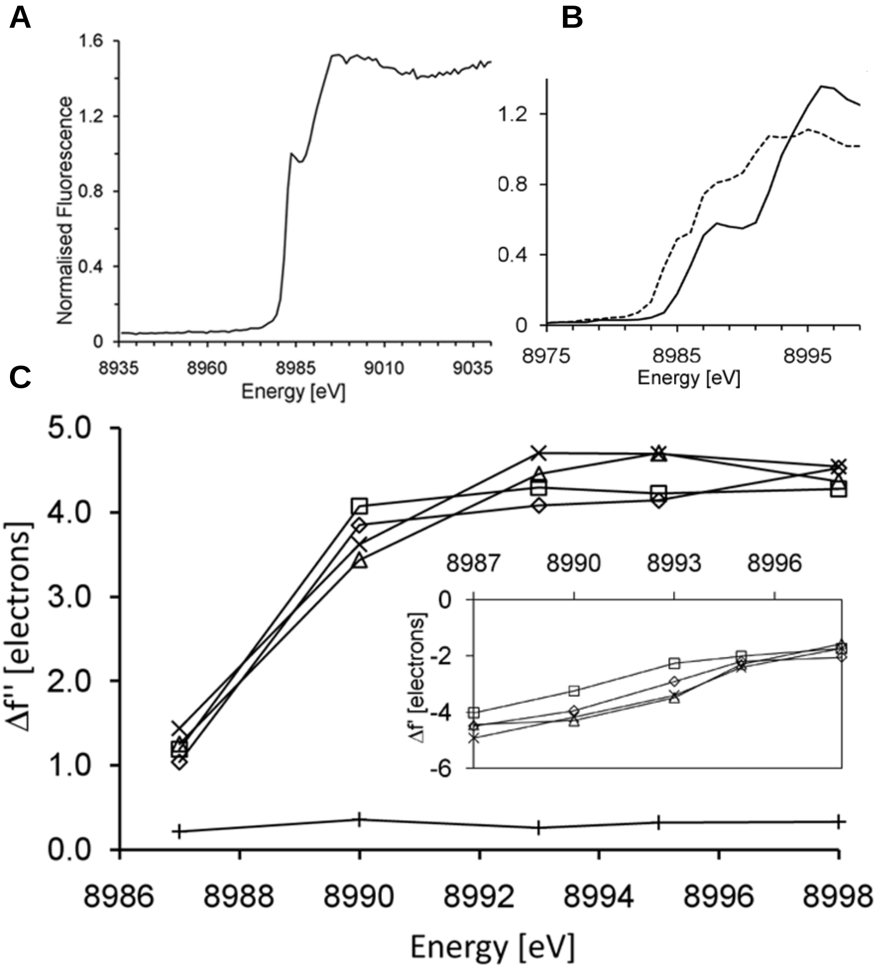


Figure 4.6 - Analysis of the $Cu_4(CopZ)_2$ metal cluster. **(A)** X-ray fluorescence emission spectrum of a single $Cu_4(CopZ)_2$ crystal normalized to the value at 8984 eV. **(B)** Normalized edge spectra for model Cu(II) complexes, with S4 (dashed line) and N4 (solid line) equatorial ligand sets **[Kau *et al* 1987]**. **(C)** Refined anomalous scattering factors Δf ″ and Δf ′ (inset) for digonal, Cu1 (□) and Cu2 (◊), and trigonal, Cu3 (Δ) and Cu4 (x), copper sites. Δf ″ scattering factors for sulfur are also given (+)

4.3.4 Comparison of $Cu_4(CopZ)_2$ P1 and $P2_1$ structures

The final $Cu_4(CopZ)_2$ structures derived from the P1 and $P2_1$ crystal forms were aligned, the all atom RMSD value from this structural alignment was 0.39 Ǻ. A closer examination of the two structures showed a small degree of flex between the dimers, with angles of 124° and 127° between the two dimers of t he $P2_1$ and P1 forms respectively. Figure 4.7 shows the alignment of the structures of the two CopZ dimers, demonstrating the close structural homology between the two models.

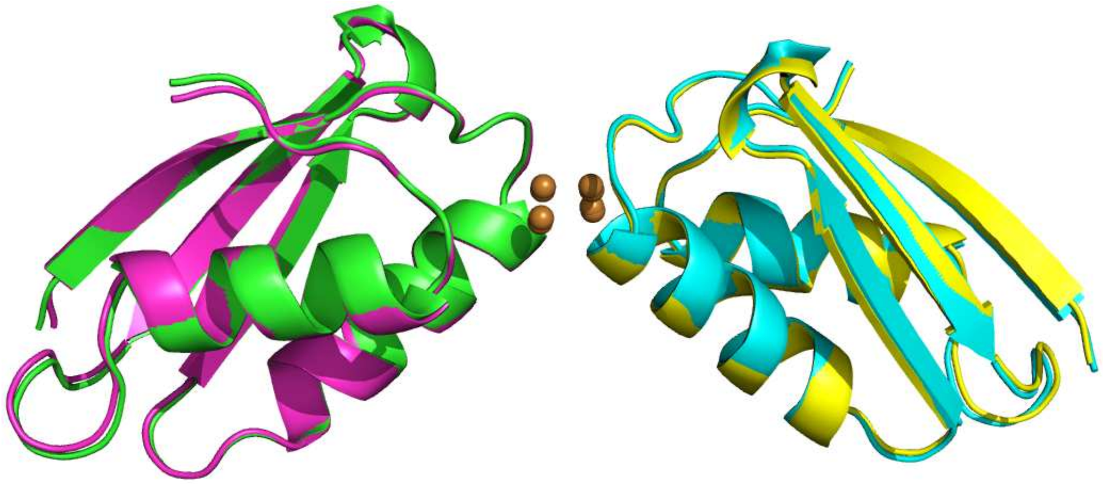


Figure 4.7 – A superposition of the $Cu_4(CopZ)_2$ $P2_1$ crystal form (Green & Cyan) and P1 crystal from (Magenta & Yellow) in a cartoon representation. Coppers are shown as orange spheres.

4.3.5 Structure of $Cu_3(CopZ)_3$

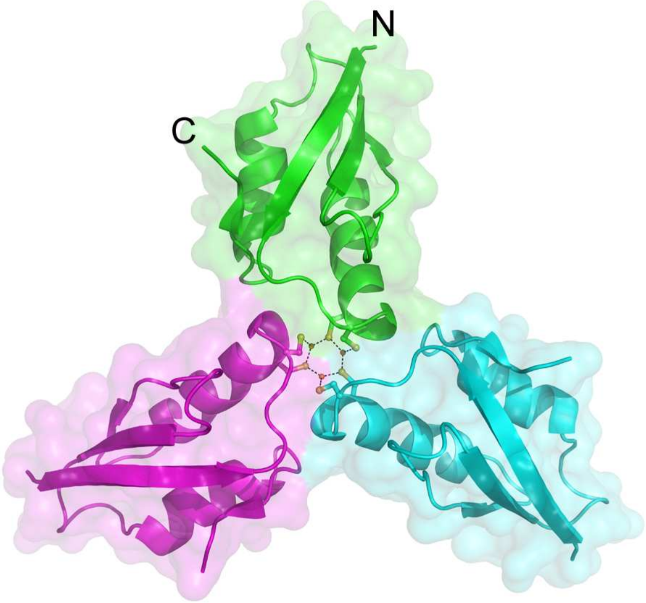


Figure 4.8 - The structure of $Cu_3(CopZ)_3$ as viewed from down the 3-fold axes. Each monomer is displayed in cartoon format, with the side chains of the copper-coordinating residues displayed as sticks. The three CopZ monomers (A, B and C) are shown along with their molecular surfaces coloured green, cyan and magenta. Copper ions are coloured orange.

The final model of the CopZ monomer contains all 69 of the amino acid residues from the primary sequence, one presumed Cu(I) ion and 20 water molecules. The principle secondary structure elements for each monomer predicted by STRIDE **[Frishman and Argos 1995]** are: two α-helices (Gln14–Glu26 and Val53–Asp62), three β-strands (Glu2–Glu9, Val30–Val33 and Lys41–Phe46), four type I β-turns (Asn36-Ala39, Leu37-Gly40, Asp47-Lys50 and Ala48-Val51), one type IV β-turn (Leu27-Val30) and one inverted γ-turn (Val67-Ala69).

The result expected from this analysis was a $Cu_1(CopZ)_2$ dimeric protein, containing two CopZ monomers binding one Cu(I) ion. However, the results of cell and symmetry tests on the monomer obtained from crystallographic experiments and the positions of the anomalous scatterers have shown the protein to be in a $Cu_3(CopZ)_3$ trimeric form, with three CopZ monomers binding three Cu(I) ions (Figure 4.8). The trigonal co-ordination of each copper ion in the cluster is provided by Cys13 and Cys16 from each CopZ monomer (distances 2.24 Å and 2.31 Å respectively) and Cys16 of a neighbouring monomer (coordination distance 2.21 Å) (Figure 4.9A). These distances are consistent with those observed in similar copper sites in proteins deposited in the PDB (~2.4 Å). The three sulfur atoms and copper ion are essentially coplanar and the most compressed of the S–Cu–S angles (107.1°) involves the sulfur atoms from Cys16 residues. The copper ions in the cluster are fully shielded from the solvent, but the sulfur atoms of residues Cys16 are solvent accessible on the near face of the trimer. Solvent access to the copper cluster via the remote face of the trimer is blocked by residues Ser12 arranged around the molecular 3-fold axis.

Two water molecules (W1 and W2) are buried in each CopZ monomer adjacent to the bound copper ion. Their low temperature factors indicate a restricted mobility. Tyr65 plays a central role in stabilizing these solvent sites, forming hydrogen bonds with each. W1 also forms further hydrogen bonds with the main chain amide nitrogen atoms of residues Ser12 and Cys13, whereas W2 forms hydrogen bonds with the main chain carbonyl oxygen of Cys16 and the side-chain of Gln63 (Figure 4.8B). These interactions appear critical to the conformation of the polypeptide spanning the copper-binding sequence motif.

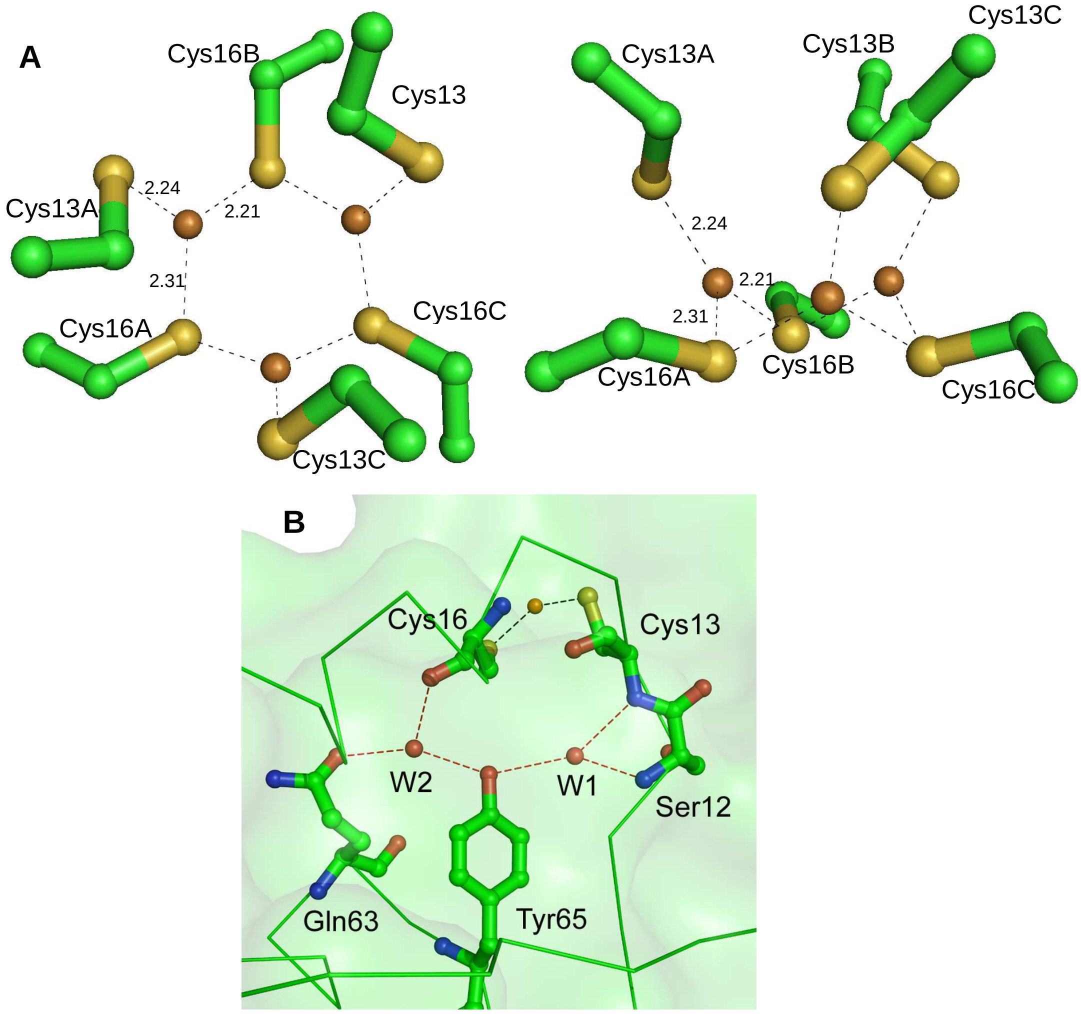


Figure 4.9 – (A) The Cu$_3$(CopZ)$_3$ copper binding motif. Bonds between copper coordinating residues Cys13 & Cys16 are marked with dashed lines. Bond lengths are in Å. **(B)** The intra-subunit interactions involving water molecules W1 and W2 and residues Ser12, Gln63 and Tyr65.

The (Φ, Ψ) torsion angles of all residues fall within the allowed regions of the Ramachandran plot. All the residues around the copper binding site are well defined within the electron density (Figure 4.10A & 4.10B), although some residues at the N-&C-termini are less well defined. To help validate the positions of the copper ions and thus the trimeric nature of the protein, an anomalous difference map was created that was found to have three intense regions of electron density around the three fold axis that were still present up to a contour level of 30 sigma. An anomalous difference fourier map was also calculated (Figure 4.10C), since these peaks were located at the copper binding site where coppers were expected and the dataset was taken at the copper k-edge, where the majority of the anomalous signal would be expected to come from copper ions, it is very likely these peaks of electron density refer to the three proposed Cu(I) ions of the Cu$_3$(CopZ)$_3$ structure.

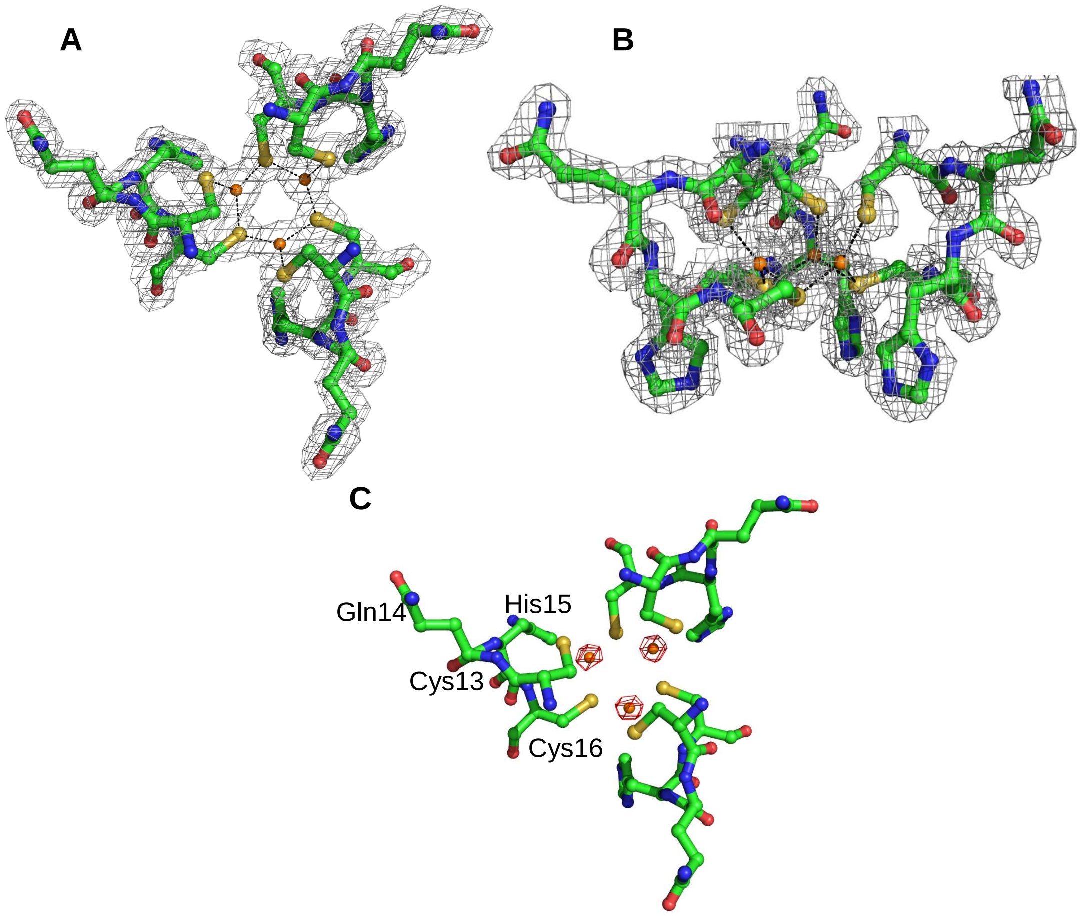


Figure 4.10 – The copper binding centre of the $Cu_3(CopZ)_3$ trimer (Cys13-Gln14-His15-Cys16), showing; **(A)** the SHELXL Fourier map (grey) orientated down the three fold axes and **(B)** orientated to the side of the trimer interface, at a contour level of 1.0 sigma, and **(C)** an anomalous difference Fourier map (red) showing the positions of the coppers at a contour level of 20 sigma.

4.3.6 Comparison of the $Cu_4(CopZ)_2$ and $Cu_3(CopZ)_3$ structures

Monomers from the $Cu_4(CopZ)_2$ and $Cu_3(CopZ)_3$ structures were superimposed, the all atom RMSD value from this structural alignment was 0.73 Å, suggesting the two monomers are structurally homologous. Figure 4.11A shows how the bulk of the two structures fit together, with the more noticeable differences around the copper binding residues shown in Figure 4.11B.

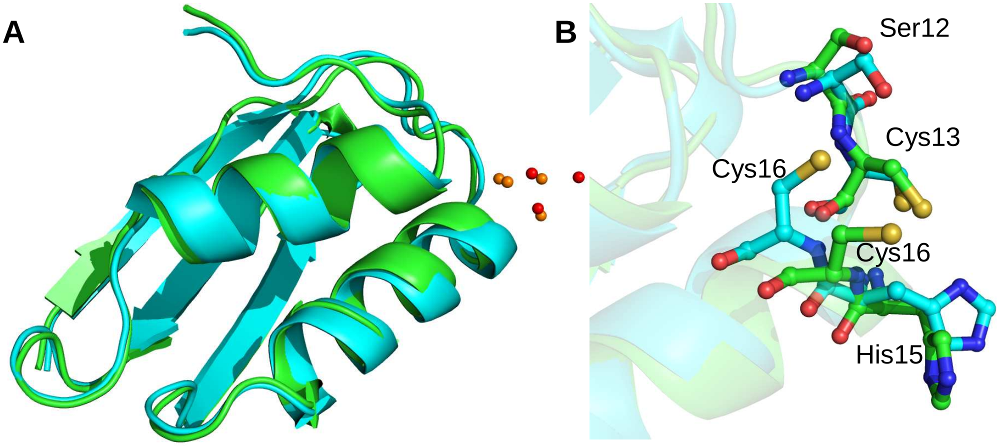


Figure 4.11 – (A) A cartoon representation of monomers from $Cu_4(CopZ)_2$ (Cyan) and $Cu_3(CopZ)_3$ (Green) CopZ structures. Also shown are the positions of the coppers in $Cu_4(CopZ)_2$ (Orange) and $Cu_3(CopZ)_3$ (Red) CopZ structures. **(B)** The copper binding residues from $Cu_4(CopZ)_2$ (Cyan) and $Cu_3(CopZ)_3$ (Green) (Ser12, Cys13, His15 & Cys16). N.B. Although Ser12 and His15 are not directly involved in copper binding in the $Cu_3(CopZ)_3$ structure, they are displayed to highlight the changes in the structures.

These differences in the copper binding residues are due to the different methods each CopZ structure uses for binding copper, with the $Cu_4(CopZ)_2$ structure needing the Ser12 and His15 residues in close proximity the metal binding site as they are involved in copper coordination. There are also differences in the positioning of the cysteine residues (particularly Cys16), due to the differing geometries of the copper clusters.

There are also noticeable differences in the positioning of some of the other residues around the copper binding site that are involved in inter-subunit interactions (Figure 4.12A). These are changes to Tyr65, which is involved in inter-subunit interactions in the $Cu_4(CopZ)_2$ structure as part of the secondary coordination sphere of the copper cluster (Figure 4.12B) and is involved with stabilising the copper coordinating cysteines in conjunction with two water molecules (Figure 4.9B), and Lys18, Asp62, Asp66, Gln14 & Gln63, that are involved in inter-subunit hydrogen-bonding interactions in the $Cu_3(CopZ)_3$ structure (Figure 4.12C).

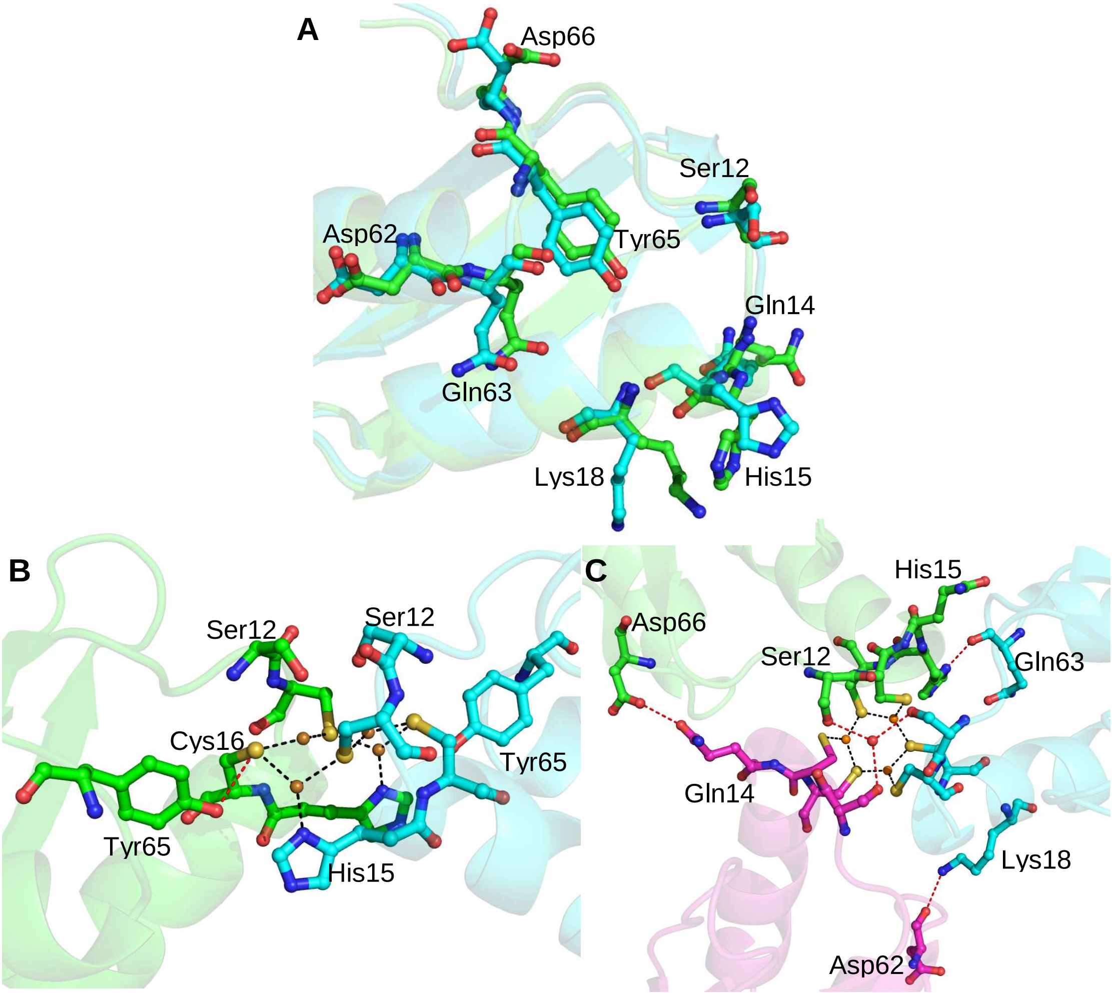
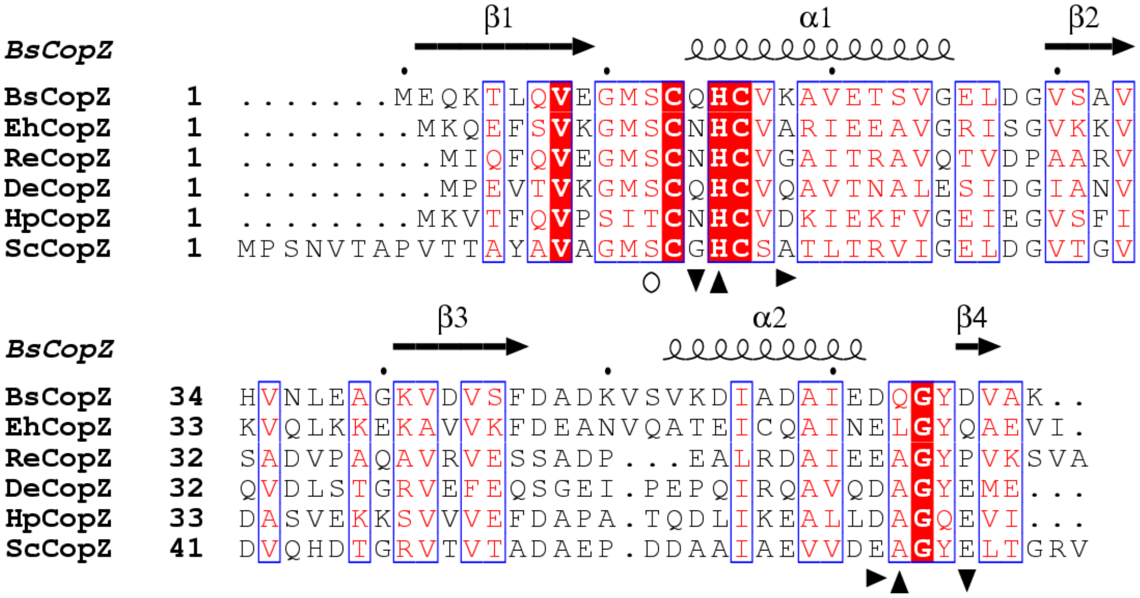


Figure 4.12 – (A) The positional changes of the inter-subunit interacting residues (Ser12, Gln14, His15, Lys18, Asp62, Gln63, Tyr65 & Asp66) of $Cu_4(CopZ)_2$ (Cyan) and $Cu_3(CopZ)_3$ (Green). **(B)** The interactions between subunits of $Cu_4(CopZ)_2$, Cu-residue interactions are represented by black dashed lines and residue-residue interactions (Tyr65–His15) by red dashed lines. **(C)** The interactions between subunits of $Cu_3(CopZ)_3$, Cu-residue interactions are represented by black dashed lines and residue-residue interactions (Ser12-H_2O, Lys18–Asp62, Asp66-Gln14 & His15–Gln63) by red dashed lines.

These results could help to explain why unexpected forms of CopZ were found in the crystal structures. It is possible that in copper-limited conditions (such as the 2 CopZ : 1 Cu(I) conditions that the $Cu_3(CopZ)_3$ crystals grew from) CopZ forms a trimeric structure that allows additional subunit-subunit interactions to hold the whole protein together. Whereas, in conditions where copper is more plentiful (such as the 2 CopZ : 2 Cu(I) conditions the $Cu_4(CopZ)_2$ crystals grew from) additional copper is utilised to form the $Cu_4(CopZ)_2$ dimer, where the binding of the copper is enough to hold the protein together and subunit-subunit interactions are minimal. There is currently no evidence for a wild type $Cu_3(CopZ)_3$ structure in solution, however, a Tyrosine – Lysine mutant has been developed by Nick Le Brun and Chloe Singleton (UEA, School of Chemical Sciences) that has been found to exist as a trimeric species. Recent work by Badarau *et al* has also identified another copper chaperone mutant that exists as a trimer, when they solved the structure of a histidine – tyrosine mutant of Atx1 from the cyanobacterium *Synechocystis* **[Badarau *et al* 2010]**.

With regards to the $Cu_3(CopZ)_3$ structure, sequence alignments between *Bacillus subtilis* CopZ and sequences from other CopZ orthologues from a variety of micro-organisms reveals conservative amino acid substitutions in the residue that mediate the hydrogen bonding interactions between the subunits that stabilise the trimer (Figure 4.13). Suggesting this trimeric form of CopZ may not be limited to *Bacillus subtilis.*



```
BsCopZ
BsCopZ   1  .......MEQKTLQVEGMSCQHCVKAVETSVGELDGVSAV
EhCopZ   1  ........MKQEFSVKGMSCNHCVARIEEAVGRISGVKKV
ReCopZ   1  .........MIQFQVEGMSCNHCVGAITRAVQTVDPAARV
DeCopZ   1  .........MPEVTVKGMSCQHCVQAVTNALESIDGIANV
HpCopZ   1  ........MKVTFQVPSITCNHCVDKIEKFVGEIEGVSFI
ScCopZ   1  MPSNVTAPVTTAYAVAGMSCGHCSATLTRVIGELDGVTGV
                               ○ ▼▲   ►

BsCopZ
BsCopZ  34  HVNLEAGKVDVSFDADKVSVKDIADAIEDQGYDVAK..
EhCopZ  33  KVQLKKEKAVVKFDEANVQATEICQAINELGYQAEVI.
ReCopZ  32  SADVPAQAVRVESSADP...EALRDAIEEAGYPVKSVA
DeCopZ  32  QVDLSTGRVEFEQSGEI.PEPQIRQAVQDAGYEME...
HpCopZ  33  DASVEKKSVVVEFDAPA.TQDLIKEALLDAGQEVI...
ScCopZ  41  DVQHDTGRVTVTADAEP.DDAAIAEVVDEAGYELTGRV
                                         ►▲   ▼
```

Figure 4.13 – Sequence alinment of CopZ orthologues from: CopZ from *Bacillus subtilis* (BsCopZ, O32221); CopZ from *Enterococcus hira*e (EhCopZ, Q47840); copper chaperone from *Ralstonia eutropha* (ReCopZ, Q0K5J5); copper chaperone from *Desulfohalobium retbaense* (DrCopZ, C1SUC1); copper-ion-binding protein from *Helicobacter pylori* G27 (HpCopZ); metal-associated protein from *Streptomyces coelicolor* A3(2) (ScCopZ, B5ZAE1). Identical residues are indicated by a red background, conservatively varied residues are boxed in blue and shown in red characters. Secondary structural elements in CopZ are indicated and labelled. Pairs of residues forming intersubunit contacts in the $Cu_3(CopZ)_3$ trimer are indicated by matching pairs of symbols (▼,▲,►). In each case, the hydrogen bond donor involves the residue in the range 14–18 (CopZ numbering). The corresponding hydrogen bond acceptor is in the residue range 62–66. Note that residue Ser12 forms water-mediated contacts with the equivalent residue in the other two subunits (○).

4.3.7 Comparison of CopZ with CopA

Sequence alignments of CopZ with the N1 (Figure 4.14A) and N2 (Figure 4.14B) domains of CopA were performed that show a conservation of the copper binding MXCXXC motif in all structures.

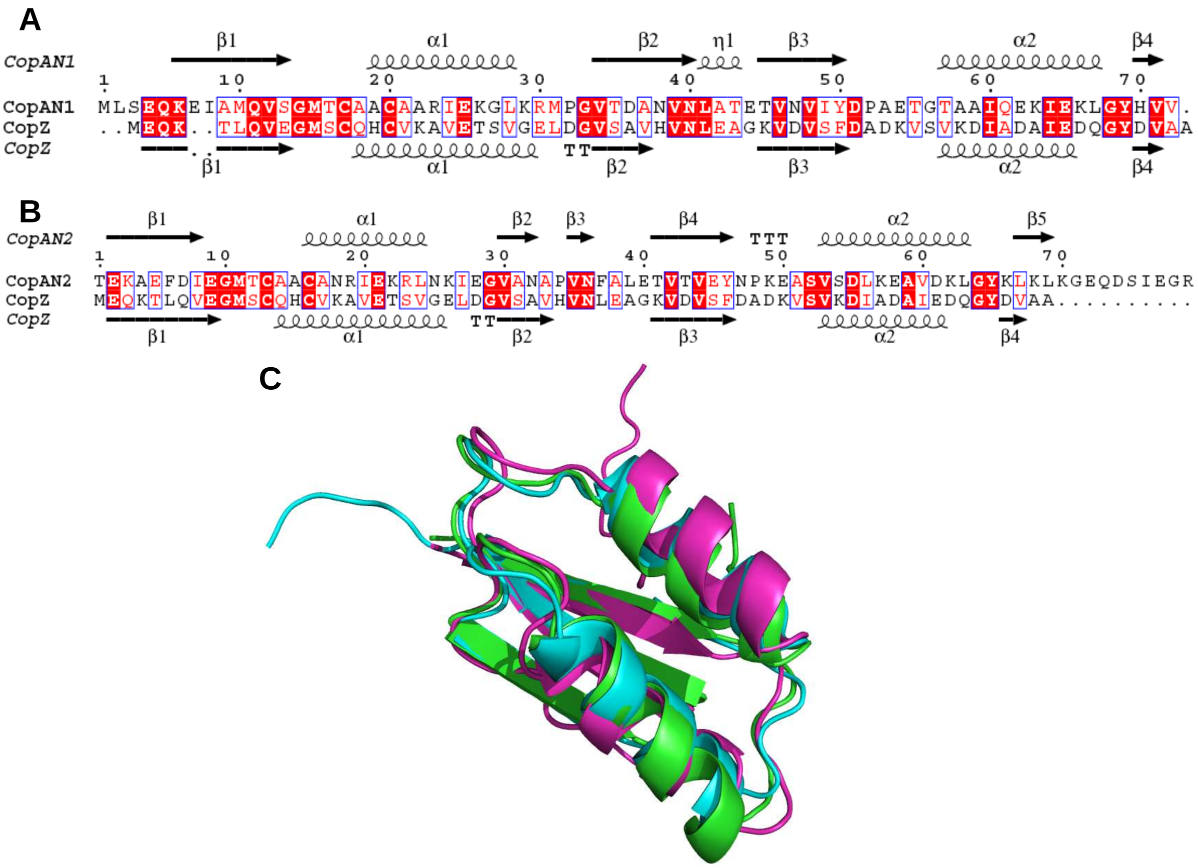

A
CopAN1
β1
α1
β2
η1
β3
α2
β4
1
10
20
30
40
50
60
70
CopAN1 MLSEQKEIAMQVSGMTCAACAARIEKGLKRMPGVTDANVNLATETVNVIYDPAETGTAAIQEKIEKLGYHVV.
CopZ ..MEQK..TLQVEGMSCQHCVKAVETSVGELDGVSAVHVNLEAGKVDVSFDADKVSVKDIADAIEDQGYDVAA
CopZ
TT
B
CopAN2
β1
α1
β2
β3
β4
TTT
α2
β5
1
10
20
30
40
50
60
70
CopAN2 TEKAEFDIEGMTCAACANRIEKRLNKIEGVANAPVNFALETVTVEYNPKEASVSDLKEAVDKLGYKLKLKGEQDSIEGR
CopZ MEQKTLQVEGMSCQHCVKAVETSVGELDGVSAVHVNLEAGKVDVSFDADKVSVKDIADAIEDQGYDVAA..........
CopZ
TT
β1
α1
β2
β3
α2
β4
C

Figure 4.14 – Sequence alignments of **(A)** CopA$_{N1}$ & CopZ and **(B)** CopA$_{N2}$ & CopZ. Sequence similarities are displayed as a blue box around red text and sequence identities as white text on a red background. Secondary structures for each sequence are also shown. **(C)** The superimposed strutcures of a CopZ monomer (Green), CopA-N1 (Cyan) and CopA-N2 (Magenta) shown in cartoon format.

A structural alignment of one monomer from the P2$_1$ CopZ structure with the N1 and N2 domains of the CopA NMR structure **[Banci *et al* 2003(1)]** shows homologous structural alignments (Figure 4.14C), with RMSDs of 0.74 Å for the N1 domain and 1.48 Å for the N2 domain, despite relatively low sequence identities of 32.9 % and 24.1 % respectively. The cysteine residues around the copper binding sites in CopZ and the CopA domains are conserved, suggesting these are the locations for copper binding in CopA.

4.3.8 How the CopZ structure compare with existing structures for CopZ and other metallochaperones?

There are a number of existing NMR structures for CopZ in the PDB, two originating from *Bacillus subtilis* (1K0V, copper bound form **[Banci *et al* 2001]** & 1P8G, apo form **[Banci *et al* 2003(2)]**) and one from *Enterococcus hirae* (1CPZ **[Wimmer *et al* 1999]**). Interestingly, the CopZ structure from *E.hirae* despite having the lowest sequence identity provided the best RMSD fit with the P2_1 $Cu_4(CopZ)_2$ crystal structure (Table 4.7).

Table 4.7 – Results of sequence and structural alignment between a monomer from the P2_1 $Cu_4(CopZ)_2$ structure and the existing CopZ structures.

Structure (PDB ID)	Sequence identities (%)	RMSD value for alignment with $Cu_3(CopZ)_3$ (Å)
1K0V	100	2.85
1P8G	100	2.86
1CPZ	42	1.74

NB – RMSD result refers to the all atom value

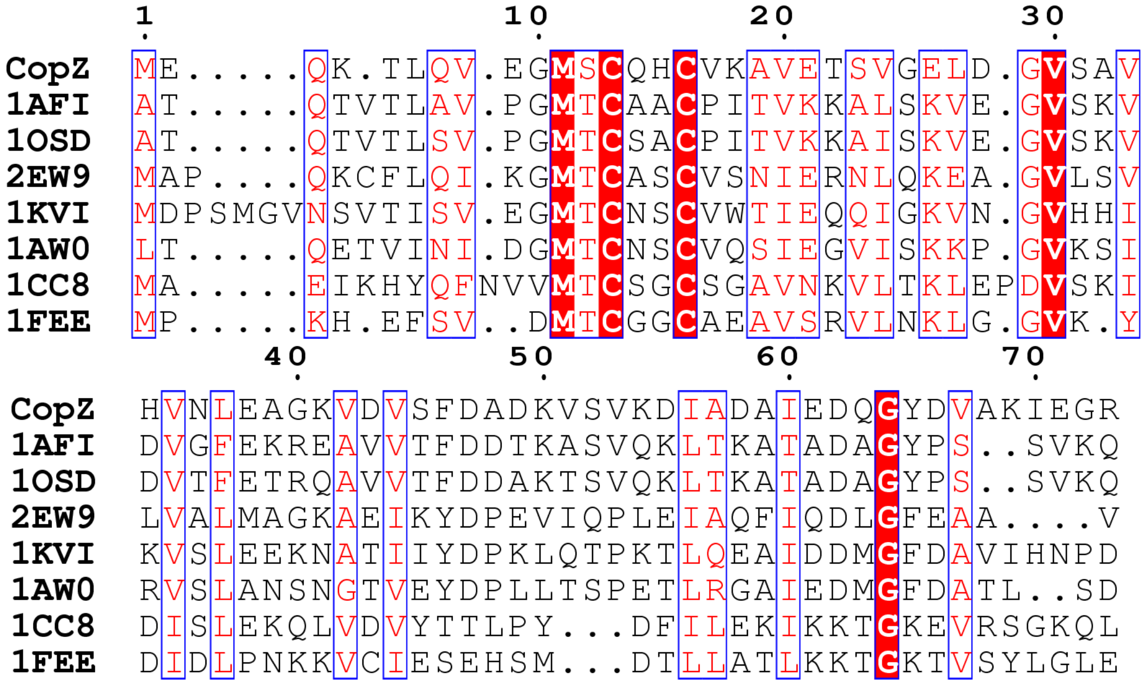
There are several structures for other metallochaperones in the PDB taken from; *Homo sapiens* **[Wernimont *et al* 2000, Gitschier *et al* 1998, DeSilva *et al* 2005, Achila *et al* 2006]**, *Ralstonia metallidurans* **[Serre *et al* 2004]**, *Saccharomyces cerevisiae* **[Rosenzweig *et al* 1999]** and *Shigella flexneri* **[Steele and Opella 1997]**. Sequence and structural alignments were performed against the P2_1 structure for CopZ on these proteins, using NEEDLE **[Needleman and Wunsch 1970]** for the sequence alignments and SUPERPOSE **[CCP4 1994, Krissinel and Henrick 2004]** for the structural alignments. A breakdown of these results can be seen in table 4.8.

Table 4.8 – Results of sequence and structural alignment between the P2_1 $Cu_4(CopZ)_2$ structure and other metallochaperones.

Structure (PDB ID)	Solved by	Description	Organism	Sequence identity to CopZ (%)	RMSD value for alignment with CopZ (Å)
1FEE	X-ray	HAH1	*Homo sapiens*	23	3.94
1CC8	X-ray	Atx1 metallochaperone	*Saccharomyces cerevisiae*	26	4.05
1AW0	NMR	Fourth metal-binding domain of the Menkes copper transporting ATPase	*Homo sapiens*	29	1.57
1KVI	NMR	First metal-binding domain of the Menkes copper transporting ATase	*Homo sapiens*	35	2.44
2EW9	NMR	Wilson protein domains 5 and 6	*Homo sapiens*	36	2.51
1OSD	X-ray	Periplasmic Mercury-binding Protein MerP	*Ralstonia metallidurans*	38	1.66
1AFI	NMR	MerP	*Shigella flexneri*	38	2.39

NB – RMSD result refers to the all atom value, X-ray = X-ray diffraction and NMR = Solution NMR

These results show how metallochaperones with a low sequence identity to CopZ can have superficially similar structures. Most noticeably the fourth metal-binding domain of the Menkes copper transporting ATPase and the *Ralstonia metallidurans* form of MerP that both give an RMSD lower than 2.0 Å form their structural alignments, giving them an apparent better fit than all of the existing CopZ structures in the PDB.



```
      1                 10          20            30
CopZ  ME.....QK.TLQV.EGMSCQHCVKAVETSVGELD.GVSAV
1AFI  AT.....QTVTLAV.PGMTCAACPITVKKALSKVE.GVSKV
1OSD  AT.....QTVTLSV.PGMTCSACPITVKKAISKVE.GVSKV
2EW9  MAP....QKCFLQI.KGMTCASCVSNIERNLQKEA.GVLSV
1KVI  MDPSMGVNSVTISV.EGMTCNSCVWTIEQQIGKVN.GVHHI
1AW0  LT.....QETVINI.DGMTCNSCVQSIEGVISKKP.GVKSI
1CC8  MA.....EIKHYQFNVVMTCSGCSGAVNKVLTKLEPDVSKI
1FEE  MP.....KH.EFSV..DMTCGGCAEAVSRVLNKLG.GVK.Y

             40          50          60         70
CopZ  HVNLEAGKVDVSFDADKVSVKDIADAIEDQGYDVAKIEGR
1AFI  DVGFEKREAVVTFDDTKASVQKLTKATADAGYPS..SVKQ
1OSD  DVTFETRQAVVTFDDAKTSVQKLTKATADAGYPS..SVKQ
2EW9  LVALMAGKAEIKYDPEVIQPLEIAQFIQDLGFEAA....V
1KVI  KVSLEEKNATIIYDPKLQTPKTLQEAIDDMGFDAVIHNPD
1AW0  RVSLANSNGTVEYDPLLTSPETLRGAIEDMGFDATL..SD
1CC8  DISLEKQLVDVYTTLPY...DFILEKIKKTGKEVRSGKQL
1FEE  DIDLPNKKVCIESEHSM...DTLLATLKKTGKTVSYLGLE
```

Figure 4.15 - A multiple sequence alignment of CopZ with other metallochaperones of known structure. Identical residues are indicated by a red background, conservatively varied residues are boxed in blue and shown in red characters.

A multiple sequence alignment (Figure 4.15) identified a conserved MXCXXC metal binding motif in each protein, as well as a glycine residue found 46-47 residues along the amino acid sequence. This glycine residue is in the vicinity of the metal binding site and although it appears to be too far away to directly affect metal binding, it does have the potential to be involved in inter-subunit interactions, as identified in the $Cu_3(CopZ)_3$ structure where the corresponding glycine residue (Gly65), along with Asp66, could potentially form a hydrogen bond with a glutamine residue (Gln14) form the adjacent monomer.

4.3.9 Structure predictions for polynuclear copper cluster proteins

4.3.9.1 Selection of existing poly-nuclear copper cluster proteins with solved structures

The search of the PDB produced 129 hits, of which only four referred to proteins with ploy-nuclear copper clusters; the two CopZ structures solved in this chapter **[Hearnshaw *et al* 2009, Singleton *et al* 2009]**, and two structures for yeast metallothionein **[Calderone *et al* 2005, Peterson *et al* 1996]**. The two CopZ structures were unsuitable for HMM based analysis as the copper cluster was located at the dimer interface, with residues from both monomers used to coordinate the cluster, therefore only the metallothionein structures were used for HMM generation.

4.3.9.2 Results of HMM creation and searches

A number of HMMs were built based on copper binding in the two yeast metallothionein structures, these HMMs accounted for the different copper packing observed in the eight copper cluster. Figure 4.16 shows the copper cluster from one of the yeast metallothionein structures and identifies one of the HMMs constructed from it.

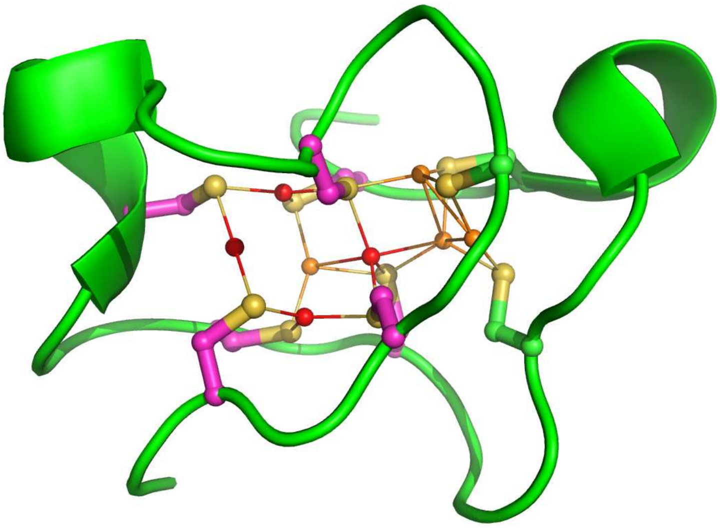


Figure 4.16 – The structure of the yeast metallothionein (PDB ID: 1RJU **[Calderone *et al* 2005]**), showing the 8 copper cluster (orange and red spheres) and coordinating cysteine residues. An example of a copper cluster structure used to build a HMM is marked by red spheres and the cysteines that coordinate it are coloured magenta.

4.3.9.3 Prediction of Ace1 tetra-nuclear copper cluster

HMM searches on the Ace1 sequence were unable to give a consistent and definite prediction for the make up of the tetra-nuclear copper cluster, however the location of the hits over a specific region of the Ace1 copper binding domain identified 6 out of the 8 cysteine residues that could be tentatively assigned as the copper coordinating residues for structural modelling, using a structure for a Cu_4S_6 ring from the Cambridge structural database as a template (Figure 4.17).

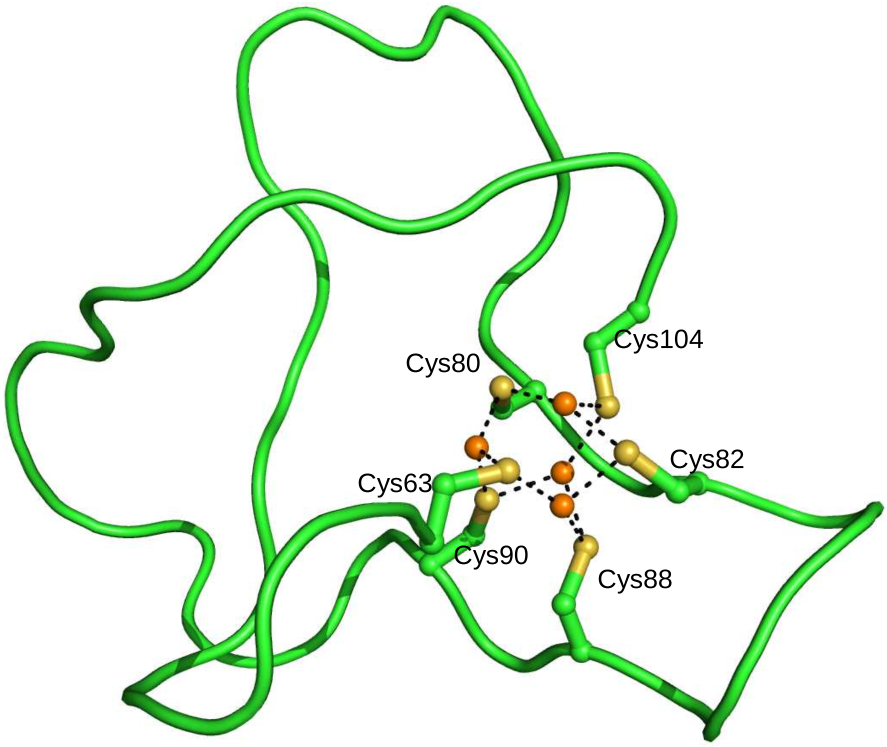


Figure 4.17 – Structure prediction for the copper binding domain of Ace1, covering residues 63-104 of the Ace1 sequence.

A Ramachandran analysis of this structure suggests it is at least protein like, with none of the residues falling into disallowed regions of the Ramachandran plot.

4.3.9.4 Prediction of Mac1 tetra-nuclear copper cluster

The prediction for the two tetra-nuclear copper clusters for Mac1 built around the C1 and C2 copper binding motifs **[Keller *et al* 2000]** were made using the cluster from the $Cu_4(CopZ)_2$ as a template (Figure 4.18), due to the lack on consistent hits with the HMM based analysis and the existence of a histidine ligand in the $Cu_4(CopZ)_2$ cluster that is also found in the C1 and C2 motifs.

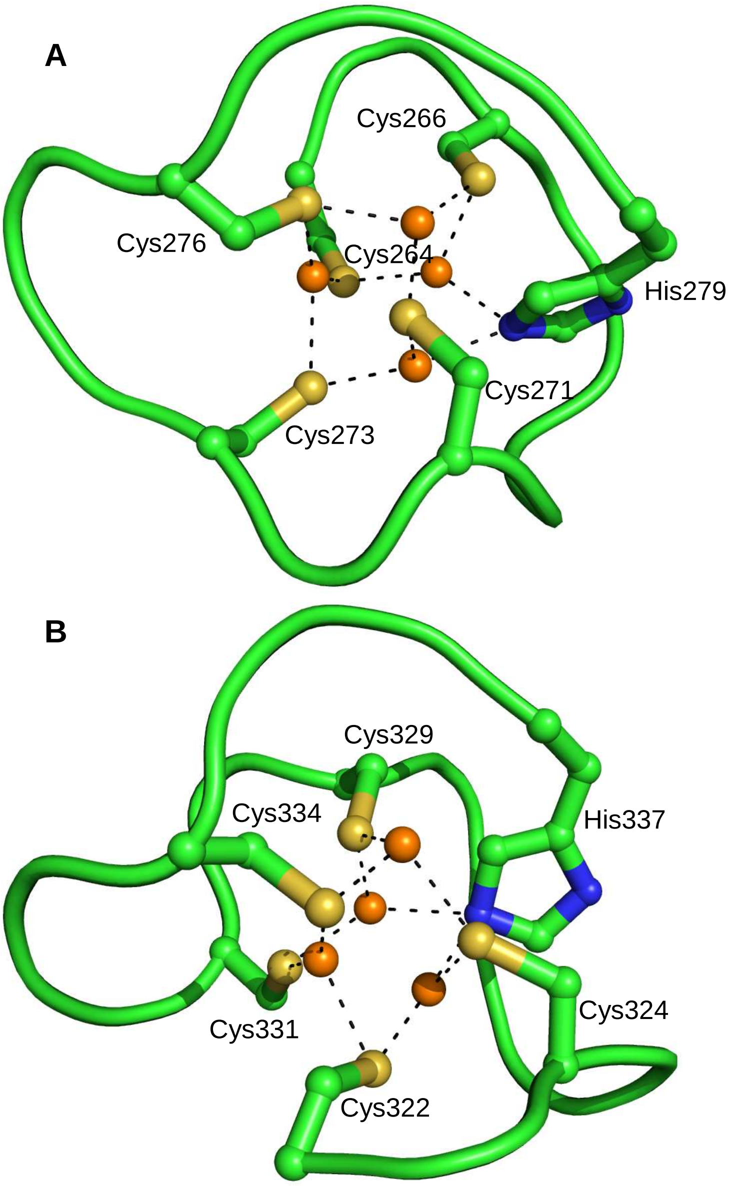


Figure 4.18 – Predicted structures for **(A)** the C1 and **(B)** C2 copper binding motifs of Mac1.

Ramachandran analyses of these models highlight some problems, with both structures having residues in the disallowed regions of the plot; the C1 and C2 based models have 2 and 3 residues respectively in disallowed regions, accounting for 15.4 and 27.3 % of the modelled structures.

4.4 Discussion

4.4.1 Why do the stoichiometries of the CopZ crystal structures not match those identified in solution experiments?

The expected stoichiometries of CopZ monomers to copper ions identified in solution were not the same as those identified in the crystal, with the $Cu_1(CopZ)_2$ solution resulting in a $Cu_3(CopZ)_3$ crystal structure and the $Cu_2(CopZ)_2$ solution resulting in a $Cu_4(CopZ)_2$ crystal structure. We must, therefore, ask why these unexpected stoichiometries were formed.

The electron density maps for the $Cu_3(CopZ)_3$ structure (Figure 4.10) show the protein is definitely trimeric due to the quality of the fit with the copper binding residues and the copper atoms themselves, so there is no major problem with the interpretation. A possible explanation for this change is a weak association between the two monomers of the $Cu_1(CopZ)_2$ dimer, causing them to separate and reform into the $Cu_3(CopZ)_3$ trimeric form under the forces exerted during crystallisation. Alternatively the $Cu_1(CopZ)_2$ dimeric form of CopZ may be energetically unfavourable in the crystal and therefore unstable, causing the protein to change into the $Cu_3(CopZ)_3$ trimeric form in solution before crystallisation. It is unclear exactly how or why CopZ changes its copper binding structure, but it is an indication of the flexibility of the protein.

4.4.2 How the CopZ structure compares with existing structures for CopZ and other metallochaperones?

As has been shown, all currently available Atx1 like metallochaperone structures share a similar structure, based around a ferredoxin-like βαββαβ-fold (Table 4.8). However, the way these monomers aggregate around varying numbers of Cu(I) ions to form higher order clusters can be quite different.

Recent work by Badarau *et al* has identified Atx1-like copper chaperone structures that bind copper in stoichiometries similar to those observed in CopZ, they solved the structures of a two copper dimer, a four copper dimer with a “head-to-head” arrangement (Figure 4.19A), a four copper dimer with a “side-to-side” arrangement (Figure 4.19A) and a four copper trimer **[Badarau *et al* 2010]**.

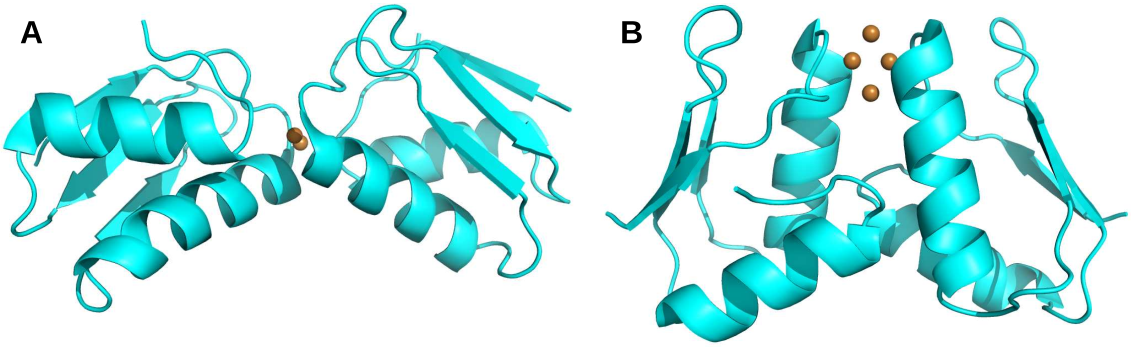


Figure 4.19 – Structures of the **(A)** head-to-head and **(B)** side-to-side dimer formations found by Badarau *et al*. [2010].

The two copper dimer coordinates its copper using the two cysteine residues from each monomer at the copper binding interface (Cys12 and Cys15), the four copper dimer with the side-to-side arrangement coordinates its coppers using the same cysteine residues and two chloride ions present in the solution, the four copper dimer with the head-to-head arrangement coordinates its coppers using the same cysteine residues and a histidine residue (His61) from each monomer (different to he His15 residue used to coordinate the four coppers in the CopZ dimer) and the four copper trimer coordinates its coppers using the two cysteine residues from each monomer.

The two CopZ structures, $Cu_4(CopZ)_2$ and $Cu_3(CopZ)_3$ have RMSDs of 3.7 and 2.1 Å with their respective Atx1 homologues (the four copper dimer with the head-to-head arrangement for $Cu_4(CopZ)_2$ and the four copper trimer for $Cu_3(CopZ)_3$) and have quite different methods of copper coordination (Figure 4.20)

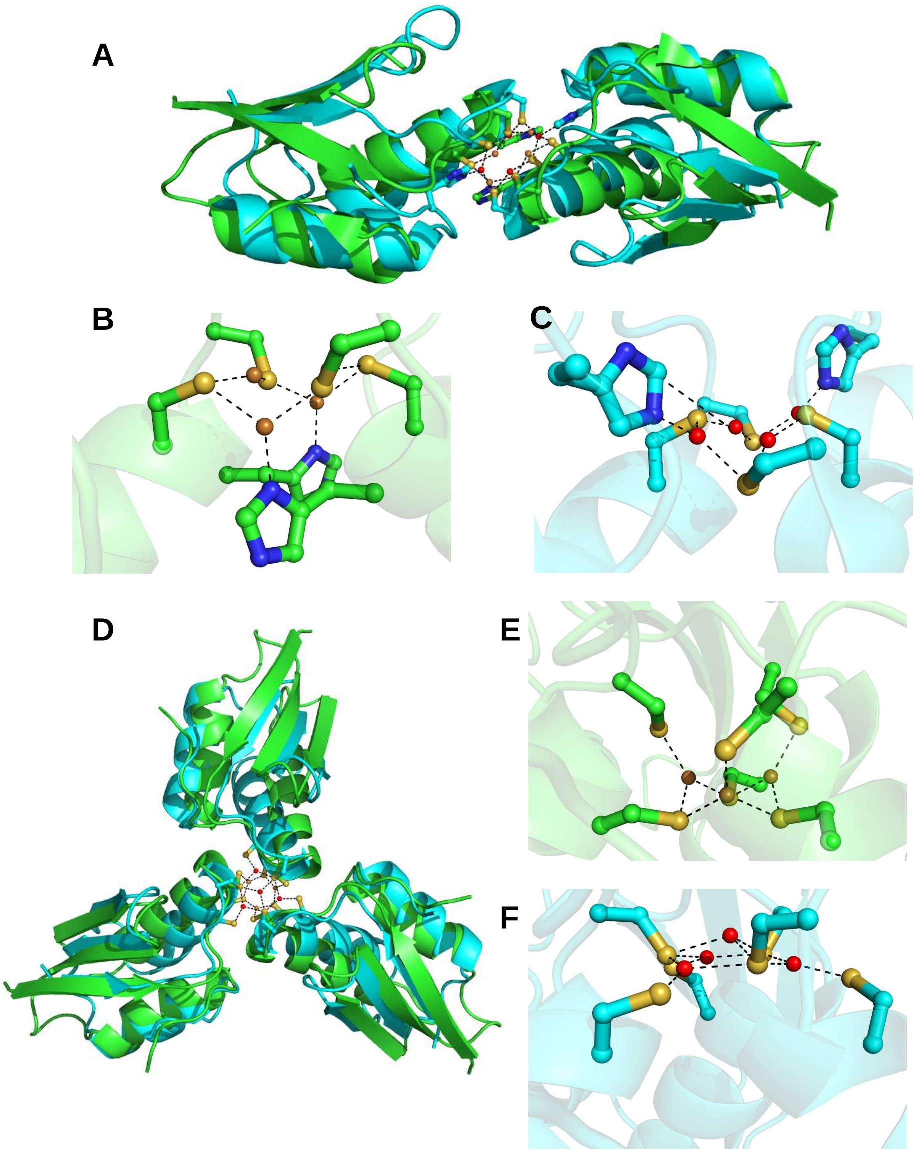


Figure 4.20 – A comparison of the CopZ structures (Green polypeptide and orange coppers) and the Badarau *et al* Atx1 structures (Cyan polypeptide and red coppers). Panel **(A)** shows the alignment of the two dimers from above the copper binding interface, panels **(B)** and **(C)** show the copper binding interface of each dimer, showing the differences in copper binding in each structure. Panel **(D)** shows the alignment of the two trimers down the 3 fold axis, panels **(E)** and **(F)** show the copper binding interface of each trimer, showing the differences in copper binding in each structure.

These structures show the versatility of the Atx1-like copper chaperones with respect to the binding of copper and how they are able to pack their individual coppers in a variety of ways and use a variety of ligands (not just the highly conserved MXCXXC motif) to coordinate their copper atoms.

4.4.3 Modelling putative copper transfer complexes

Multhaup *et al* have shown that CopZ can bind to CopA in both the presence and absence of copper **[Multhaup *et al* 2001]**. This binding is much stronger with copper present, as demonstrated by the observed 15-fold decrease in the dissociation rate, k_d. However, little is known about how this binding could occur, other than it appears to be the CXXC motif in CopA that is responsible for binding copper. The high-resolution structure of a complex involving three CopZ monomers reported here not only demonstrates an extraordinary flexibility in Cu(I) co-ordination and monomer–monomer interactions, but also raises the possibility that a transient trimeric heterocomplex between CopZ and CopA could occur as part of a novel Cu(I)-transfer mechanism. Molecular modelling was used to test this hypothesis.

Models were constructed demonstrating potential structures for how CopA may bind to CopZ in the presence of Cu(I), using the $Cu_3(CopZ)_3$ and $Cu_4(CopZ)_2$ structures as templates for modelling the interaction. Two models were created for the CopA/$Cu_3(CopZ)_3$ hybrid structures, with either the N1 ($Cu_3(CopZ)_2(CopA^{N1})$) (Figure 4.21A) or N2 ($Cu_3(CopZ)_2(CopA^{N2})$) (Figure 4.21B) subunits of the CopA NMR structure replacing one of the monomers of the $Cu_3(CopZ)_3$ structure. Two models were also created for the $Cu_4(CopZ)_2$ hybrid structures, again with the N1 ($Cu_4(CopZ)_1(CopA^{N1})$) (Figure 4.21C) or the N2 ($Cu_4(CopZ)_1(CopA^{N2})$) (Figure 4.21D) subunit of the CopA NMR structure replacing one of the monomers of the $Cu_4(CopZ)_2$ structure. All these models seem to be plausible as there are no steric clashes between the domains of CopZ and CopA in any structure.

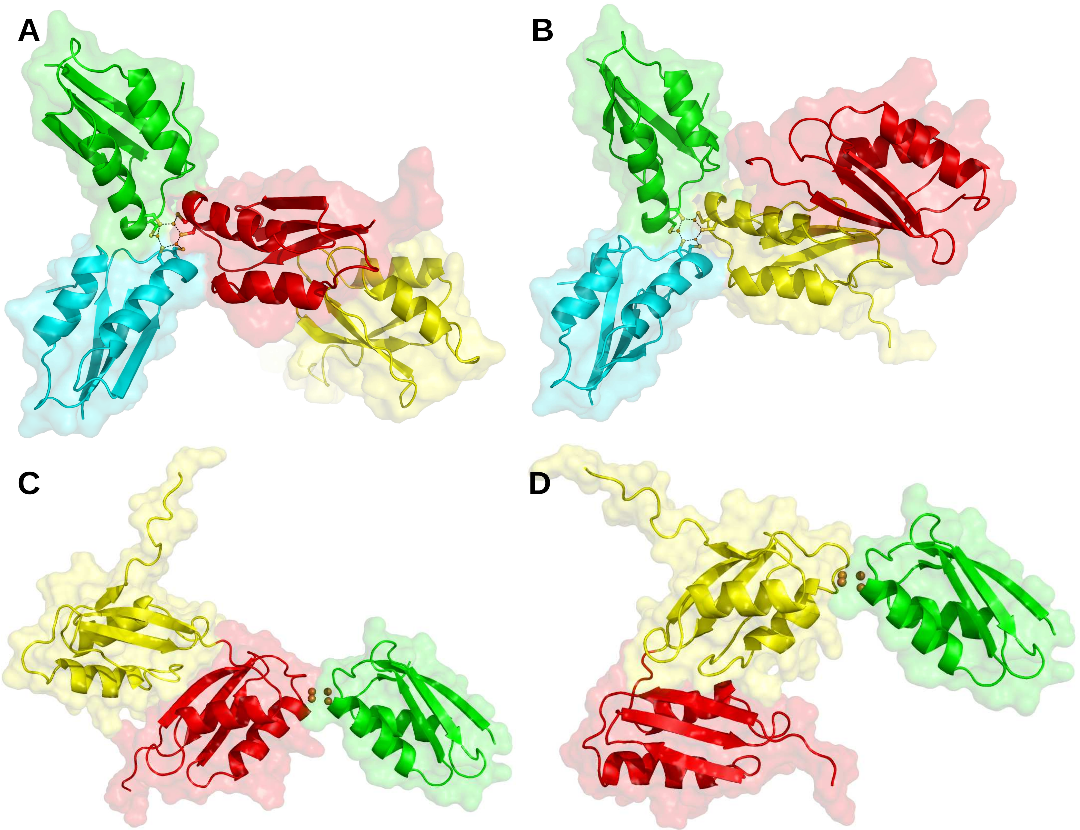


Figure 4.21 – (A) Models for $Cu_3(CopZ)_2(CopA^{N1})$ and **(B)** $Cu_3(CopZ)_2(CopA^{N2})$. **(C)** Models for $Cu_4(CopZ)_1(CopA^{N1})$ and **(D)** $Cu_4(CopZ)_1(CopA^{N2})$. The N1- and N2-domains of CopA are coloured red and yellow respectively. Copper ions are shown as orange spheres.

Two structures for CopA were modelled using the Swiss Model server **[Arnold *et al* 2006]** and utilising a monomer from the $Cu_3(CopZ)_3$ and $Cu_4(CopZ)_2$ structures as a template. It is arguably the case that this template would create a more accurate structure for modelling side chain interactions than the NMR structure **[Banci *et al* 2003]**. Figure 4.22 shows copper binding sites of the CopA/CopZ hybrid structures created using these CopA models.

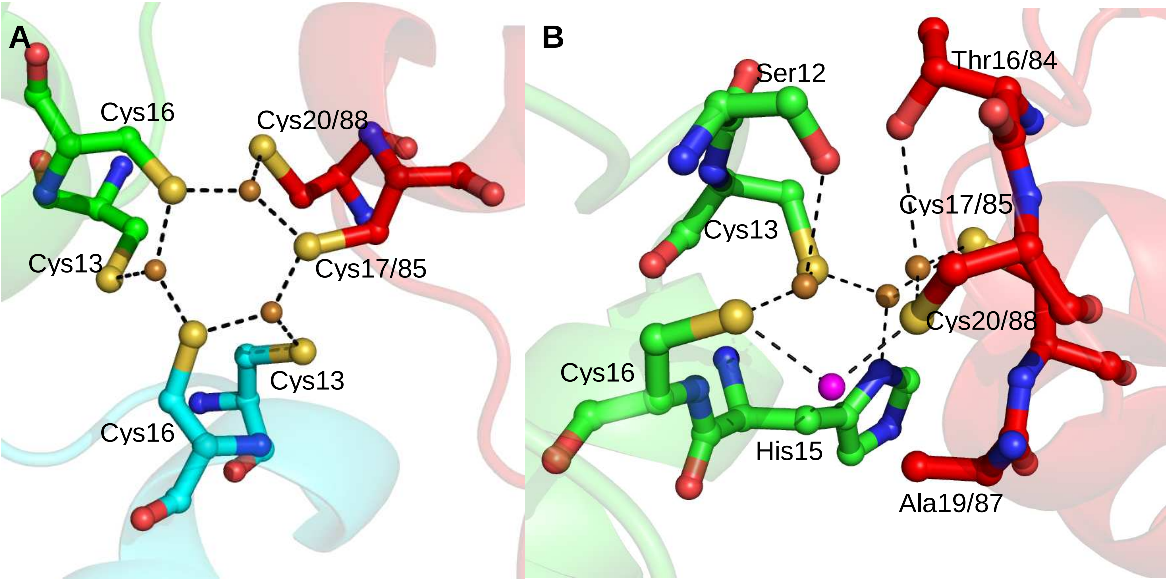


Figure 4.22 – The copper binding site of proposed CopA/CopZ hybrid structures **(A)** The CopA(I)/$Cu_3(CopZ)_3$(II) hybrid structure. CopA residues are coloured red, CopZ monomers green and cyan, and coppers orange. **(B)** The CopA(I)/$Cu_4(CopZ)_2$(I) hybrid structure. CopA is coloured green and CopZ red. Orange spheres indicate coppers that have sufficient coordination via neighbouring residues; the magenta sphere indicates the copper that does not. (Since both the N1 & N2 subunits have the same binding site, only one is shown, but the residue numbers for both are displayed).

The CopA(I)/$Cu_3(CopZ)_3$(II) models were able to maintain all coordinating reactions mediating the trinuclear copper cluster with both the N1 or N2 subunit of CopA at the trimer interface, through the sulfurs of cysteines 17 & 20 or 85 & 88 respectively (Figure 4.22A). However, the CopA(I)/$Cu_4(CopZ)_2$(I) model is unable to maintain the tetranuclear copper cluster, due to the absence of a suitable residue to provide the coordination offered by His15 of CopZ. Three of the four copper ions could be coordinated once again by cysteines 17 and 20 or 85 and 88 (depending on the CopA domain present at the interface), and a threonine residue conserved in both domains (Thr 16/84) would be able to provide the secondary coordination afforded by Serine 12 in CopZ (Figure 4.22B).

It has been shown that CopA can bind CopZ not only in the presence of copper, but also in the absence of copper **[Multhaup *et al* 2001]**, and to accomplish this there must be some other protein-protein interactions, apart from those mediating for formation of the copper cluster. Like the $Cu_4(CopZ)_2$ crystal structure, the CopA(I)/$Cu_4(CopZ)_2$(I) hybrid model did not contain any additional intersubunit interactions beyond those coordinating the copper cluster. There are however a number of potential intersubunit hydrogen-bonding interactions in the hybrid models based on the $Cu_3(CopZ)_3$ structure, where the N1 and N2 domains of CopA are bound at the interface, containing 5 and 6 hydrogen-bonding interactions respectively (Figure 4.23). The $Cu_3(CopZ)_2(CopA^{N1})$ model has the 3 hydrogen-bonds found between the CopZ monomers already identified in the $Cu_3(CopZ)_3$ structure (Figure 4.12C) and has two hydrogen-bonds at the CopA-CopZ interface, between; CopA-Leu67↔CopZ-His15 and CopA-Lys68↔CopZ-Lys18 (Figure 4.23A). The $Cu_3(CopZ)_2(CopA^{N2})$ model also has the 3 hydrogen-bonds found between the CopZ monomers already identified in the $Cu_3(CopZ)_3$ structure (Figure 4.12C) and has three hydrogen-bonds at the CopA-CopZ interface, between; CopA-Asn90↔CopZ-Gln63, CopA-Lys124↔CopZ-Lys18 and CopA-Leu125↔CopZ-His15 (Figure 4.23B).

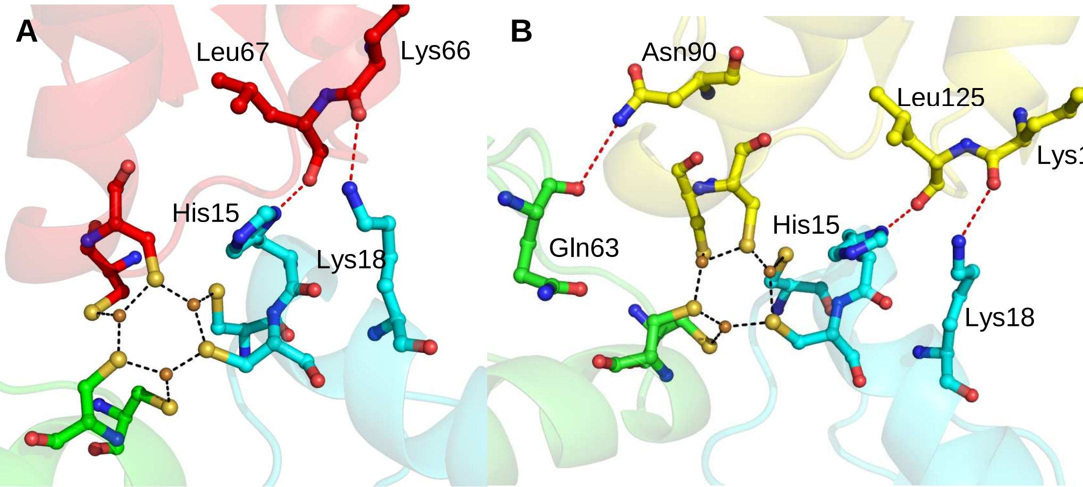


Figure 4.23 – Hydrogen bonds formed between the CopZ and CopA subunits of **(A)** $Cu_3(CopZ)_2(CopA^{N1})$ and **(B)** $Cu_3(CopZ)_2(CopA^{N2})$ structures.

The specific interactions between molecular subunits at the interfaces in the trimeric complexes are limited. The total surface area lost on formation of the three modelled complexes from individual subunits ranged from 702 to 783 $Å^2$. Of this, approx. 60% was contributed by non-polar and neutral atoms. The usual indicators of permanency for protein complexes (e.g. **[Ponsting *et al* 2000]**) cannot be used for situations such as this where a major driver for stability is presumably the formation of specific Cu(I) ion to cysteine thiolate co-ordinate bonds. However, the relatively low value of the solvent-accessible area lost on complex formation and the low number of inter-subunit hydrogen bonds are at least consistent with a tentative classification of the modelled complexes as transient.

A potential method to test the validity of these models would be to prepare protein solutions containing stable forms of differing mixtures of CopA bound to CopZ and copper ions for crystallisation experiments, leading to X-ray data collection to ascertain a 3D structure. However the predicted transient nature of this complex is likely to make isolating a complex stable enough to form protein crystals quite challenging, if this is the case, surface plasmon resonance **[Van Der Merwe 2001]** or analytical ultracentrifugation are other potential techniques that could be analysed to assess CopA-CopZ complex formation.

The residues identified as potentially important for inter-subunit binding between CopZ and CopA (His15, Lys18 & Gln63 from CopZ, Lys66 & Leu67 from $CopA^{N1}$ and Asn90, Lys134 & Leu135 from $CopA^{N2}$) could be tested by site directed mutagenesis, mutating these residues and performing solution experiments capable of analysing protein binding, such as surface plasmon resonance **[Van Der Merwe 2001]** or analytical ultracentrifugation.

4.4.4 Comparison of $Cu_4(CopZ)_2$ with other tetra-copper cluster containing proteins and structure predications for these proteins

Mac1 and Ace1 are two copper-regulated transcription factors from *Saccharomyces cerevisiae* **[Keller *et al* 2000]**. X-ray absorption spectroscopy on the copper regulatory domains of these proteins has revealed remarkably similar tetra-copper clusters **[Brown *et al* 2002]**. The precise coordination of these copper binding sites is currently unknown, as no structure is available. However, suggestions have been put forward by Brown *et al* as to the layout of these copper clusters **[Brown *et al* 2002]**. Mac1 is thought to bind copper via a CXCXXXXCXCXXCXXH motif providing trigonal coordination for each copper though five cysteine residues and one histidine (Figure 4.24A). Copper binding in Ace1 is thought to be coordinated by 6 (Figure 4.24B) or eight (Figure 4.24C) cysteine residues over a 60 residue Cys-rich domain, providing each copper with a trigonal coordination.

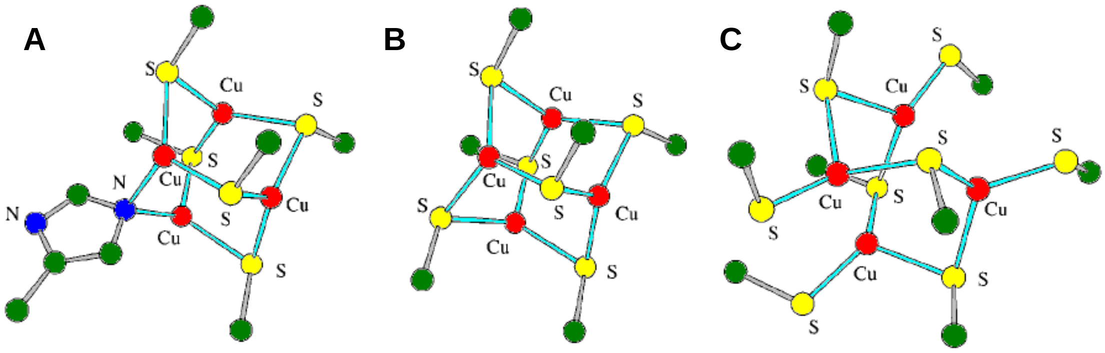


Figure 4.24 – Proposed structures for copper binding sites from **(A)** Mac1, **(B)** Ace1, utilising 6 cysteine residues and **(C)** Ace1, utilising 8 cysteine residues. **[Brown *et al* 2002]** Copper ions are coloured red, sulfur yellow, carbon green and nitrogen blue. Peptide bonds are coloured grey and cluster coordinating bonds cyan.

Structure predictions for Mac1 and Ace1 using the HMM based prediction methodology had limited success in predicting the structures of poly-nuclear coppers clusters. This is primarily due to the lack of existing structures for poly-nuclear copper cluster containing proteins, with yeast metallothionein providing the only template for HMM construction. This method may have more success in the future once the structures of more poly-nuclear coppers cluster containing proteins have been solved. The structures that have been proposed for Mac1 (Figure 4.18) and Ace1 (Figure 4.17) do at least give some indication as to the possible structure of the polypeptide that coordinates the coppers in these tetranuclear clusters.

Chapter 5 - Structural studies of the Sulfide Dehydrogenase Flavoprotein SoxF of *Paracoccus pantotrophus* and insights into its role in the *sox* cycle

5.1 Introduction

5.1.1 FAD structure and properties

Flavin Adenine Dinucleotide (FAD) is a redox cofactor involved in metabolic reactions that can exist in two different redox states, which it converts between by accepting or donating electrons. The molecule itself consists of a riboflavin moiety (vitamin B_2) bound to the phosphate group of and ADP molecule (Figure 5.1A) and undergoes its redox reactions on the isoalloxazine rings of the riboflavin subunit (Figure 5.1B).

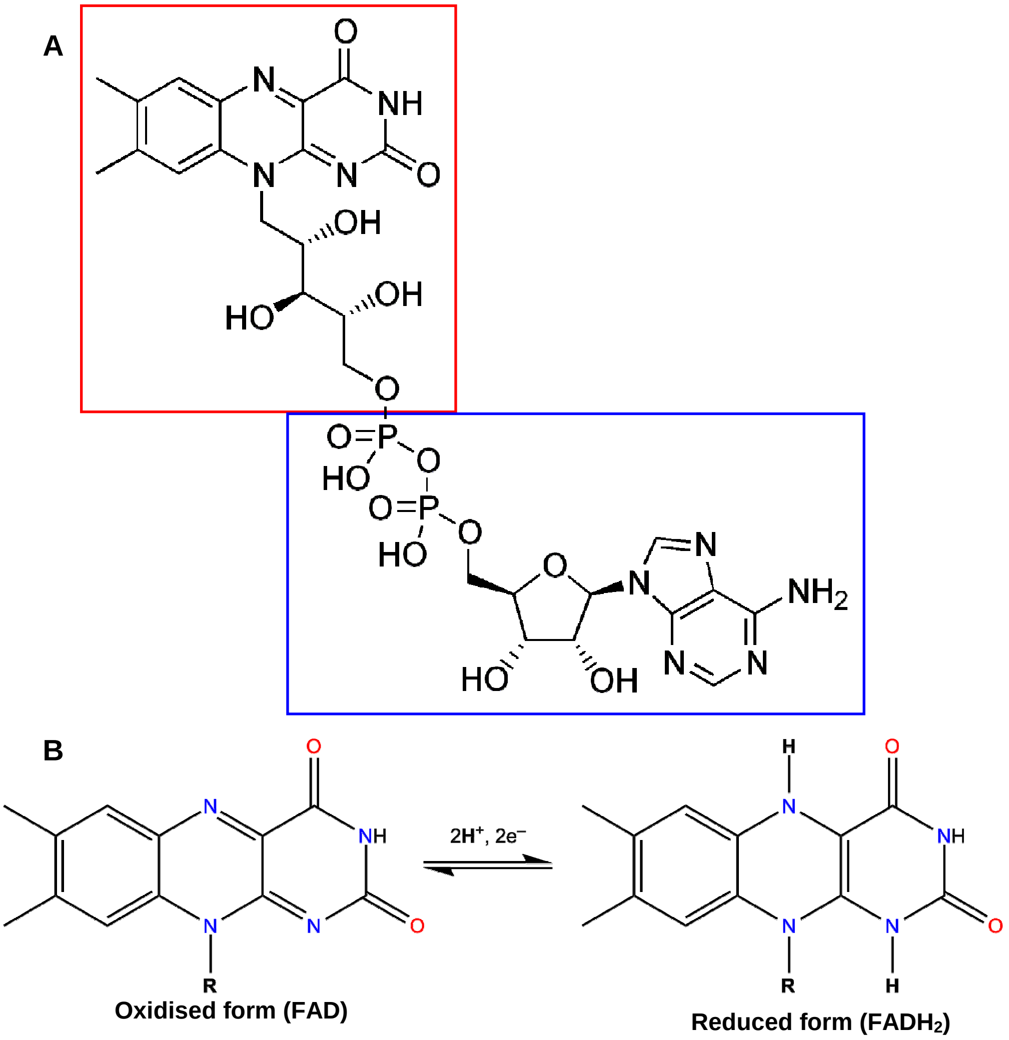


Figure 5.1 – The chemical structure of FAD **(A)** with the riboflavin subunit highlighted by a red box and the ADP subunit by a blue box. **(B)** The equilibrium between the oxidised and reduced forms of FAD, showing where the electrons are added/removed.

Flavin binding proteins or flavoproteins are involved in a wide range of biological processes, including; sulfur oxidisation, bioluminescence, photosynthesis, DNA repair, apoptosis and the removal of radicals leading to oxidative stress.

5.1.2 Succinate dehydrogenase

Succinate dehydrogenase (SDH) is an example of a flavoprotein. SDH is an enzyme complex bound to the inner mitochondrial membrane of mammalian mitochondria and many bacterial cells. SDH is part of both the citric acid cycle. In the citric acid cycle it is responsible for the oxidation of succinate to fumarate and the reduction of ubiquinone to ubiquinol. In the electron transport chain it is responsible for the delivery of electrons to the quinone pool.

Eukaryotic SDH consists of four subunits (Figure 5.2). These are arranged into a hydrophilic head that protrudes into the matrix of the mitochondrion, consisting of two subunits, a flavoprotein (Sdh1) and an iron-sulfur protein (Sdh2), which form the catalytic core of the complex, and a hydrophobic membrane anchor that is embedded into the inner mitochondrial matrix with a short segment protruding into the soluble inner membrane space, consisting of two subunits, Sdh3 and Shd4, that bind a B-type heme at the subunit interface with each subunit providing one of the two axial His ligands **[Sun *et al* 2005]**.

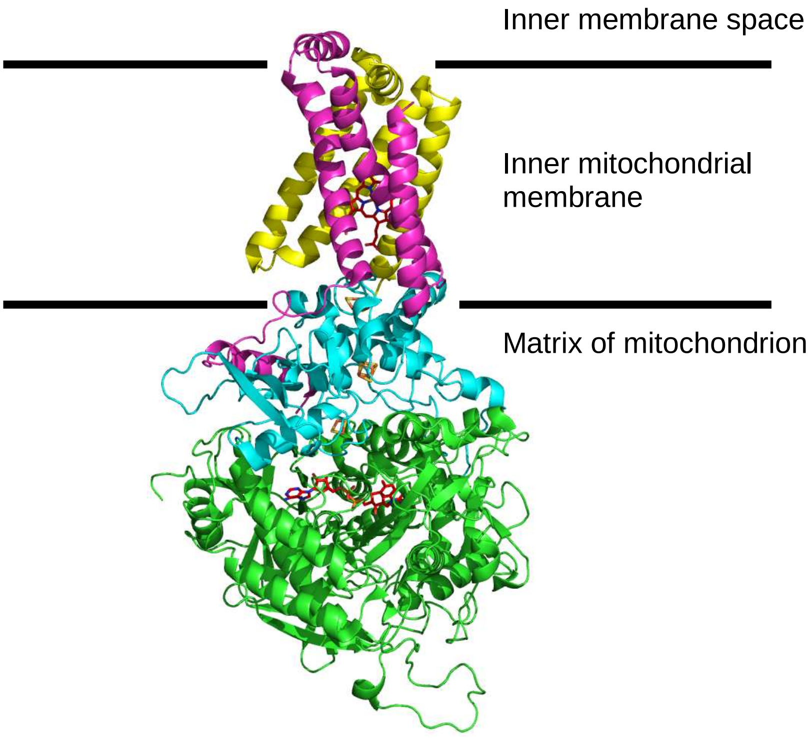


Figure 5.2 – Structure of the succinate dehydrogenase enzyme complex (PDB ID: 3ABV **[Harada *et al* 2011]**) and its position in the mitochondrial membrane. The subunits that form the hydrophilic head, Sdh1 and Sdh2, are coloured green and cyan respectively, the subunits that form the hydrophobic membrane anchor, Sdh3 and Sdh4, are coloured magenta and yellow respectively. The FAD, heme and iron sulfur cofactors are displayed as sticks.

The binding site for succinate oxidation is found on Sdh1. The side chains of residues Thr254, His354, and Arg399 stabilise the succinate molecule while FAD oxidises it and carries the electrons to the first of the iron-sulfur clusters in Sdh2. The electrons are tunnelled through Sdh2 along the iron-sulfur cluster relay to one of two potential ubiquinone binding sites; the higher affinity Q_P site, formed by residues from Shd2, Shd3 and Sdh4 and the lower affinity Q_D site, formed by residues from Sdh3 and Sdh4, where ubiquinone is reduced to ubiquinol **[Yankovskaya *et al* 2003, Sun *et al* 2005]**.

The role of the B-type heme associated with Sdh3 and Sdh4 remains unclear. It has been shown that reduction of ubiquinone can still take place without the heme moiety and that the affect on the catalytic activity of the complex is minimal **[Oyedotun *et al* 2007]**, suggesting the heme is not needed for ubiquinone reduction. Rutter *et al* have speculated that the ubiquinone reduction is able to take place at the Q_P site without using the heme, but that the heme would be needed to mediate the transfer of electrons to the Q_D site to allow ubiquinone reduction at the lower affinity site **[Rutter *et al* 2010]**.

5.1.3 Sulfur oxidising flavoproteins

Sulfur, the 10th most abundant element in the universe, is a brittle, yellow, non-metallic element that occurs in all living matter as a component of methionine and cysteine amino acids, it has critical roles in both climate and in the health of various ecosystems **[Environmental Literacy Council 2006]**.

Most of the Earth's sulfur is contained in rocks and salts or buried deep in the ocean in oceanic sediments. Sulfur is also found in the atmosphere and can enter through both natural and human sources. Natural resources include; volcanic eruptions, bacterial processes, evaporation from water, or decaying organisms. While human sources for atmospheric sulfur are primarily a consequence of industrial processes where sulfur dioxide (SO_2) and hydrogen sulfide (H_2S) gases are emitted on a wide scale.

When sulfur dioxide enters the atmosphere it reacts with oxygen to produce sulfur trioxide gas (SO_3), or with other chemicals in the atmosphere to produce sulfur salts. Sulfur dioxide can also react with water to produce sulfuric acid (H_2SO_4). All these particles will settle back onto earth, or react with rain and fall back as acid deposition. The particles will then be absorbed by plants again and are released back into the atmosphere, so that the sulfur cycle will start over again **[Environmental Literacy Council 2006]**.

5.1.3.1 Sulfide:Quinone oxidoreductases

The Sulfide:Quinone oxidoreductases (SQRs) are another example of flavoproteins. Homologues of the SQRs are found in all domains of life except plants and play a physiological role in both sulfide detoxification and energy transduction **[Marcia *et al* 2009]**. The SQRs oxidize sulfide ions (S$_2^{-}$, HS^{-}) to zero valent sulfur, thought to be released from the protein as a polysulfide chain of up to 10 sulfur atoms **[Greisbeck *et al* 2002]**.

Several SQR structures have been solved **[Brito *et al* 2009, Marcia *et al* 2009, Cherney *et al* 2010]** that have identified the catalytic cysteine residues at the active site and shown how they are able to bind varying numbers of sulfur atoms, this fits with the belief that the SQRs are responsible for creating long chains of sulfur atoms. Figure 5.3 shows an example of an SQR structure and some of the different active site compositions that have been discovered thus far.

The exact mechanism for sulfide oxidation is currently unknown. Several mechanisms have been put forward **[Brito *et al* 2009, Cherney *et al* 2010, Marcia *et al* 2010]**, with a common theme between all of them of the FAD cofactor acting as the electron donor/acceptor for the mechanism, highlighting the vital role of FAD in the system.

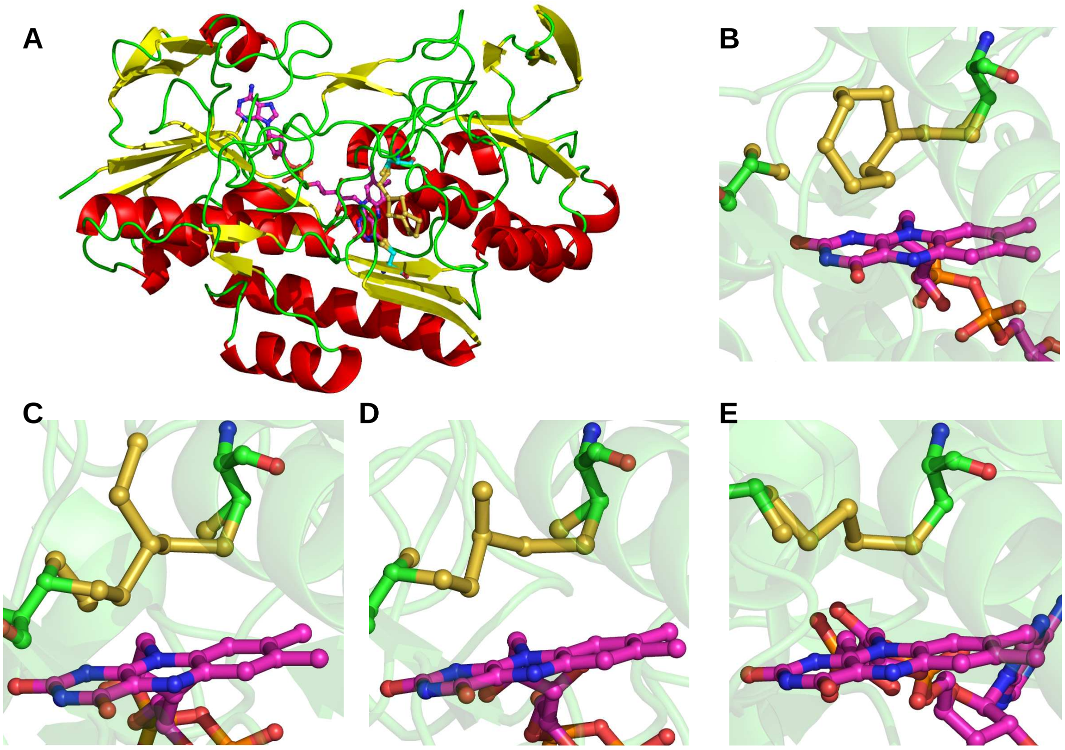


Figure 5.3 – The structure of an SQR and a selection of SQR active sites whose structures have been published. **(A)** Shows the overall secondary structure of a monomer of the SQR from *Aquifex aeolicus* (PDB ID: 3HYV **[Marcia *et al* 2009]**), with α-helices coloured red, β-sheets yellow and turns and loops green, the FAD cofactor is coloured magenta and the active cysteine residues are coloured cyan. **(B)** Shows the active site of the *Aquifex aeolicus* SQR with a chain of nine sulfurs with an S8 ring that have built up on cysteine 156. **(C)** Shows the active site of the *Acidithiobacillus ferrooxidans* SQR (PDB ID: 3KPG **[Cherney *et al* 2010]**) with a chain of five sulfurs between the two cysteine residues. **(D)** Shows a different sulfur composition in the active site of the *Acidithiobacillus ferrooxidans* SQR (PDB ID: 3KPI **[Cherney *et al* 2010]**) with a chain of four sulfurs between the two cysteine residues. (E) Shows the active site of the *Acidianus ambivalens* SQR (PDB ID: 3H8I **[Brito *et al* 2009]**) with a chain of two sulfurs between the two cysteine residues.

5.1.3.2 The sox system

The α-Proteobacterium *Paracoccus pantotrophus* is an example of an organism that can oxidise inorganic sulfur species to sulfate, via its sulfur oxidizing, or "sox" system, which is found in both photosynthetic and non-photosynthetic sulfur-oxidizing Eubacteria **[Sauve et al 2007]**. This oxidation of inorganic sulfur species to sulphate by bacteria, such as *Paracoccus pantotrophus*, is a vital part of the global sulfur cycle **[Freidrich *et al* 2005]** and is important for agriculture (through the oxidation of inorganic reduced compounds), waste water treatment (thorough the oxidation of toxic hydrogen sulfuide to relatively harmless sulfate) and biomining (through mineral decomposition) **[Rawlings 2002]** The sox gene cluster of *P.pantotrophus* comprises 15 genes, organised into three transcriptional units; *soxRS*, *soxVW* and *soxXYZABCDEFGH* (Figure 5.4). The gene *soxR* codes for a DNA-binding repressor protein of the AsrR family and *soxS* codes for a periplasmic thioredoxin that has been shown to be essential for full expression **[Rother *et al* 2005]**. The *soxVW* genes comprise a transcriptional unit *soxV* that codes for the membrane protein SoxV, a channel protein with six transmembrane helices, responsible for transport of reductant and *soxW* codes for a periplasmic thioredoxin **[Fredrich 2008]**. The *soxXYZABCDEFGH* genes code for the 7 core proteins of the sox cycle (SoxXYZABCD), a small c-type cytochrome (SoxE), a flavoprotein with sulfite dehyrogenase activity (SoxF), a protein with two zinc binding motifs (SoxG) and a protein with two metal binding motifs (SoxH) **[Rother *et al* 2001]**.

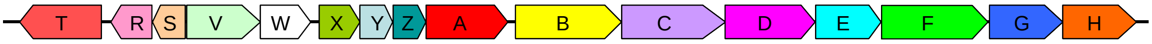


Figure 5.4 - Schematic map of the *sox* gene cluster of *P.pantotrophus*.

The current model for the Sox pathway from *P.pantotrophus* is shown in Figure 5.5. SoxAX initiates oxidation of thiosulfate to form SoxY-thiocysteine-S-sulfate, SoxB hydrolyzes sulfate from the thiocysteine-S-sulfate residue to give S-thiocysteine, SoxCD then oxidizes the outer sulfur atom to SoxY-cysteine-S-sulfate and finally, sulfate can again be hydrolyzed and removed by SoxB to regenerate the cysteine residue of SoxY **[Friedrich *et al* 2001]**.

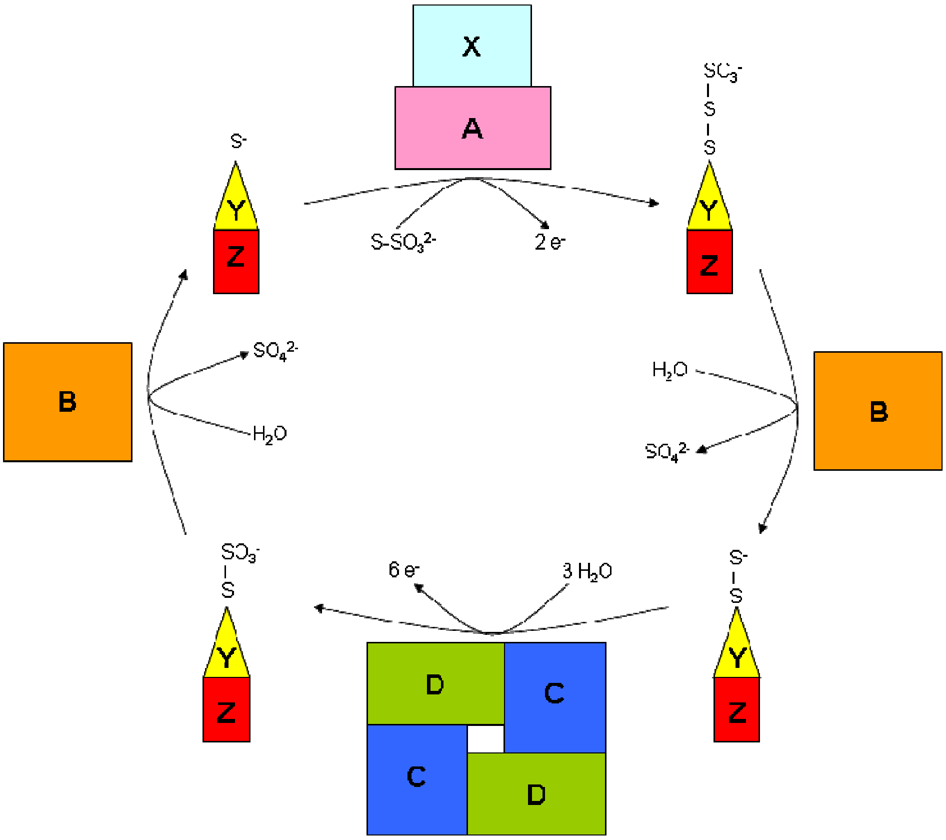


Figure 5.5 – The sox system of *Paracoccus pantotrophus*. The capital letters indicate the Sox proteins according to their gene designation. X and A, Cytochrome complex SoxAX; B, dimanganese protein SoxB; C and D, heterotetrameric molybdoprotein-cytochrome *c* complex Sox(CD)$_2$; Y and Z the heterodimeric complex that carries the intermediates between the other Sox proteins **[Friedrich *et al* 2001]**.

SoxF of *P.pantotrophus* is a 42,797 Da, monomeric, FAD containing periplasmic protein, closely related to the flavoprotein subunits of flavocytochromes from chemolithotrophic and phototrophic sulfur-oxidizing bacterium **[Quentmeier *et al* 2004]**. SoxF has sulfite-dehydrogenase activity. It has been found to be non-essential for the activity of the Sox pathway, however, knocking out the SoxF gene does result in a reduced rate of thiosulfate oxidation, this rate can be enhanced *in vivo* via the addition of SoxF to the proteins of the Sox system **[Bardischewsky et al 2006]**, suggesting that SoxF does influence the sox system in some way.

This chapter will present X-ray crystal structures for the native form of SoxF (SoxF-native) and the sulfite-inhibited (SoxF-SO$_3^{2-}$) and sulfur-inhibited (SoxF-S$_n$) forms. It will also show the results of the docking of a GGCGG pentapeptide that mimics the C-terminal of SoxY into the active site cavity of SoxF leading to a proposed mechanism for SoxF mediated reactivating of SoxYZ via the refolding of the SoxY C-terminus, thus characterising SoxF's influence on the sox system. A homology model for the small c-type cytochrome SoxE (presumed to act as the electron acceptor for SoxF) will also be shown, along with a model for a SoxEF complex.

5.2 Materials and Methods

5.2.1 SoxF crystallisation and structure determination

5.2.1.1 Crystallisation

N-terminal His-tagged SoxF was provided by Prof. Cornelius Friedrich (University of Dortmund). The concentration of the protein solution in 10 mM Tris pH 7.5, 1 mM MgSO4, 0.1 mM sodium thiosulfate was determined to be 5.5 mg ml^{-1} via the method of Bradford **[Bradford 1976]**. Initial crystallisation screening experiments were carried out at 4℃ and 10℃ using the screens of Jancarik & Kim **[Jancarik and Kim 1991]** and Cudney *et al*. **[Cudney *et al* 1994]** and the hanging drop vapour diffusion technique. Each experiment utilised a 3.5 µl drop containing equal volumes of concentrated protein solution and screen solution. Each hanging drop was equilibrated against a 700 µl reservoir of screen solution. Emerald green crystals of a plate morphology (Figure 5.6) appeared in drops grown at 4℃ after 7 days from a crystal growth solution containing; 0.1 M MES pH 6.5 and 12 % w/v PEG 20,000. Optimization around this condition revealed that the crystals grew reproducibly across a PEG 20,000 concentration gradient of 10-14 % but grew best under the original conditions.

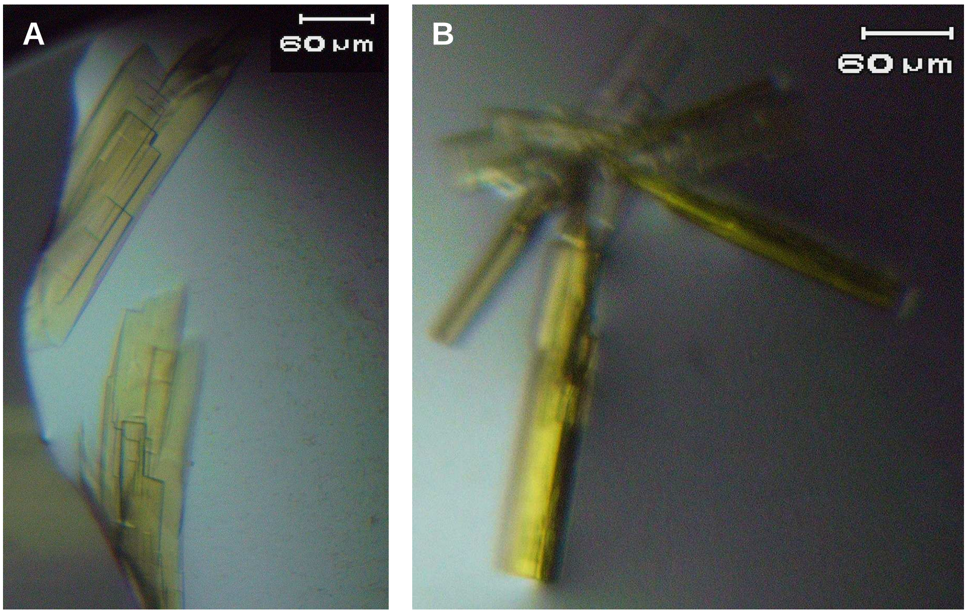


Figure 5.6 – Crystals of wild type SoxF protein, grown using the seeded sitting drop vapour diffusion technique. **(A)** Grown in 0.1 M sodium cacodylate pH 6.5, 30 % w/v PEG 8000. **(B)** Grown in 0.1 M MES pH 6.5, 12 % w/v PEG 20,000.

Wild-type SoxF was also provided by Prof. Cornelius Friedrich (University of Dortmund). The concentration of the protein solution in 10 mM Tris pH 7.5, 1 mM MgSO4, 0.1 mM sodium thiosulfate was determined to be 7.5 mg ml^{-1} via the method of Bradford **[Bradford 1976]**. Initial crystallisation screen experiments were carried out under the same conditions used with the His-tagged protein, but unfortunately, no crystals grew from these experiments.

Cross microseeding experiments were carried out with the wild type protein at 4℃ and 16℃ using the screens of Jancarik & Kim **[Jancarik and Kim 1991]** and Cudney *et al*. **[Cudney *et al* 1994]** and an Oryx Nano protein crystallisation robot (Douglas Instruments Ltd). A seed stock was created using crystals of His-tagged SoxF and a seed bead (Hampton Research). The desired seed crystal was removed from the drop and placed into a microcentrifuge tube containing 50 µl of crystal stabilising solution (0.1 M MES pH 6.5, 12 % w/v PEG 20,000) and the seed bead, before vortexing in 10 seconds intervals for a total 90 seconds, returning the solution to the ice after each vortex.

Sitting drop vapour diffusion seeding experiments were set up in 96-well plates using 0.2 µl of concentrated protein solution, 0.1 µl of seed stock and 0.2 µl of screen solution per drop. Each sitting drop was equilibrated against a 50 µl reservoir of screen solution. Emerald green crystals appeared within 6 days in drops grown at 16℃ in two different crystal growth conditions one in; 0.2 M sodium acetate, 0.1 M sodium cacodylate pH 6.5, 30 % w/v PEG 8000, and the other in the same conditions as the His-tagged protein (0.1 M MES pH 6.5, 12 % w/v PEG 20,000). These crystals are currently waiting for available synchrotron beamtime for data collection.

5.2.1.2 Crystal harvest and X-ray data collection

Crystal harvesting was carried out at 4℃. Crystal s were transferred to a cryoprotecting solution (0.1 M MES pH 6.5, 12 % w/v PEG 20,000, 30 % (w/v) ethylene glycol) and allowed to equilibrate for one minute. The crystals used for data collection were of plate morphology with dimensions ranging from 80-200 µm x 80-200 µm and were mounted in a free standing film using a LithoLoop (Molecular Dimensions Ltd) and cryocooled by immediate immersion into liquid nitrogen. In an attempt to solve structures of inhibited SoxF, a number of crystals were soaked in a solution containing the mother liquor and either 50 mM sodium metabisulfite or 1 mM sulphur in the form of polysulfide, created by dissolving sulfur in boiling sodium hydroxide, both of which have been shown to inhibit (in the case of sulfur, K_i = 1.3 µM) or inactivate (in the case of sulfite) SoxF activity **[Quentmeier *et al* 2004]**, before harvesting via transferring to a cryoprotecting solution (0.1 M MES pH 6.5, 12 % w/v PEG 20,000, 30 % (w/v) ethylene glycol) and cryocooling by immediate immersion into liquid nitrogen.

X-ray data was collected at the Diamond Light Source (beamline I02) using an ADSC Q315 CCD detector. Datasets were collected for the native and soaked crystals. The native dataset (SoxF-native) was taken at a wavelength of 0.979 Å, with a detector distance of 290.3 mm and an exposure time of 7.5 seconds per image, 180×1.0° oscillations about the goniometer Φ axis were recorded. The dataset taken from a sulfite-soaked crystal (SoxF-SO_3^-) was collected at a wavelength of 0.977 Å, with a detector distance of 374.2 mm and an exposure time of 0.348 seconds per image, 410×0.9° oscillations about the goniometer Φ axis were recorded. The dataset taken from a sulfur-soaked crystal (SoxF-S_n) was collected at a wavelength of 0.979 Å, with a detector distance of 373.4 mm and an exposure time of 1 second per image, 200×0.6° oscillations about the goniometer Φ axis were recorded.

5.2.1.3 Structure determination and refinement

Analysis of the three datasets collected from SoxF crystals processed with MOSFLM **[CCP4 1994]** was carried out with POINTLESS **[Evans 2005]**. This suggested the space group for each was C2. The datasets were subsequently scaled using SCALA **[CCP4 1994, Kabsch 1988]** and from these analyses the space group was confirmed to be C2 for each dataset. Molecular replacement was carried out with PHASER **[McCoy *et al* 2007]** using a modelled structure, created with MODELLER 9v4 **[Eswar *et al* 2006]**, based on the 2.5 Å resolution crystal structure of the homologous protein FccB, a sulfide dehydrogenase from *Allochromatium Vinosum* (PDB ID: 1FCD **[Chen *et al* 1994]**). Initial refinements and simulated annealing were carried out using PHENIX **[Adams *et al* 2010]**, and COOT **[Emsley and Cowtan 2004]** was used for map interpretation and remodelling of the structures, before final refinements and addition of water molecules using PHENIX.

The final structure of SoxF-native from this procedure, refined using data over the full resolution range (44.4-2.2 Å), had an R-factor of 21.9%, and an Rfree of 27.4%. The final structure of SoxF-SO_3^{2-}, refined using data over the full resolution range (50-2.8 Å), had an R-factor of 22.6 % and an Rfree of 30%. The final structure of SoxF-S_n, refined using data over the full resolution range (50-2.8 Å), had an R-factor of 24.1 % and an Rfree of 30 %. For full data collection and refinement parameters see Tables 5.1 and 5.2.

5.2.2 Statistics from data collections and structural refinement

Tables 5.1 and 5.2 summarise the data collection and refinement statistics for the native, sulfite soaked and sulfur soaked structures of the His-tagged SoxF protein.

Table 5.1 – Data collection statistics for each SoxF dataset

Dataset	SoxF - Native	SoxF – SO_3^{2-}	SoxF - S_n
Beamline	DLS I02	DLS I02	DLS I02
Space group	C2	C2	C2
Cell Parameters a , b , c (Å) α , β , γ (°)	 152.7, 76.2, 89.0 90, 121.1, 90	 152.4, 76.4, 88.5 90, 121.0, 90	 151.9, 75.9, 88.3 90, 121.0, 90
Wavelength (Å)	0.979	0.977	0.979
Resolution (Å)	44.4–2.2 (2.32 – 2.2)	50–2.8 (2.95 – 2.8)	50–2.8 (2.95-2.8)
Completeness (%)	99.5 (99.3)	99.7 (96.9)	96.3 (95.9)
R_{sym} (%)	13.6 (31.0)	19.9 (45.7)	10.3 (18.6)
<I/σI>	6.9 (3.7)	9.6 (2.5)	7.5 (3.4)
Independent reflections	44173 (6390)	21407 (3130)	20489 (2939)
Multiplicity	3.5 (3.6)	3.0 (3.1)	2.4 (2.5)
Overall temperature factor ($Å^2$)	19.3	33.6	32.5

Numbers in brackets represent data in the high resolution shell

Table 5.2 – Refinement statistics for each SoxF dataset

Dataset	SoxF – Native	SoxF – SO_3^-	SoxF – S_n
SoxF monomers per AU*	2	2	2
Refined structure Total atoms Water molecules	 6460 430	 6165 150	 6083 107
R_{cryst} (%)	21.9	22.6	24.1
R_{free} (%)	27	30	30
Ramachandran Analysis (%) Most favoured Additional allowed Generously allowed Disallowed	 86.2 13.1 0.5 0.3	 80.8 17.2 1.4 0.6	 79.8 18.0 1.8 0.3
RMS deviations Bonds (Å) Angles (°) Planes (Å)	 0.01 1.09 0.01	 0.01 1.52 0.01	 0.04 2.03 0.02
Mean Atomic B-value ($Å^2$)	13.3	11.5	11.6

*AU = Asymmetric unit

5.2.3 Modelling of the Di-heme subunit

A model for the di-heme subunit, SoxE, was generated by comparative structural modelling using MODELLER 9v4 **[Eswar *et al* 2006]**. The crystal structure of an oxidised recombinant cytochrome c_4 from *Pseudomonas stutzeri* (PDB ID: 1M70 **[Kadziola *et al* 1995]**), which has a 21.3 % sequence identity to SoxE, was used as a template. The cytochrome c_4 from *Pseudomonas stutzeri* was used as a template for modelling, rather than the di-heme subunit FccA of the flavocytochrome c sulfide dehydrogenase, because it had a greater sequence identity to SoxE. Additional distance restraints were placed on the heme binding cysteine and histidine ligands, as these have been discovered to be well conserved in hemoproteins (see Chapter 2), to improve the quality of the model

5.2.4 Modelling of a transient SoxEF encounter complex

A model for a transient SoxEF encounter complex between SoxE and SoxF was created using the FccAB complex from *Allochromatium vinosum* as a template. The homology modelled SoxE structure was superimposed onto the di-heme subunit FccA and the 2.2 Å resolution native SoxF crystal structure was superimposed into the flaviun binding subunit FccB. All superpositions were carried out using the alignment function of PyMOL **[DeLano 2002]**, which performs a Cα alignment.

5.2.5 Searches for homologous proteins

To search for proteins with structures that were homologous to SoxF a protein BLAST search **[Altschul *et al* 1990]** was carried out against the PDB database. Global sequence alignments and thus identities were calculated using the NEEDLE pairwise sequence alignment algorythm **[Needleman and Wunsch 1970]**.

A group of SoxF orthologues had been identified by Freidrich *et al* **[Freidrich *et al* 2005]**, the sequences of these proteins were aligned using MUSCLE **[Edgar 2004]**, and CONSURF **[Glaser *et al* 2003]** was used to produce a PDB file with the residues coloured in accordance with their conservation in the sequence alignment.

5.3 Results

5.3.1 The Crystal Structure of SoxF-native

The final model of the two SoxF monomers found in the asymmetric unit contain 786 amino acid residues from the primary sequence (393 per monomer), two flavin adenine dinucleotide (FAD) molecules that are bound covalently to the apoprotein by an 8-α-methyl(S-cysteinyl) thioether linkage with Cys43 and 430 water molecules (Figure 5.7). The protein is comprised of two domains, a FAD binding domain (Pfam ID: PF09242) found in the flavocytochrome c sulphide dehydrogenases and a FAD-dependent pyridine nucleotide-disulphide oxidoreductase domain (Pfam ID: PF07992) found in both class I and class II oxidoreductases and also NADH oxidases and peroxidases.

The principal secondary structure elements for each monomer as determined by the program STRIDE **[Frishman and Argos 1995]** are: nine α helices (Gly13-Arg23, Ser46-Gly51, Tyr63-Ala68, Pro133-Ala144, Pro164-Thr180, Gln198-Tyr209, Arg256-Ile260, Ala301-Leu319 and Ala371-Phe392), four 3_{10} helices (Phe56-Leu59, Pro111-Ser113, Leu119-Ala121 and Pro229-Ser231) and 25 β strands (Lys4-Ile8, Asp30-Val34, Val39-Thr41, Gln60-Gly62, Ala72-Val74, Ala78-Val81, Thr87-Leu90, Val95-Pro97, Arg100-Leu103, Ile107-Phe109, Val150-Val154, Lys185-Leu189, Val213-Ile216, Val225-Arg228, Glu233-Val236, Thr239-Lys242, Cys245-Val248, Gln252-Ala254, Ala270-Pro271, Lys278-Ser279, Asp282-Ile287, Ser292-Ala293, Tyr329-Ala338, Asp341-Asn352 and Arg355-Ile363). The (Φ, Ψ) torsion angles of all but one residue (Arg323 from monomer B) fall within the allowed regions of the Ramachandran plot.

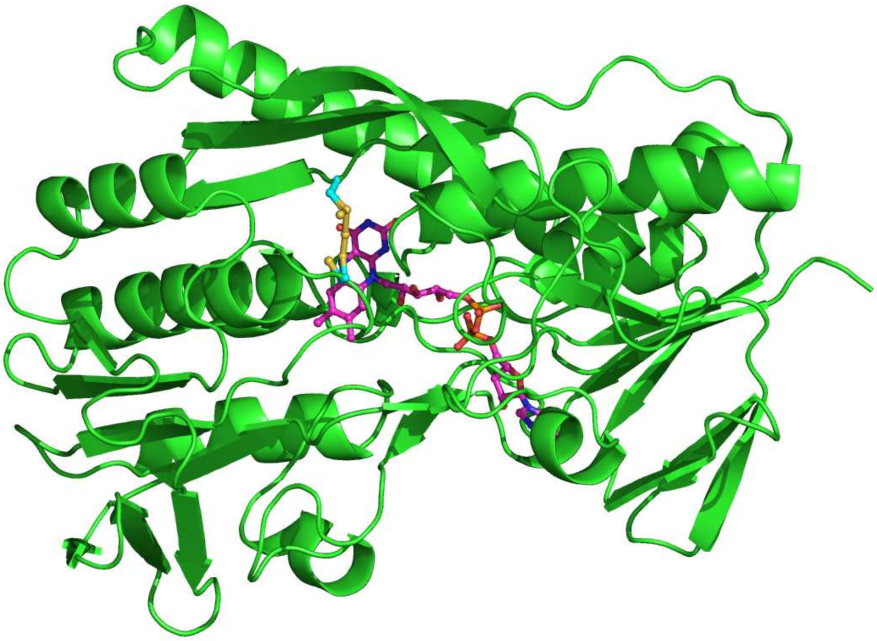


Figure 5.7 – Structure of SoxF displayed in cartoon format, the side chains of the active site residues (Cys162 and Cys333) and the FAD cofactor displayed as sticks and coloured cyan and magenta respectively.

The sequencing of the *P.denitrificans* SoxF suggested residues 162 and 333 (the active site residues) are both cysteine residues **[Wodara *et al* 1997]**, but closer inspection of the difference Fourier electron density maps has suggested that Cys333 has undergone a post translational modification to form a cysteine persulfide (cysteine residue with an extra sulfur attached to the SD sulfur atom). These active site residues also have alternate conformations, one of which results in the formation of a trisulfide bridge, the refined occupancies of the two conformers were 0.41 in favour of the bridged (oxidised) conformation and 0.59 in favour of the broken (reduced) conformation (Figure 5.8).

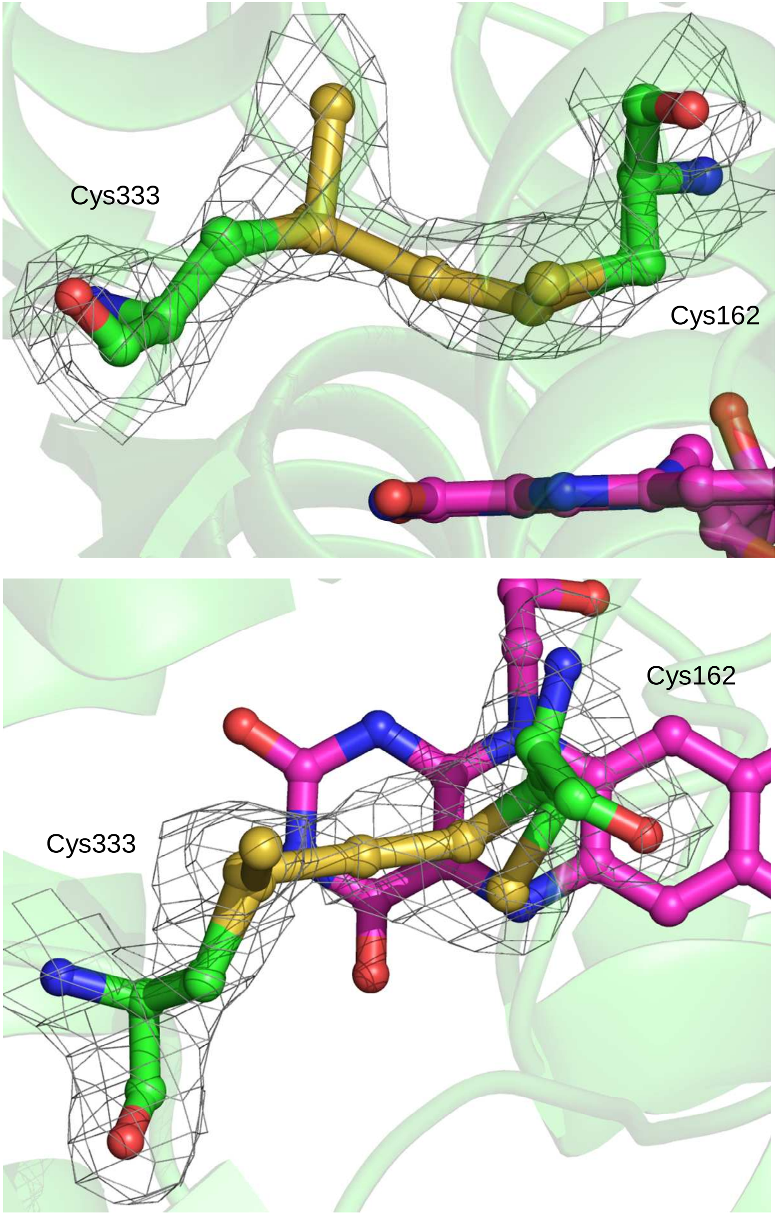


Figure 5.8 – Orthogonal views of the active site of SoxF-native, displaying the alternative conformations of the active site residues (Cys 162 and 333) and the double difference Fourier electron density map around them, at a contour level of 1.1 sigma. The protein backbone is displayed in cartoon format and the FAD cofactor coloured magenta.

Single difference Fourier electron density maps were also calculated for the SoxF-native dataset, with a cysteine residue at position 333 rather than a cysteine persulfide (Figure 5.9), to provide further evidence for this interpretation of the data, i.e. the presence of dual conformations for the active site residues that result in the formation of a trisulfide bridge via a cysteine persulfide at position 333. The regions of positive single difference density located between the cysteine two residues and above Cys333; where the middle sulfur of the trisulfide bridge and SD atom of the cysteine persulfide respectively would be located support this interpretation of the data. The regions of negative single difference density located over the SG atoms of the cysteine residues suggest their occupancies are too high, also supporting this interpretation of the data, since the dual conformers of the cysteine residues result in lower individual occupancies.

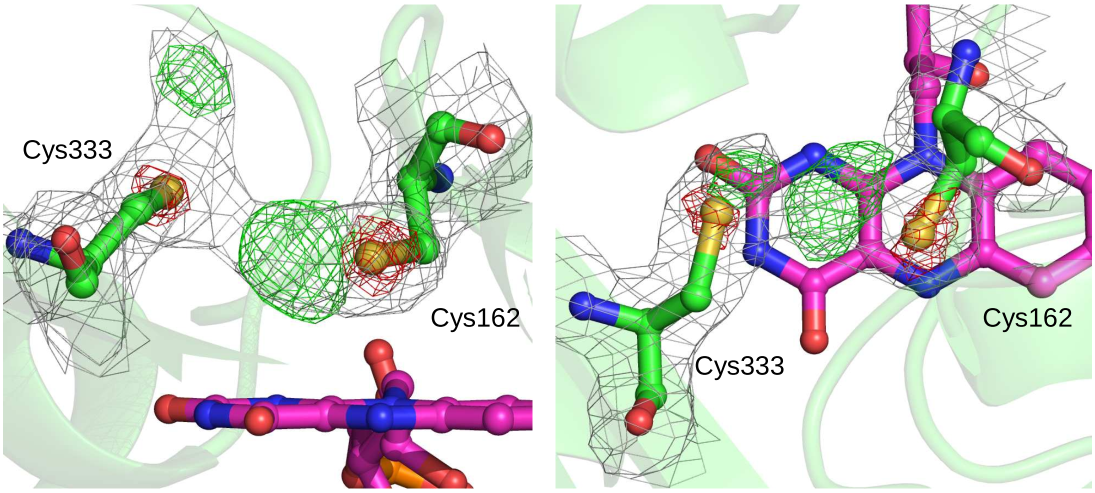


Figure 5.9 – Orthogonal views of the active site of SoxF-native, with a cysteine residue present as position 333 rather than a cysteine persulfide. The double difference Fourier electron density map at a contour level of 1.1 sigma is coloured Grey, the positive single difference Fourier electron density map at a contour level of 5 sigma is coloured is coloured green and the negative single difference Fourier electron density map at a contour level of -4 is coloured red.

5.3.2 Structures of SoxF-SO_3^{2-} and SoxF-S_n

The structures for SoxF-SO_3^{2-} and SoxF-S_n were superficially similar to that of the native structure, with RMSD's of 0.33 and 0.29 Å respectively after structural alignment, based on the coordinates of the Cα atoms. Both structures contained two SoxF monomers in the asymmetric unit, both contained all 786 amino acid residues from the primary sequence (393 per monomer) and two FAD molecules. The sulfite-soaked structure contained 182 water molecules and the sulfur-soaked structure 131. The secondary structure elements for each structure are the same as those found in the native structure. For the SoxF-SO_3^{2-} structure, the (Φ, Ψ) torsion angles of all but three residues (Arg83 from each monomer and Arg323 from the second monomer fall within the allowed regions of the Ramachandran plot. For the SoxF-S_n structure, the (Φ, Ψ) torsion angles of all but four residues (Arg83 and Arg323 from each monomer) fall within the allowed regions of the Ramachandran plot.

The geometries of the active sites for each structure are markedly different; both still appear to contain one cysteine and one post translationally modified cysteine residue, rather than two apo cysteine residues, but this is where the similarities end. The SoxF-SO_3^{2-} structure, like the native structure, contains alternate conformers for the residues of the active site, one of which forms a trisulfide bridge. However, unlike the native structure the modified cysteine residue, resulting in a cysteine persulfide, is found at position 162 rather than 333 and the alternate conformation for residue Cys333 is the sulfite bound form of cysteine, cysteine-s-sulfonate (Figure 5.10).

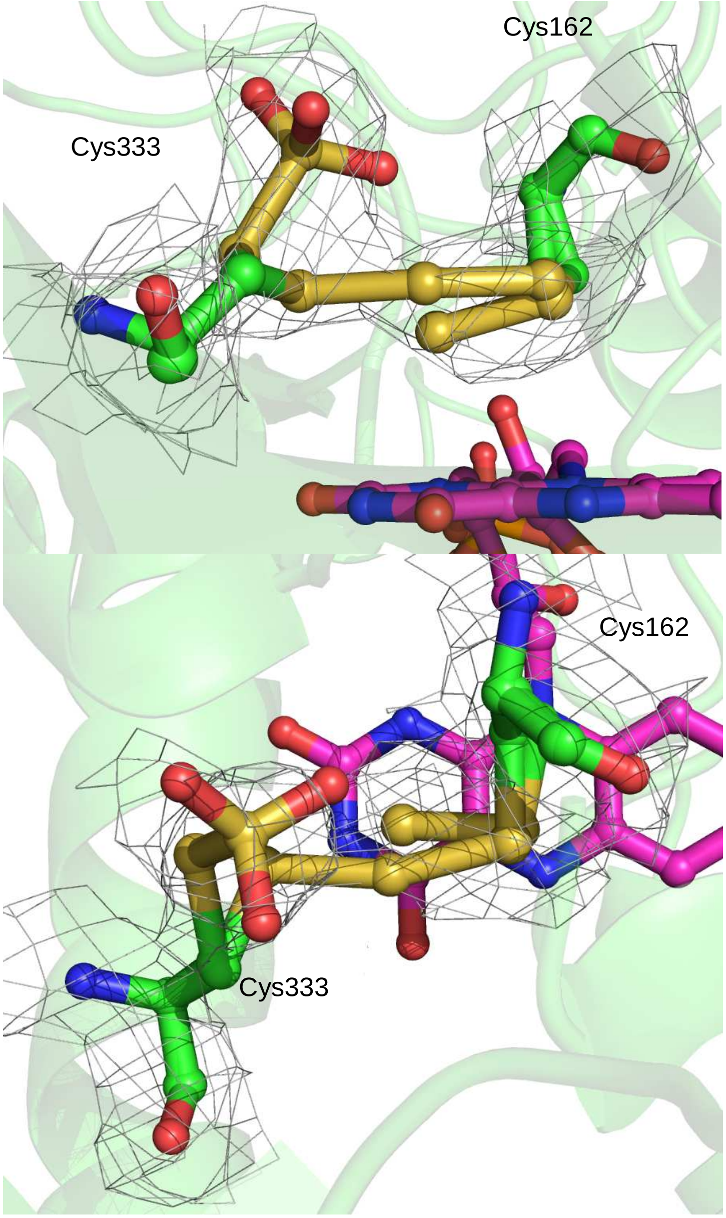


Figure 5.10 – Orthogonal views of the active site of SoxF-SO_3^{2-}, displaying the alternative conformations of the active site residues (Cys 162 and 333) and the double difference Fourier electron density map around them, at a contour level of 1.1 sigma. Cys162 has been modified to a cysteine persulfide and Cys333 to a cysteine-s-sulfonate. The protein backbone is displayed in cartoon format and the FAD cofactor coloured magenta.

A single difference Fourier electron density map was also calculated for the SoxF-SO_3^{2-} dataset, with a cysteine persulfide at position 333 rather than a cysteine-s-sulfonate (Figure 5.11), to provide further evidence for this interpretation of the data, i.e. the presence of the cysteine-s-sulfonate at position 333. The region of positive electron density over the SD atom of the cysteine persulfide at position 333 suggests there is more than just sulfur present at in this region, supporting this interpretation of the presence of a cysteine-s-sulfonate, as does the lack of any single difference density around the active site residues of the refined structure.

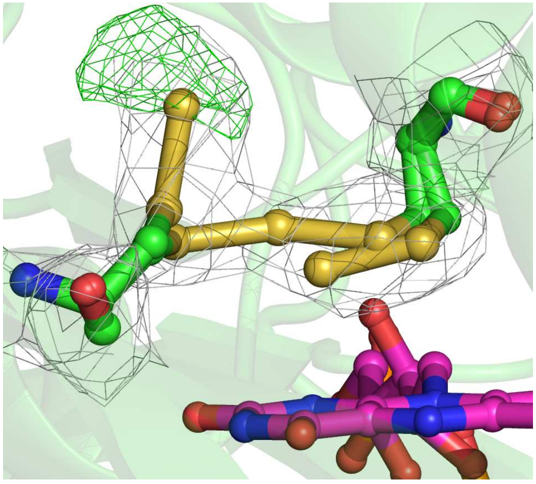


Figure 5.11 – The active site of SoxF-SO_3^{2-} with a cysteine persulfide at position 333, rather than a cysteine-s-sulfonate. The double difference Fourier electron density map at a contour level of 1.1 sigma is coloured Grey and the positive single difference Fourier electron density map at a contour level of 2.3 sigma is coloured is coloured green.

The SoxF-S_n structure, unlike the SoxF-native and SoxF-SO_3^{2-} structures, does not contain alternate conformers for the two active site cysteine residues. Furthermore, the electron density suggests there is no bridge formed between them. As was observed in the native structure, residue 162 is a cysteine and residue 333 has undergone a post translational modification. However, unlike the native structure residue 333 has been modified to a cysteine-s-trisulfane (a cysteine residue with a chain of three additional sulfur atoms attached to its SD sulfur atom) rather than a cysteine persulfide. (Figure 5.12)

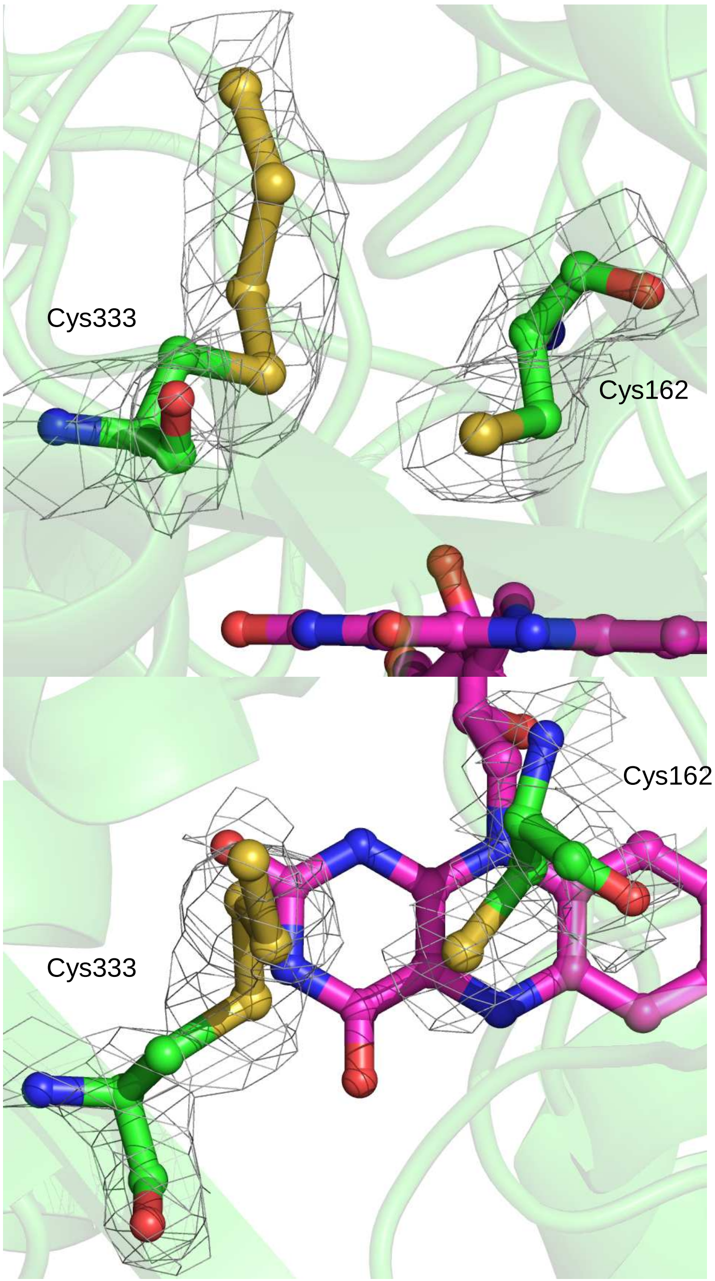


Figure 5.12 – Orthogonal views of the active site of SoxF-S_n, displaying the active site residues (Cys 162 and 333) and the double difference Fourier electron density map around them, at a contour level of 1.1 sigma. Cys333 has been modified to a cysteine-s-trisulfane. The protein backbone is displayed in cartoon format and the FAD cofactor coloured magenta.

5.3.3 Modelling a SoxF-SoxYZ complex

It has been speculated that SoxF is capable of binding the SoxYZ heterodimer, either as a single SoxYZ heterodimer or a SoxYZ heterotetramer involving two SoxYZ heterodimers. To test if two SoxYZ heterodimers could bind with SoxF a SoxYZ heterotetramer was docked into the SoxF active site cleft to analyse if there were any clear polypeptide clashes that would preclude complex formation. The dimeric structure of SoxY from *Chlorobium limicola f. thiosulfatophilum* (PDB ID: 2NNF **[Stout *et al* 2007]**) was used as a template for SoxYZ dimer packing. The results suggested that a heterotetramer could physically fit into the active site cleft as no clashes were found; however the active cysteine residues on the C-termini of the SoxY molecules could not approach close enough to the SoxF active site with two heterodimers present, making a single heterodimer more likely.

To give further insight to the nature of the SoxF-SoxYZ interaction a GGCGG pentapeptide structure that mimics the sequence found on the C-terminal of SoxY was docked into the active site, using a pathway predicted by CAVER **[Petrek *et al* 2006]** as a mould for docking. This was done to assess how many pentapeptide sequences could be docked into the active site and the specific binding interactions that could take place. A number of pentapeptide structures were generated based on secondary structure elements and fit to the caver path, with the best shown in Figure 5.13.

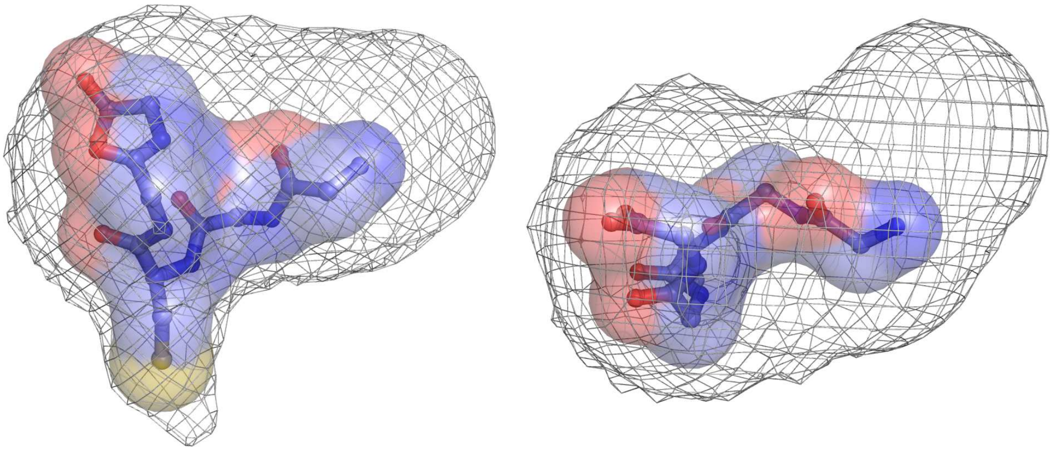


Figure 5.13 – Orthogonal views of the GGCGG pentapeptide sequence that replicates the sequence found on the C-terminal of SoxY (displayed as sticks with a transparent protein surface) and how it fits within the CAVER pathway (displayed as a mesh). Demonstrating how only one pentapeptide is able to fit into the active site pathway.

Figure 5.14 shows how the pentapeptide fits into the binding cavity in relation to the SoxF structure. The SG atom on the pentapeptide cysteine approaches to within 2.1 Å of the cysteine persulfide SD atom on Cys333 of SoxF. Also shown are the residues that line the binding pocket of SoxF; Asn158, Pro163, Pro164, Lys194, Ser196, Asp296, Pro298, Leu345 and Ile363.

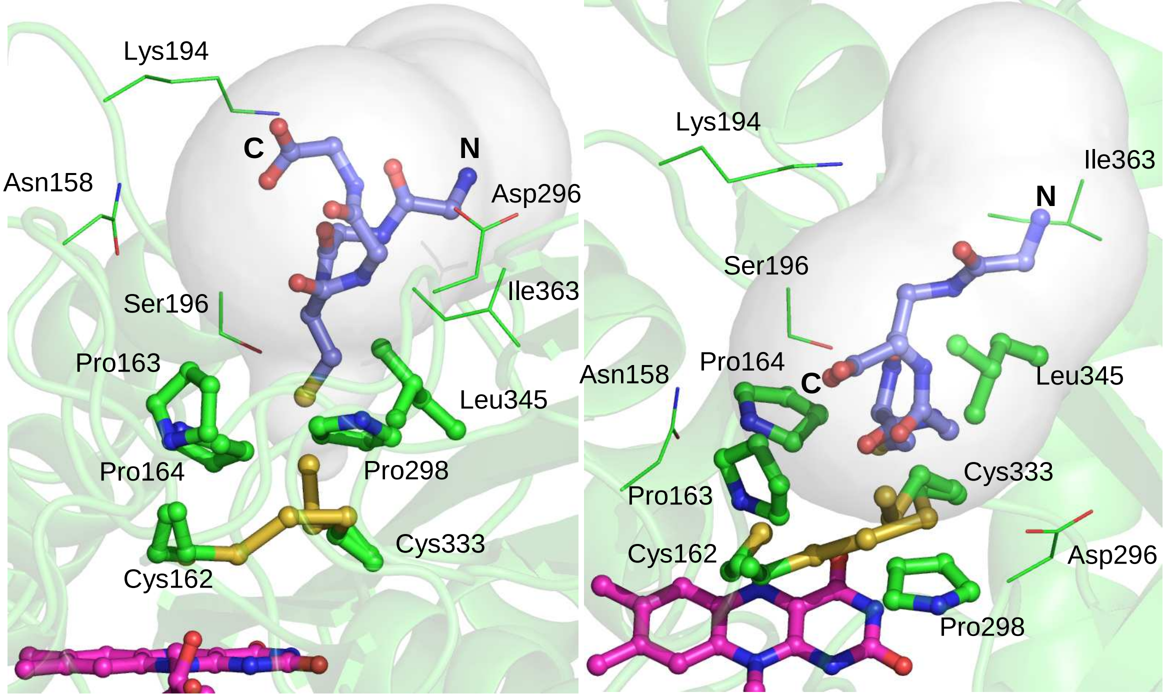


Figure 5.14 – Orthogonal views of the GGCGG pentapeptide (with its N and C termini labelled) and how it could fit into the SoxF binding cavity. The cysteine SG sulfur on the pentapeptide approaches to a distance of 2.1Å from the cysteine persulfide SD sulfur on Cys333 of SoxF. The side chains of the active site cysteine residues and the residues that form the "plug hole" are displayed as sticks, the other side chains of the binding pocket residues are labelled and displayed as lines, and the FAD cofactor coloured magenta.

All these results points towards a single SoxYZ heterodimer interacting with SoxF rather than a heterotetramer, since there was not enough space in the predicted active site cavity to accommodate more than one pentapeptide. Further evidence can be seen in the positioning of the three proline residues (Pro 162, 164 and 298, Figure 5.14) that appear to form a rigid "plug hole" around the reactive persulfide residue, providing enough space for a single thiol group from the approaching SoxY cysteine, but not enough for a second.

5.3.4 How SoxF compares with its homologues

Analysis of the SoxF sequence with CONSURF **[Glaser *et al* 2003]**, utilising 10 homologous sequences as identified by Freidrich *et al* **[Freidrich *et al* 2005]**, indicated conserved regions in the sequences. There was strong sequence conservation in the vicinity of the active site (Figure 5.15A) and an area of strong sequence conservation at the SoxE binding face (Figure 5.15B).

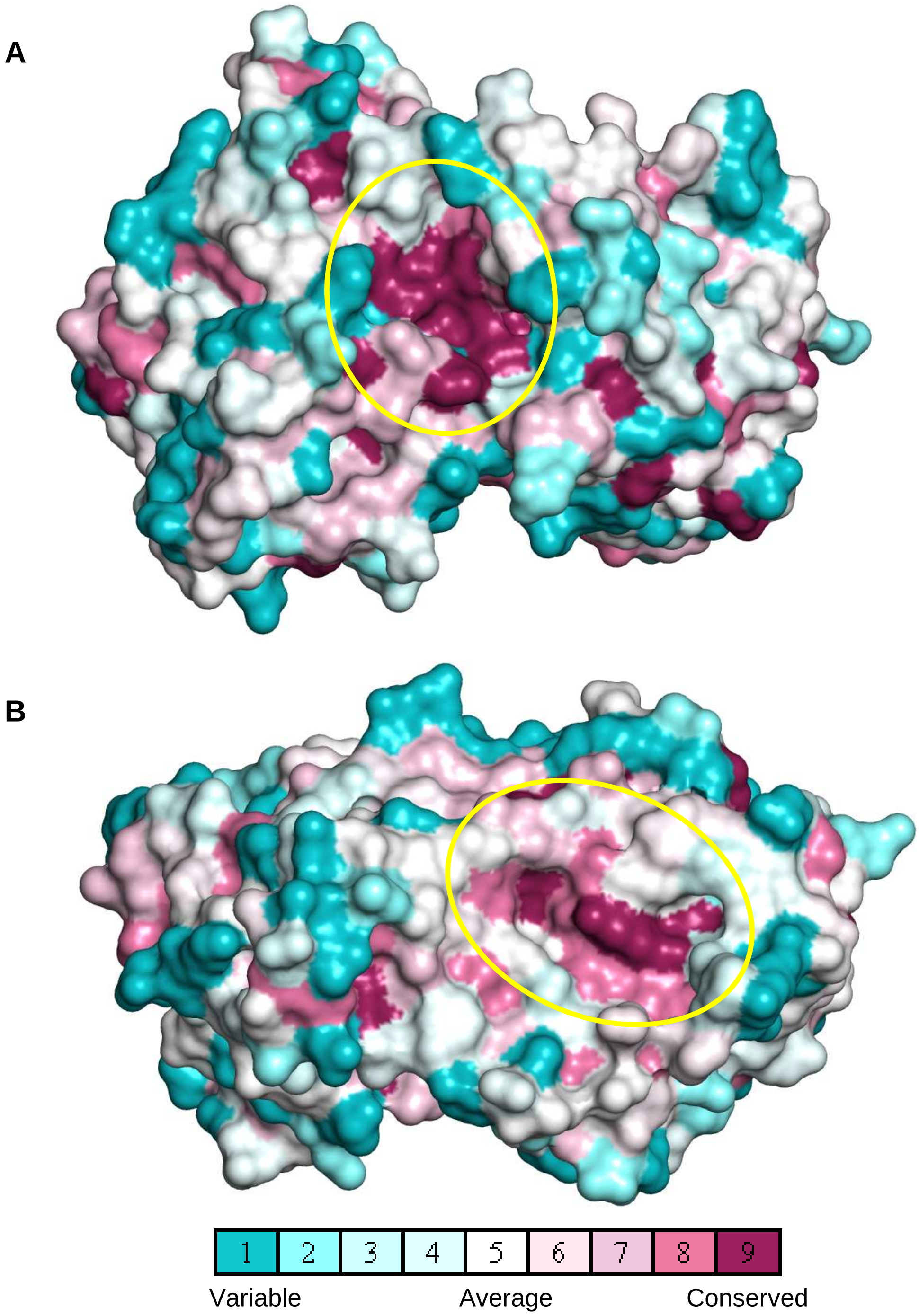


Figure 5.15 – The protein surface of SoxF as seen from **(A)** above the active site at the SoxY binding face and **(B)** the SoxE binding face, displayed in the CONSURF colouration (key above figure legend). The conservation of sequence around the active site and SoxE binding region (both circled) can be seen.

A multiple sequence alignment of these SoxF orthologues (Figure 5.16) indicates conservation of the active site cysteines, the three proline residues of the rigid “plug hole” surrounding the reactive persulfide residue and the serine of the active lining the active site cavity, and conservative amino acid substitutions of the “plug hole” forming leucine and other residues lining the active site cavity.

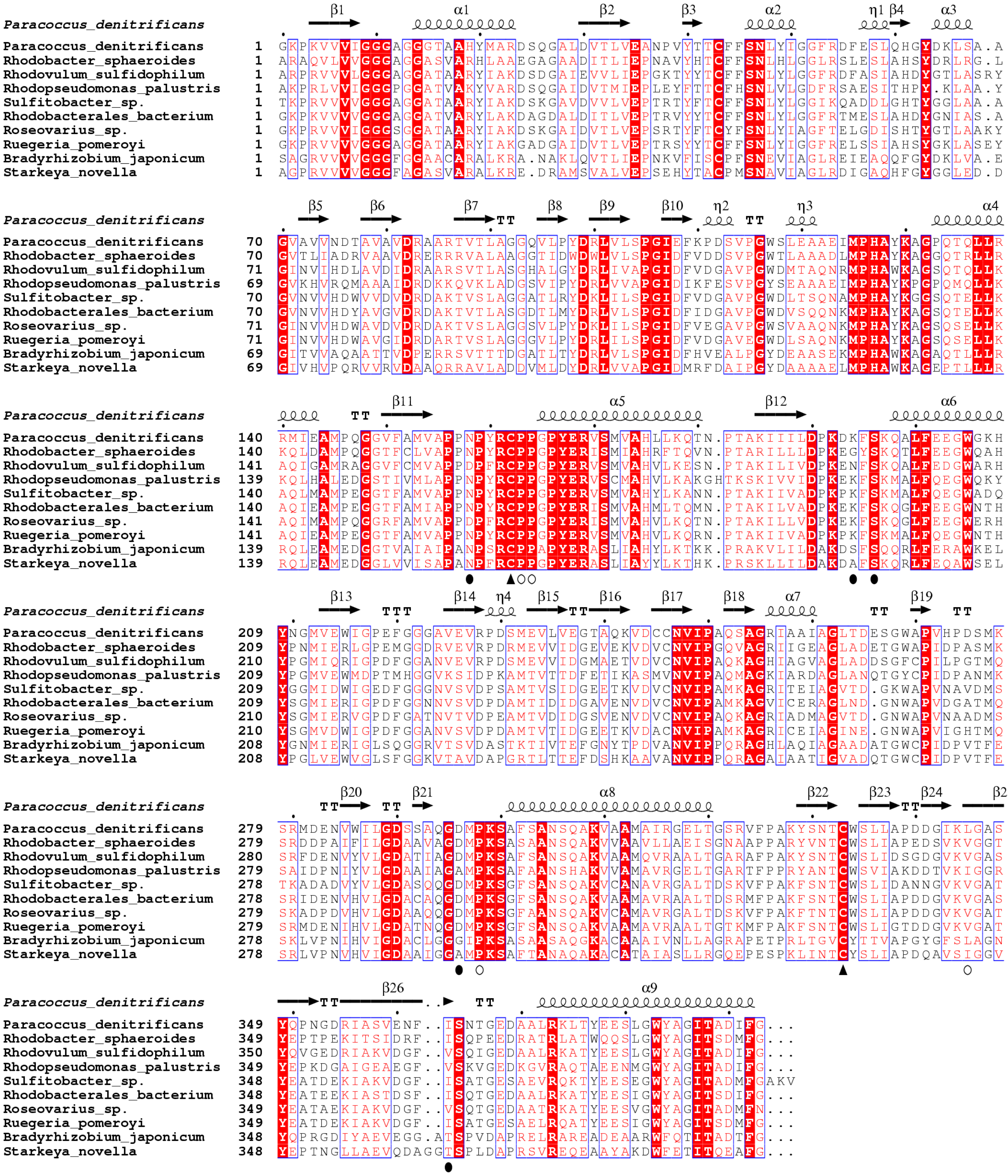


Figure 5.16 – Multiple sequence alignment of SoxF orthologues. Active site cysteines (▲), residues of the active site cavity that form the "plug hole" around the reactive cysteine 333 (○) and the other residues surrounding the active site cavity (●) are all marked. The secondary structure annotation is from the 2.2 Å resolution SoxF-native structure. Identical residues are indicated by a red background, conservatively varied residues are boxed in blue and shown in red characters.

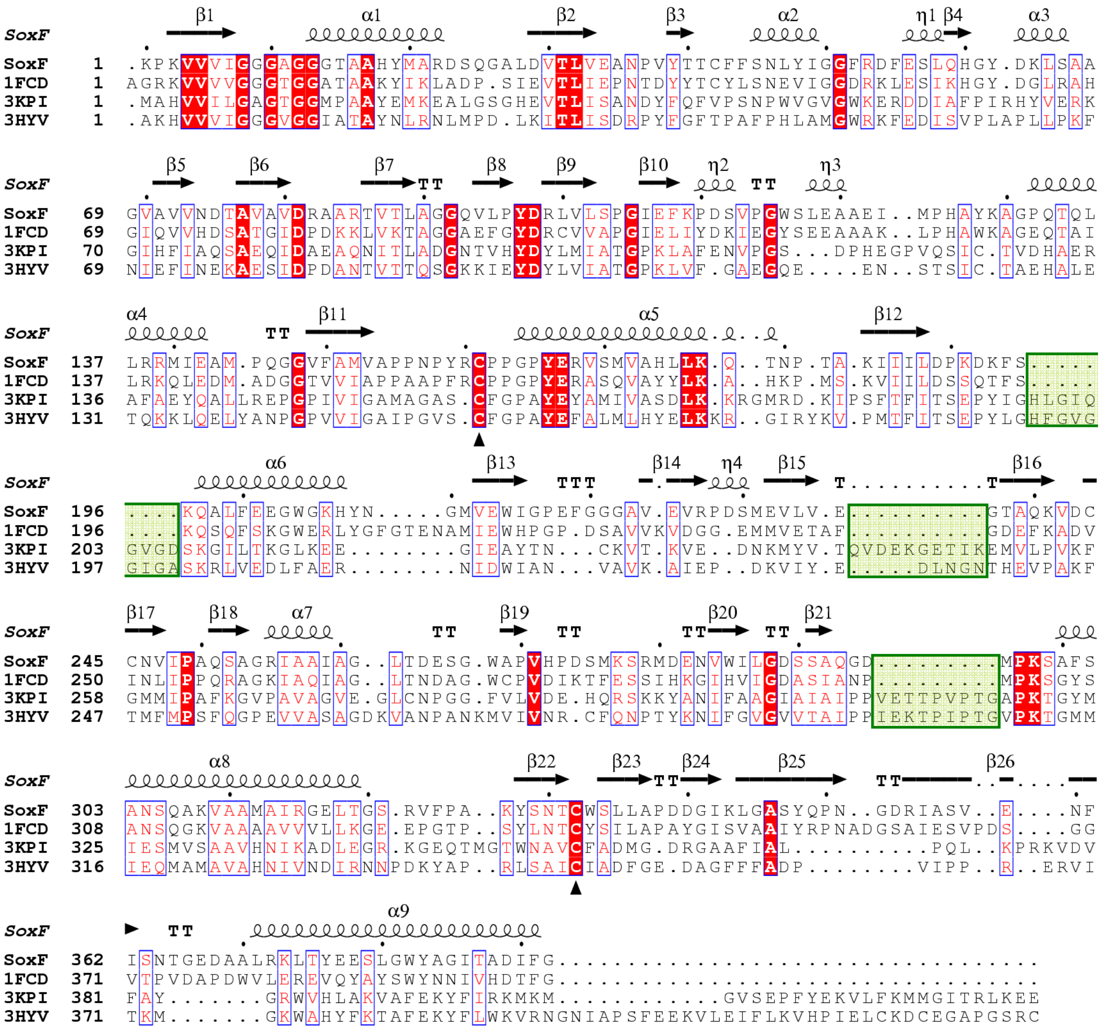
Four SoxF homologues with solved structures were identified by a BLAST search of the PDB, a breakdown of which can be seen in Table 5.3. All structures used are native protein structures, i.e. the crystals had not undergone any soaking procedures prior to data collection.

Table 5.3 – A breakdown of structures homologous to SoxF

PDB ID	Description	Organism	Sequence identity to SoxF (%)	RMSD value for alignment with SoxF (Å)
1FCD	Flavocytochrome c sulfide dehydrogenase	*Allochromatium vinosum*	42.4	0.76
3KPI	sulfide:quinone oxidoreductase	*Acidithiobacillus ferrooxidans*	19.0	2.46
3HYV	sulfide:quinone oxidoreductase	*Aquifex aeolicus*	19.2	3.56

A multiple sequence alignment of these structural homologues (Figure 5.17) identified conserved catalytic cysteine residues in each sequence and insertions in the sulfide:quinone oxidoreductases (SQRs) relative to SoxF and the flavocytochrome c sulfide dehydrogenase that form the “capping loop” identified by Marcia *et al* that guarantees exclusive access of sulfite in the SQRs **[Marcia *et al* 2009]**.

A closer inspection of the active sites of SoxF and its structural homologues reveals differences in the stoichiometries of the active site cysteine. As has been previously stated, the active site of SoxF-native is comprised of a cysteine (Cys162) and a post translationally modified cysteine to a cysteine persulfide (Cys333) that have different conformers, one of which results in the formation of a trisulfide bridge (Figure 5.18A). The flavocytochrome c sulfide dehydrogenase (PDB ID: 1FCD **[Chen *et al* 1994]**) contains a single conformation of cysteine residues in a disulfide bridge (Figure 5.18B). The active site cysteines of the SQR’s both contain alternate conformers. Like SoxF (PDB ID: 3KPI **[Cherney *et al* 2010]**) contains a cysteine and cysteine persulfide, although the persulfide is found on the opposite cysteine (Cys160), the alternate conformation for this active site is a branched polysulfide bridge containing six sulfur atoms (Figure 5.18C). The other SQR (PDB ID: 3HYV **[Marcia *et al* 2009]**) also contains a cysteine and cysteine persulfide, with the persulfide on the opposite cysteine (Cys156) to SoxF, the alternate conformation for this active site is that it has an eight sulfur ring attached to Cys156 (Figure 5.18D). These findings suggest that it is the N-terminal cysteines (Cys160 and Cys156) that are the reactive active site residues in the SQRs, as opposed to the C-terminal cysteine (Cys333) in SoxF.



```
SoxF    1  .KPKVVVIGGGAGGGTAAHYMARDSQGALDVTLVEANPVYTTCFFSNLYIGGFRDFESLQHGY.DKLSAA
1FCD    1  AGRKVVVVGGGTGGATAAKYIKLADP.SIEVTLIEPNTDYYTCYLSNEVIGGDRKLESIKHGY.DGLRAH
3KPI    1  .MAHVVILGAGTGGMPAAYEMKEALGSGHEVTLISANDYFQFVPSNPWVGVGWKERDDIAFPIRHYVERK
3HYV    1  .AKHVVVIGGGVGGIATAYNLRNLMPD.LKITLISDRPYFGFTPAFPHLAMGWRKFEDISVPLAPLLPKF

SoxF   69  GVAVVNDTAVAVDRAARTVTLAGGQVLPYDRLVLSPGIEFKPDSVPGWSLEAAEI..MPHAYKAGPQTQL
1FCD   69  GIQVVHDSATGIDPDKKLVKTAGGAEFGYDRCVVAPGIELIYDKIEGYSEEAAAK..LPHAWKAGEQTAI
3KPI   70  GIHFIAQSAEQIDAEAQNITLADGNTVHYDYLMIATGPKLAFENVPGS...DPHEGPVQSIC.TVDHAER
3HYV   69  NIEFINEKAESIDPDANTVTTQSGKKIEYDYLVIATGPKLVF.GAEGQE....EN..STSIC.TAEHALE

SoxF  137  LRRMIEAM.PQGGVFAMVAPPNPYRCPPGPYERVSMVAHLLK.Q..TNP.TA.KIIILDPKDKFS.....
1FCD  137  LRKQLEDM.ADGGTVVIAPPAAPFRCPPGPYERASQVAYYLK.A..HKP.MS.KVIILDSSQTFS.....
3KPI  136  AFAEYQALLREPGPIVIGAMAGAS.CFGPAYEYAMIVASDLK.KRGMRD.KIPSFTFITSEPYIGHLGIQ
3HYV  131  TQKKLQELYANPGPVVIGAIPGVS.CFGPAYEFALMLHYELKKR..GIRYKV.PMTFITSEPYLGHFGVG

SoxF  196  ....KQALFEEGWGKHYN.....GMVEWIGPEFGGGAV.EVRPDSMEVLV.E..........GTAQKVDC
1FCD  196  ....KQSQFSKGWERLYGFGTENAMIEWHPGP.DSAVVKVDGG.EMMVETAF..........GDEFKADV
3KPI  203  GVGDSKGILTKGLKEE........GIEAYTN...CKVT.KVE..DNKMYV.TQVDEKGETIKEMVLPVKF
3HYV  197  GIGASKRLVEDLFAER........NIDWIAN...VAVK.AIEP..DKVIY.E.....DLNGNTHEVPAKF

SoxF  245  CNVIPAQSAGRIAAIAG..LTDESG.WAPVHPDSMKSRMDENVWILGDSSAQGD.........MPKSAFS
1FCD  250  INLIPPQRAGKIAQIAG..LTNDAG.WCPVDIKTFESSIHKGIHVIGDASIANP.........MPKSGYS
3KPI  258  GMMIPAFKGVPAVAGVE.GLCNPGG.FVLVDE.HQRSKKYANIFAAGIAIAIPPVETTPVPTGAPKTGYM
3HYV  247  TMFMPSFQGPEVVASAGDKVANPANKMVIVNR.CFQNPTYKNIFGVGVVTAIPPIEKTPIPTGVPKTGMM

SoxF  303  ANSQAKVAAMAIRGELTGS.RVFPA..KYSNTCWSLLAPDDGIKLGASYQPN..GDRIASV..E....NF
1FCD  308  ANSQGKVAAAAVVVLLKGE.EPGTP..SYLNTCYSILAPAYGISVAAIYRPNADGSAIESVPDS....GG
3KPI  325  IESMVSAAVHNIKADLEGR.KGEQTMGTWNAVCFADMG.DRGAAFIAL..........PQL..KPRKVDV
3HYV  316  IEQMAMAVAHNIVNDIRNNPDKYAP..RLSAICIADFGE.DAGFFFADP........VIPP..R..ERVI

SoxF  362  ISNTGEDAALRKLTYEESLGWYAGITADIFG...............................
1FCD  371  VTPVDAPDWVLEREVQYAYSWYNNIVHDTFG...............................
3KPI  381  FAY.......GRWVHLAKVAFEKYFIRKMKM............GVSEPFYEKVLFKMMGITRLKEE
3HYV  371  TKM.......GKWAHYFKTAFEKYFLWKVRNGNIAPSFEEKVLEIFLKVHPIELCKDCEGAPGSRC
```

Figure 5.17 – Dali **[Holm and Rosenström 2010]** alignment of SoxF homologues identified by BLAST search of the PDB. 1FCD is a flavocytochrome c sulfide dehydrogenase from *Chromatium vinosum*, 3KPI and 3HYV are sulfide:quinone oxidoreductases from *Acidithiobacillus ferrooxidans* and *Aquifex aeolicus* respectively. The active site cysteines are marked with a ▲ and the insertions in the SQRs that correspond to the "capping loops" are highlighted green. The secondary structure annotation is from the 2.2 Å resolution native SoxF structure. Identical residues are indicated by a red background, conservatively varied residues are boxed in blue and shown in red characters.

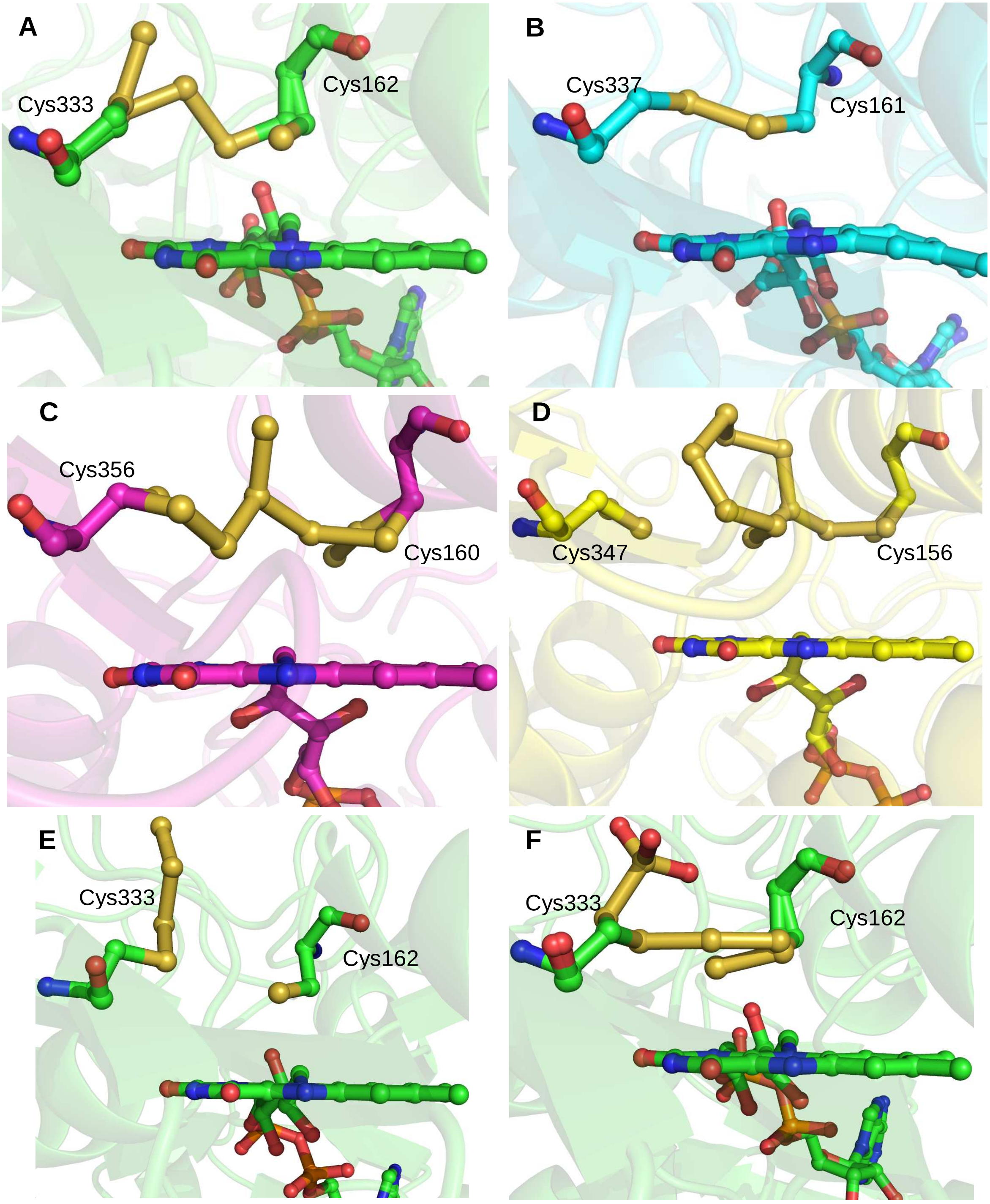


Figure 5.18 – The active sites of *P. Pantotrophus* SoxF and its homologues, showing the alternate cysteine and sulfur geometries in; **(A)** SoxF-native, **(B)** *Chromatium vinosum* flavocytochrome c sulfide dehydrogenase (1FCD), **(C)** *Acidithiobacillus ferrooxidans* sulfide:quinone oxidoreductase (3KPI) and **(D)** *Aquifex aeolicus* sulfide:quinone oxidoreductase (3HYV). All structures are of the native form of each protein. For comparison the active sites of **(E)** the SoxF-S_n and **(F)** SoxF-SO_3^{2-} structures are also shown.

The addition of the capping loops in the SQR structures results in a clear differences in the pathways between SoxF and the SQRs (Figure 5.19).

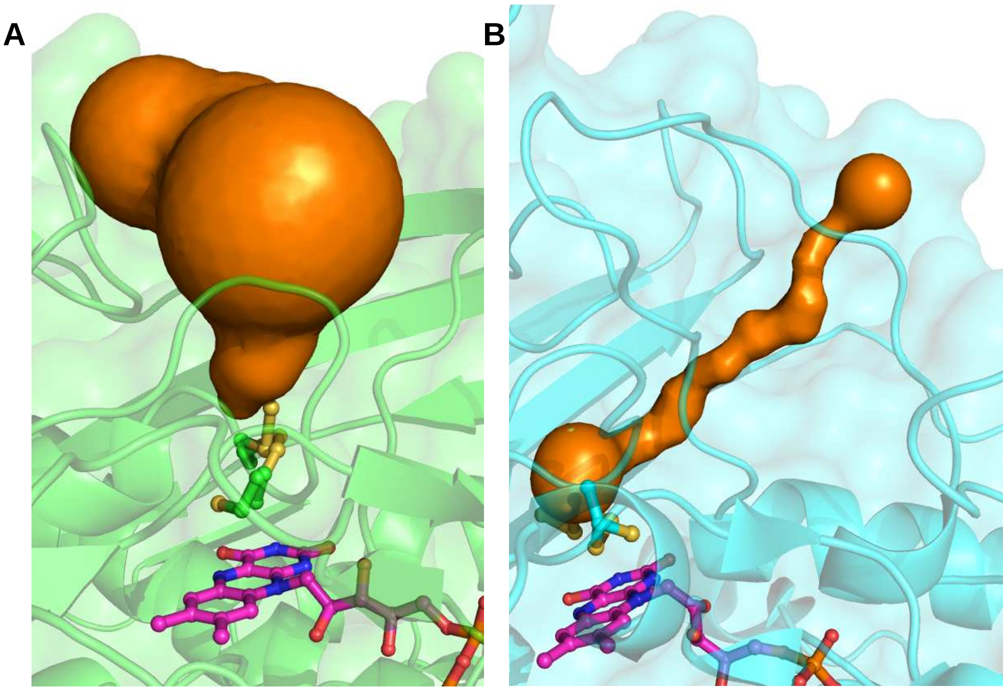
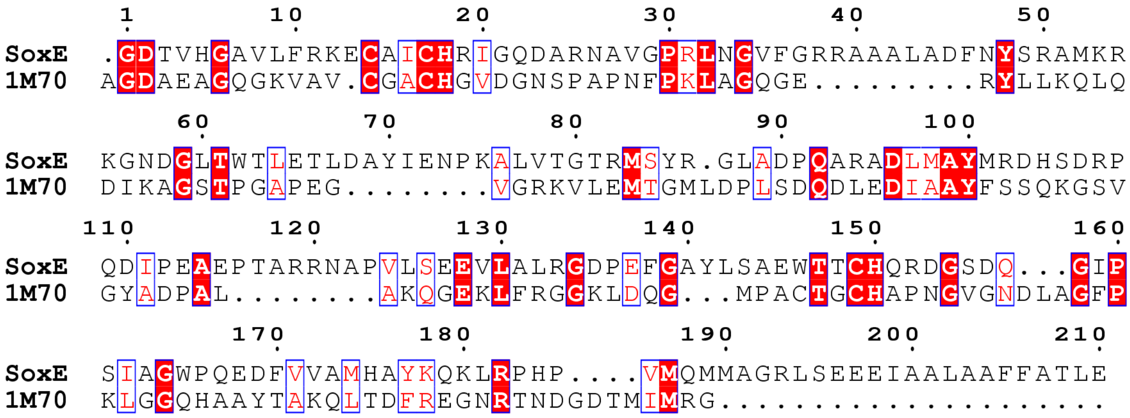


Figure 5.19 – CAVER pathways calculated for **(A)** *P.pantotrophus* SoxF and **(B)** *Acidithiobacillus ferrooxidans* sulfide:quinone oxidoreductase, demonstrating the significant differences in active site cavities in the two structures. The proteins are displayed in cartoon format with a transparent surface representation also shown, active site residues are displayed as sticks, FAD cofactors coloured magenta and active site cavities coloured orange.

5.3.5 A Homology Model for the small c-type cytochrome SoxE, presumed subunit of a transient SoxEF encounter complex

SoxF is thought to interact with a c-type cytochrome subunit SoxE **[Friedrich *et al* 2000, Quentmeier *et al* 2004]**, presently no crystal structure exists for this protein, so MODELLER was used to model the structure of SoxE using the structure of a homologous protein, cytochrome c_4 from *Pseudomonas stutzeri* (PDB ID: 1M70 **[Kadziola *et al* 1995]**), as a template. Figure 5.20 shows a sequence alignment between SoxE and 1M70.



```
            1         10        20        30        40        50
SoxE  .GDTVHGAVLFRKECAICHRIGQDARNAVGPRLNGVFGRRAAALADFNYSRAMKR
1M70  AGDAEAGQGKVAV.CGACHGVDGNSPAPNFPKLAGQGE.........RYLLKQLQ

               60        70        80         90      100
SoxE  KGNDGLTWTLETLDAYIENPKALVTGTRMSYR.GLADPQARADLMAYMRDHSDRP
1M70  DIKAGSTPGAPEG........VGRKVLEMTGMLDPLSDQDLEDIAAYFSSQKGSV

     110       120       130       140       150         160
SoxE  QDIPEAEPTARRNAPVLSEEVLALRGDPEFGAYLSAEWTTCHQRDGSDQ...GIP
1M70  GYADPAL........AKQGEKLFRGGKLDQG...MPACTGCHAPNGVGNDLAGFP

                 170       180           190       200       210
SoxE  SIAGWPQEDFVVAMHAYKQKLRPHP....VMQMMAGRLSEEEIAALAAFFATLE
1M70  KLGGQHAAYTAKQLTDFREGNRTNDGDTMIMRG....................
```

Figure 5.20 – Sequence alignment between the target sequence, SoxE, and the sequence for the template structure, 1M70. Identical residues are indicated by a red background, conservatively varied residues are boxed in blue and shown in red characters.

The final structure contained 210 of the 213 residues from the primary sequence and two heme groups (Figure 5.21). The principal secondary structure elements as determined by the program STRIDE **[Frishman and Argos 1995]** are: eight α helices (Cys14-His18, Tyr48-Leu60, Pro91-Asp103, Leu125-Leu132, Trp146-His150, Gln167-Gln179, Val186-Gly192 and Glu196-Thr208) and one 3_{10} helix (Thr81-Met83). The (Φ, Ψ) torsion angles of all but four residues (Arg19, Lys75, Ala118 and Gln151) fall within the allowed regions of the Ramachandran plot.

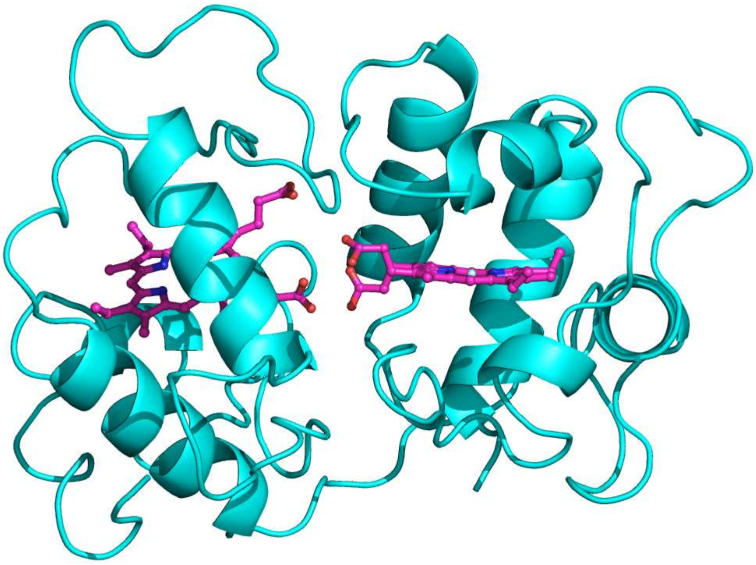


Figure 5.21 – Homology modelled structure for the cytochrome c subunit SoxE, presumed electron acceptor for SoxF. The protein is displayed in cartoon format, with heme cofactors coloured magenta.

Both of the hemes were found to be His-Met ligated and interestingly the second appears to have a highly unusual WXXCH heme binding motif (Figure 5.22), with the tryptophan residue presumably providing some form of heme stabilisation to compensate for the loss of the cysteine residue.

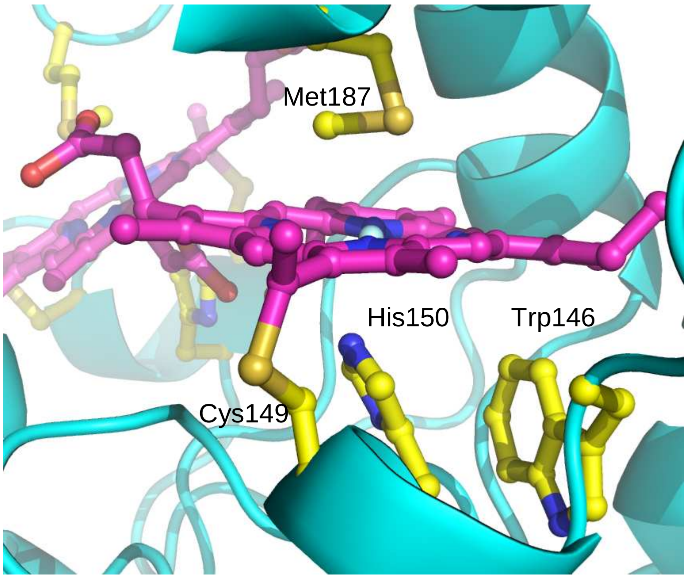


Figure 5.22 – The unusual WXXCH heme coordination proposed for the second heme of SoxE. The protein is displayed in cartoon format; the heme coordinating residues are displayed as yellow sticks and the heme cofactors as magenta sticks

5.3.6 Modelling of the presumed transient SoxE-SoxF complex

SoxE is thought to associate with SoxF **[Wodara *et al* 1997]**, it is thought this interaction occurs in a way homologous to that observed in FccAB of *Allochromatium vinosum* **[Chen *et al* 1994]**, although the heterodimer formed between SoxE and SoxF in *P.denitrificans* is thought not to be as tight as that formed between FccA and FccB in *Allochromatium vinosum* since they have not been successfully purified together. To model the *P.denitrificans* heterodimer, the model for SoxE and crystal structure of the 2.2 Å resolution SoxF were superimposed onto the di-heme cytochrome and flavin binding subunits of the *Allochromatium Vinosum* structure (Figure 5.23), with RMSDs of 0.73Å and 0.76Å, respectively.

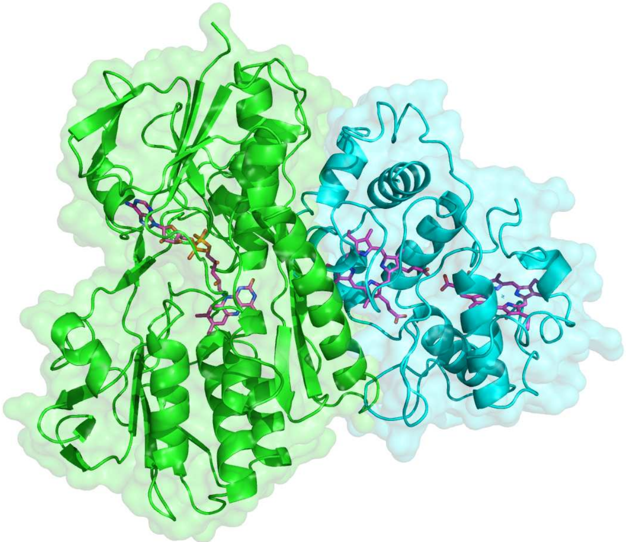


Figure 5.23 – A model for the transient SoxE-SoxF encounter complex from *P.denitrificans.* SoxE is coloured cyan, SoxF green and the cofactors magenta.

There were two main chain-main chain clashes between the two models on the dimer interface. This is likely to be due to the SoxE structure being a homology model rather than a crystal structure, and in fact the two main chain clashes between the models occur in loop regions of the SoxE structure where there is an insertion relative to the template structure used to build the model.

A potential path for electron transfer between the two subunits can be seen between the FAD of SoxF and interface facing heme of SoxE (Figure 5.24). Two tryptophan residues (Trp 334 and 383) are involved in this pathway, although it is unclear whether the electrons are passed from Trp383 to Trp334 during transfer, or if these residues are present to provide an appropriate electron tunnelling environment since the distance between the FAD and heme cofactors of 12.2 Ǻ is short enough to allow direct electron transfer.

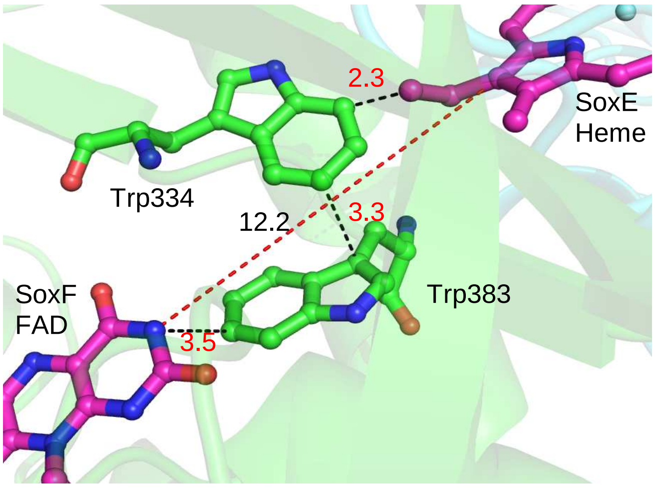


Figure 5.24 – Potential electron transfer path between SoxE and SoxF via tryptophans 334 and 383. Electron transfer distances are shown are in Ångstroms.

This predicted electron transfer pathway fits with the caver prediction for the cavity leading from the active site to the protein surface (Figure 5.25), since they do not occupy the same space. The channel itself fits with the model for the protein complex as a whole, as it does not interfere with the SoxF-SoxE interface.

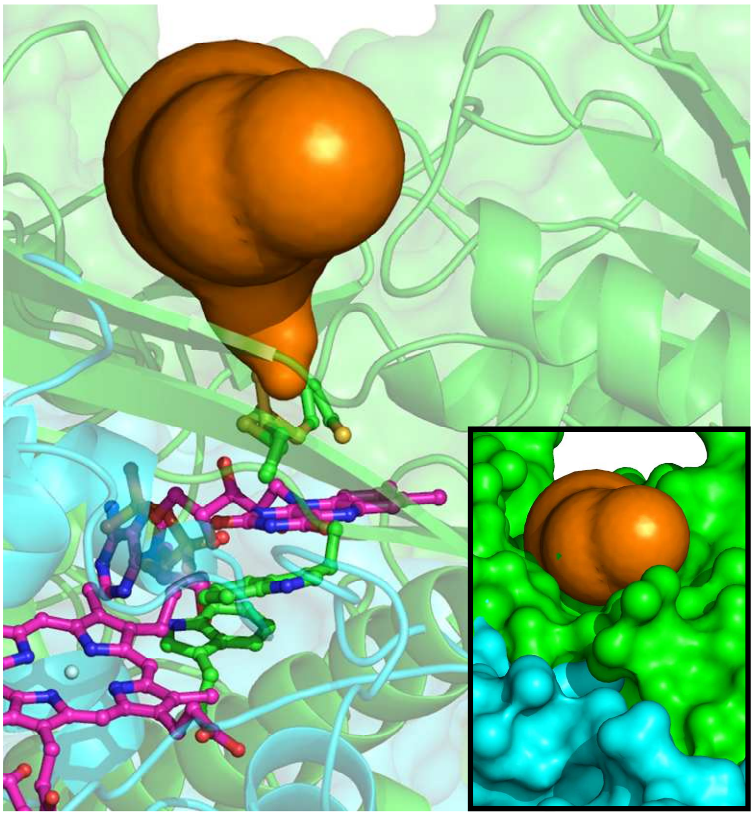


Figure 5.25 – The active site cavity as predicted by CAVER **[Petr *et al* 2008]**. Active site cysteines and residues along the electron transfer chain between SoxE (cyan) and SoxF (green) are shown as sticks, cofactors are coloured magenta and the cavity is the orange. The insert shows where the cavity protrudes on the protein surface.

Electrostatic surfaces were calculated for SoxE and SoxF, these results seem to fit with the hypothesis for the heterodimeric complex, with SoxF having a positive surface at the binding site and SoxE having a negative surface (Figure 5.26).

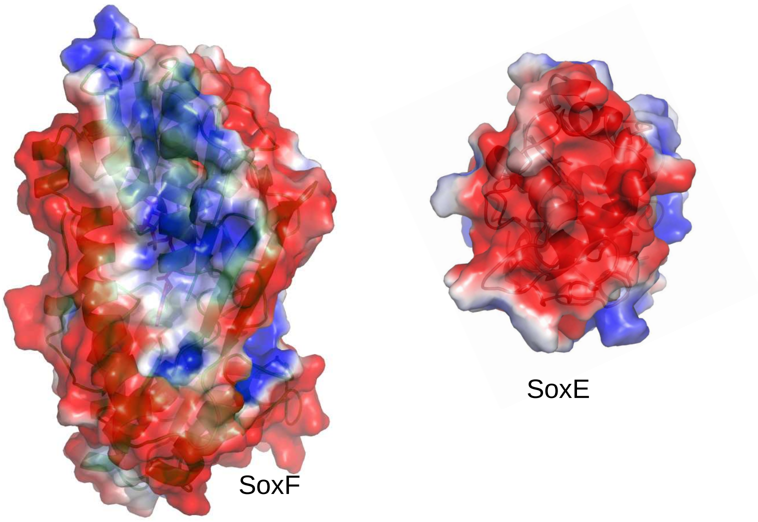


Figure 5.26 – The electrostatic surfaces of SoxF and SoxE from -2.00 (red) to 2.00 (blue) k_b T e_c^{-1}., showing the electrostatic surfaces of the proteins at their interaction faces.

5.4 Discussion

5.4.1 Post-translational Modifications to the active site cysteine residues in native and inhibited SoxF

The soaking of SoxF crystals in sodium metabisulfite or dissolved polysulfide was performed to solve the structure of the inhibited enzyme. RMSD values of 0.23 and 0.21 Å for the Cα alignments of the SoxF-SO_3^{2-} and SoxF-S_n structures with the SoxF-native structure show there is globally very little difference between the inhibited and active forms of SoxF. However, a closer look at the active site residues (Cys162 and Cys333) from each structure (Figure 5.27) highlights more significant differences.

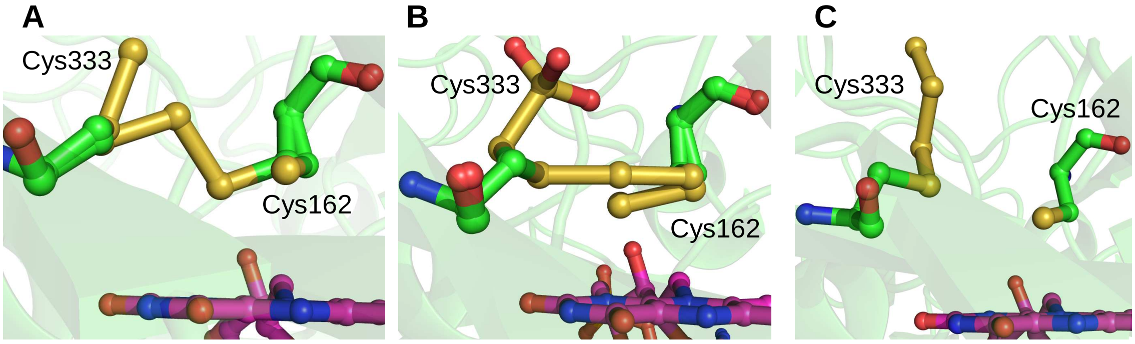


Figure 5.27 – The differences in the active site residues between **(A)** SoxF-native **(B)** SoxF-SO_3^{2-} and **(C)** SoxF-S_n. The active site cysteines are displayed as sticks, the surrounding protein backbone as transparent cartoon and FAD cofactors are coloured magenta.

SoxF-SO_3^{2-} is likely to have become inhibited via the process of sulfitolysis. Sulfite (SO_3^{2-}) has the ability to attack and break disulfide bridges via sulfitolysis **[Häberlein 1994, Drescher *et al* 1998]**, so it is not unreasonable to suggest that SO_3^{2-} could attack and break the trisulfide bridge between cysteines 162 and 333 in SoxF. This attack could potentially occur at two different positions, identified as 1 and 2 (Scheme 5.1). However analysis of the double difference Fourier electron density map would suggest SO_3^{2-} only attacks at position 2 of the bridge, since there is no electron density to suggest the SO_3^{2-} moiety is present on cysteine 162, only cysteine 333. Presumably, the reaction with SO_3^{2-} inactivates SoxF by blocking the reactive cysteine 333.

Scheme 5.1:

$$\underset{162}{R}{-}S\overset{1}{-}S\overset{2}{-}S{-}\underset{333}{R} + SO_3^{2-} \longrightarrow \underset{162}{R}{-}S{-}SH \qquad {}^{2-}O_3S{-}S{-}\underset{333}{R} + H_2$$

SoxF-S_n is likely to have become inhibited by the action of the polysulfide chains (HS_n^-). HS_n^- has the ability to attack persulfurated cysteine residues adding elongated sulfur chains to them, as identified in the Sud proteins **[Klimmek *et al* 1998, Klimmek *et al* 1999]**. X-ray data has shown that addition of HS_n^- to SoxF crystals results in the addition of sulfur to the persulfurated cysteine 333 (Scheme 5.2).

Scheme 5.2:

$$\underset{162}{\mathrm{R{-}SH}}\quad \underset{333}{\mathrm{HS{-}S{-}R}} + \mathrm{HS_n^-} \longrightarrow \underset{162}{\mathrm{R{-}SH}}\quad \underset{333}{\mathrm{{}^-S_{n+1}{-}S{-}R}}$$

Both these inhibited SoxF structures suggest that it is the persulfurated cysteine 333 that is the site of activity in SoxF's active site, since both forms of inhibition result in modifications to this specific residue.

5.4.2 Interaction between SoxF and SoxYZ

It has been speculated that SoxF is capable of binding the SoxYZ, either as a single SoxYZ heterodimer, or a SoxYZ heterotetramer involving two SoxYZ heterodimers. The docking of a SoxYZ heterotetramer to SoxF suggested that SoxYZ was unlikely to bind to SoxF in this form, since although there were no peptide clashes, the active cysteine of SoxY could not be brought close enough to the SoxF active site. This hypothesis is backed up by the results obtained from fitting a GGCGG pentapeptide sequence, which matches the sequence found at the C-terminal of SoxY, into the SoxF active site cavity. The result suggested only a single SoxYZ dimer would interact with SoxF as there was only sufficient space for one pentapeptide in the cavity (Figure 5.8) and the ring of proline resides that appear to form a rigid “plug hole” around the active persulfide residue (Figure 5.9). Based on this information, it is plausible that SoxF binds only a single SoxYZ heterodimer

So what could be happening with regards to SoxYZ binding to SoxF? A mechanism for SoxF-mediated activation of SoxY. This mechanism is based on; the observation of a mixed oxidized/reduced trisulfide bridge (ratio 42%:58% from occupancy refinement) involving cysteines 333 and 162 in SoxF, the observation of a mixture of trisulfide bridge and cysteine-S-sulfonate at position 333/cysteine persulfide at position 162 in the sulfite-soaked crystals (Scheme 5.1 and Figure 5.27B), the observation of a cysteine-S-trisulfane at position 333/cysteine at position 162 with no evidence for a trisulfide bridge in the sulphur-soaked crystals (Scheme 5.2 and Figure 5.27C), the absence of a stable SoxFE complex after purification (unlike the case for FccBA), and the structure of the active site cavity that restricts access to the trisulfite by bulky thiol-bearing moieties such as SoxY. The structures of the native and soaked forms of SoxF suggest an equilibrium exists between the oxidised and reduced form of the trisulfide in isolated SoxF, the absence of a stable complex with di-heme SoxE or any suitable cytochrome c ensures that electrons distribute between the oxidised and reduced forms of the trisulfite bridge and the FAD cofactor, and the structure of the active site cavity, suggests that it is the cysteine persulfide at position 333 of the reduced trisulfide bridge which is the most likely site of attachment of SoxY via the cysteine at position 110.

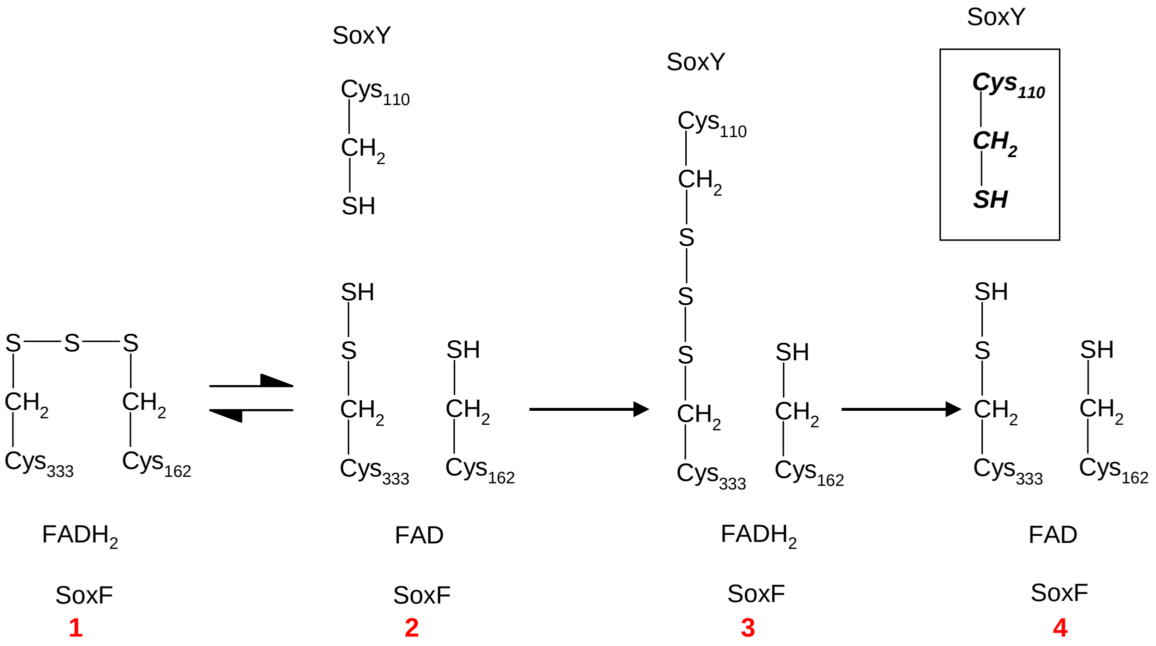


Figure 5.28 - A schematic illustrating the formation of a SoxZYF complex linked via a trisulfide bridge (numbers in red are used to identify each complex). Complex 1 and 2 (the oxidised and reduced forms of the SoxF trisulfide bridge) are in equilibrium, Cys108 from SoxY is able to attack the reduced trisulfite bridge (complex 2) forming an intermolecular trisulfide bridge between SoxF and SoxYZ (complex 3), the finger-like insertion in SoxZ relative to SoxY may obscure the site of cytochrome c binding to SoxF, isolating the system and precluding loss of electrons, the C-terminal region of SoxY undergoes a change in conformation during binding to the SoxF active site cavity, "activated" SoxYZ is released (complex 4) and the SoxF trisulfide bridge returns to its equilibrium between the oxidised and reduced states.

The schematic shown in Figure 5.28 illustrates the formation of a SoxZYF complex linked via a trisulfide bridge. Following the formation of an intermolecular trisulfide bridge between SoxF and SoxYZ long finger-like insertion in SoxZ relative to SoxY may obscure the site of cytochrome c binding to SoxF, isolating the system and precluding loss of electrons. The C-terminal region of SoxY undergoes a change in conformation during binding to the SoxF active site cavity. Release of "activated" SoxYZ follows. Figure 5.29 shows a structural schematic of this interaction.

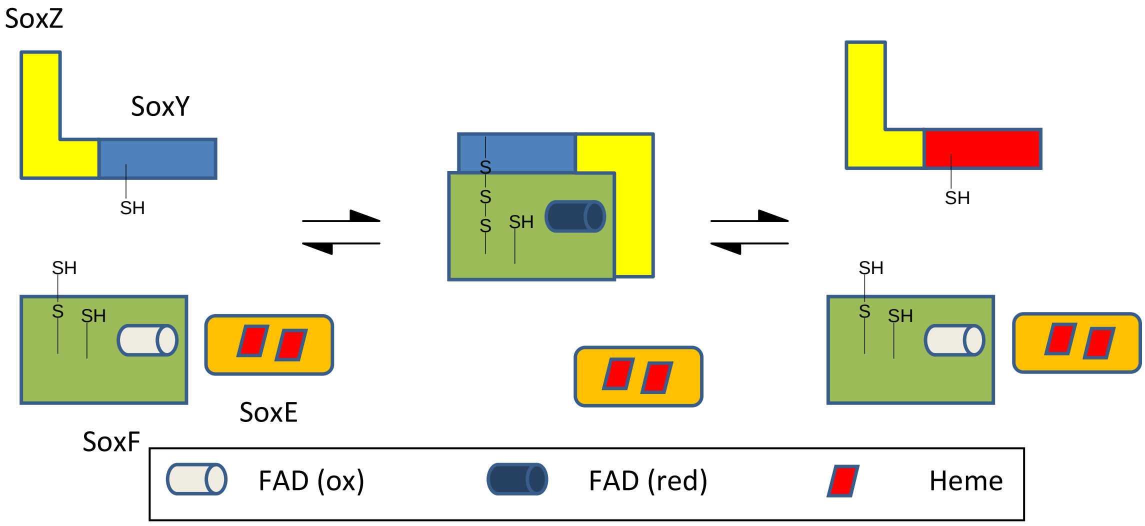


Figure 5.29 – Structural schematic of the SoxZYF complex showing. SoxYZ binds with SoxF via an intermolecular trisulfide bridge, blocking the SoxE binding site of SoxF with the long finger-like insertion in SoxZ, the C-terminal of SoxY is then modified and "activated" SoxYZ released.

5.4.3 Could the mixture of trisulfide plus cysteine and cycteine-s-sulfane/persulfide be a result of photoreduction

It has been shown that disulfide bridges can be broken as a result of photoreduction due to synchrotron radiation **[Weik *et al* 2000, Alphey *et al* 2003]**. To test whether such photoreduction has occurred in the SoxF structures, leading to the mixture of species in the active sites, a method used by Robets *et al* could be used employed.

Roberts *et al* had solved the structure of the N-terminal domain of the *Salmonalla typhimurium* flavoprotein AhpF with data collected at synchrotron sources **[Wood *et al* 2001]**, they found that the active site cysteine residues did not form a disulfide bridge, but appeared to be in close nonbonded interaction (~3 Å separation), despite the protein used for crystallisation being in the oxidised state and crystallised in the absence of any reducing agent. To test whether this disulfide bridge reduction was as a result of synchrotron radiation two datasets were collected from a single large crystal, one using their own laboratory X-ray source and then one at a synchrotron. They found that the disulfide bridge was oxidised in the dataset collected using their in-house X-ray source, but it reduced during synchrotron data collection **[Roberts *et al* 2005]**. In order for the hypothesise put forward in this chapter to be possible, this test would need to show that the trisulfide bridge was reduced before synchrotron data collection.

5.4.4 Comparisons of SoxF with it structural homologues

RMSD values for alignments of SoxF with its structural homologues (Table 5.3) suggest that the structures are superficially similar, as was observed with the differences between the native and soaked structures for SoxF. The significant differences become apparent when looking at the active sites. Figure 5.18 shows how the orientations and sulfur bindings of the two active site cysteine residues differs between each protein, with the persulferated cysteine in these structures being the equivalent of Cys162 from SoxF rather than Cys333.

There is also a noticeable difference in the pathways as predicted by CAVER, with the large cavity capable of accommodating at least a pentapeptide molecule observed in SoxF replaced by a much tighter path in the *Acidithiobacillus ferrooxidans* SQR (Figure 5.30). This change in active site cavity is likely due to two insertions and a deletion in the SQR enzyme that result in the blocking of the large cavity seen in SoxF. The changes in active site cavity also fit with the known function of the SQRs as enzymes that catalyse the reaction of sulfide ions to elemental sulfur **[Greisbeck *et al* 2002]**, which would not require a cavity the size of that observed in SoxF.

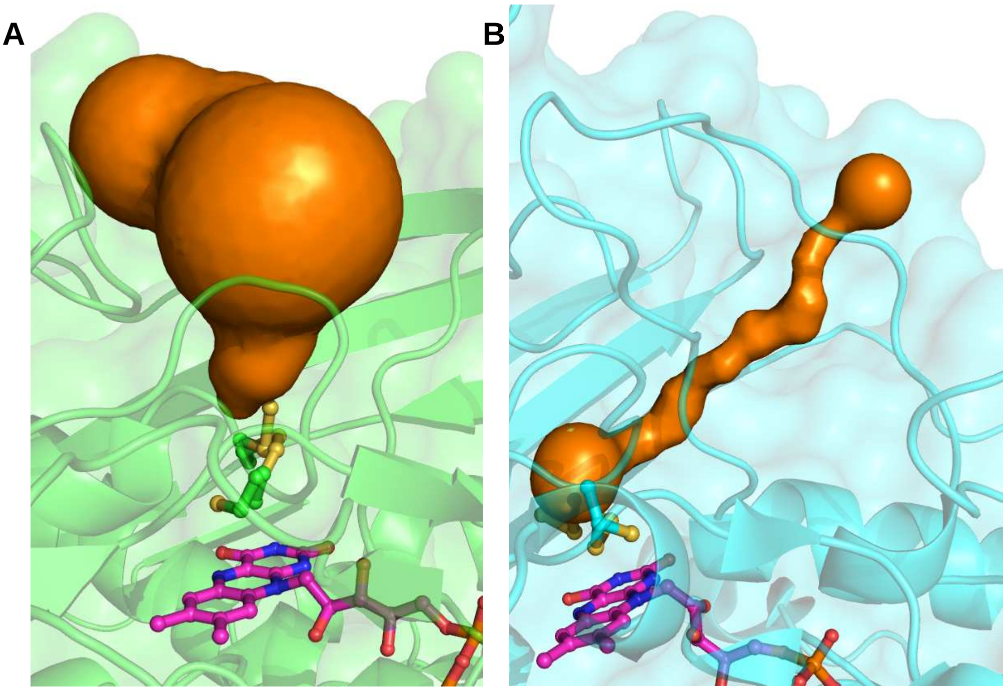


Figure 5.30 – Caver pathways from **(A)** *Paracoccus pantotrophus* SoxF and **(B)** *Acidithiobacillus ferrooxidans* sulfide:quinone oxidoreductase, demonstrating the significant differences in active sites between the two structures.

5.4.5 Reasons for no tight complex with the cytochrome subunit, observed in the *Allochromatium vinosum* sulfite dehydrogenase

An electrostatic surface potentials for FccA and FccB (Figure 5.31) suggest that the electrostatic attraction between the two subunits is inferior to that seen in the SoxE-SoxF complex (Figure 5.26) since the charges to not match up as clearly, indicating the differences in binding strength are more likely to be due to reside-residue contacts on the protein surface, or interactions between the FAD and heme cofactors.

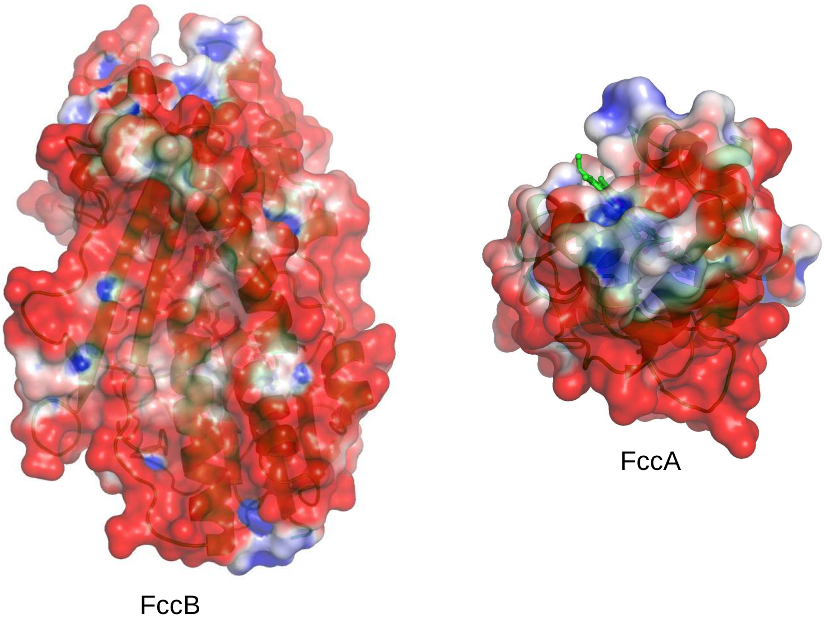


Figure 5.31 – The electrostatic surface potentials for FccA and FccB from -2.00 (red) to 2.00 (blue) k_b T e_c^{-1}, showing the interacting faces of each subunit.

The SoxE-SoxF complex does have two regions where main chain-main chain clashes occur between the two structures (Figure 5.32) that could point towards a reason for a more transient complex between SoxE and SoxF.

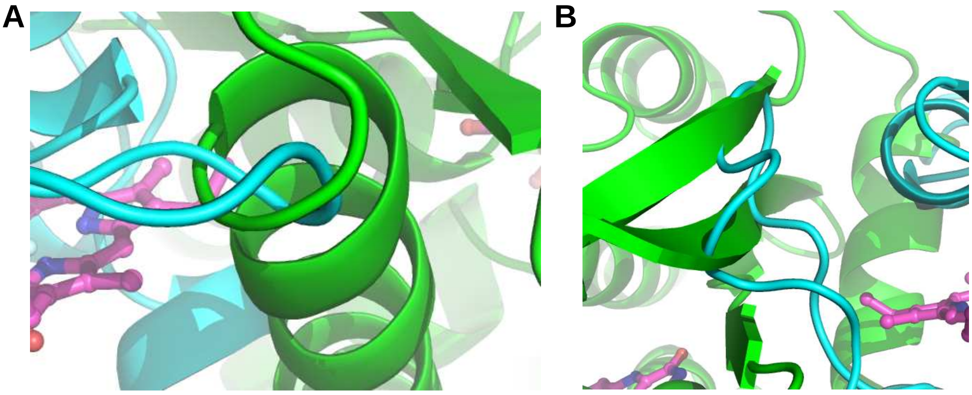


Figure 5.32 – The regions between SoxE (cyan) and SoxF (green) where the main chain-main chain clashes occur, at **(A)** residues 71-73 and **(B)** 136-143 of the modelled SoxE structure.

The interface between FccA and FccB has 13 hydrogen bonds, including one slat bridge, connecting the two subunits **[Chen *et al* 1994]**, only 8 hydrogen bonds could be identified at the interface between SoxE and SoxF, suggesting another reason for the more transient nature of the complex.

There is also a change in the electron transfer paths between the SoxE-SoxF (Figure 5.24) and FccA-FccB (Figure 5.33) complexes, with the FccA-FccB complex bringing the FAD and heme cofactors closer together (11.0Å, compared in 12.2Å).

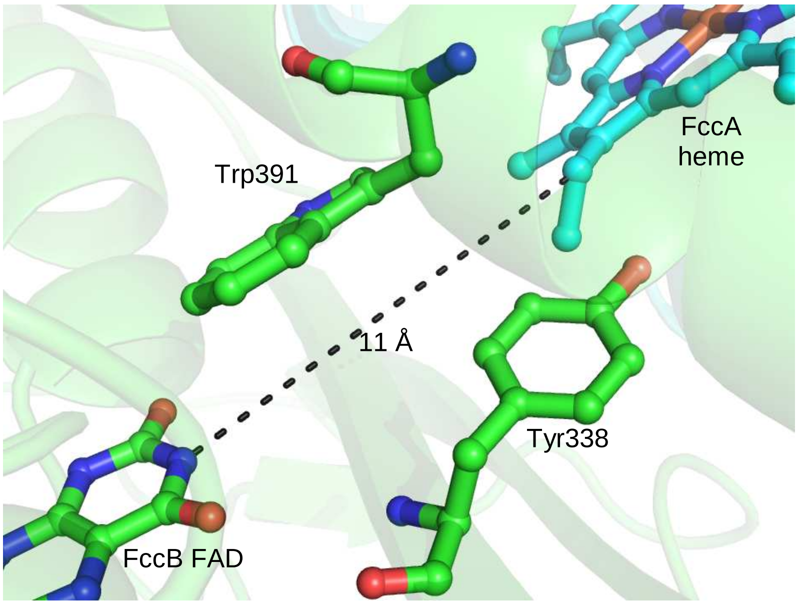


Figure 5.33 – The electron transfer path between FccB (green) and FccA (cyan), showing the distance between the FAD and heme cofactors and the residues located along the path.

Chapter 6 - Conclusions and future work

6.1 Heme packing and hemoprotein prediction

The results of the heme packing analysis in Chapter 2 has shown that specific heme motifs are conserved between proteins within each SCOP family and that, with the exception of the di-heme elbow pair motif which was found in the c_3-like and di-heme elbow SCOP families, each SCOP family contains a unique set of heme packing motifs. The analysis has also shown that, based on the proteins used in this analysis, heme packings from each family can be constructed from a small number of heme motifs. The cytochrome c_3–like SCOP family is populated by hemoproteins with four distinct motifs; cytochrome c_3, cytochrome c_7, 9-heme cytochrome and 16-heme cytochrome *c*, some of which have been shown to spear in more than one family (Figure 1.12). The di-heme elbow motif SCOP family contains proteins from a wide range of protein structures, however, the analysis of the heme packings has shown that the heme substructures from all of these proteins can be constructed from sequential packing of the parallel pair and di-heme elbow pair motifs, with the incorporation of an active site heme pair if the protein has enzymatic function.

The sequences that coordinate each heme motif were extracted, separated into subclusters based on their sequence length, polypeptide structure or phylogeny, aligned and used to build HMMs. This work has demonstrated that these HMMs can be used to predict the heme substructure of hemoproteins and provide templates for polypeptide modelling. This structure prediction methodology has been shown to produce models that agree with the structures determined by X-ray crystallography. In the case of the STC test case, this is true over the entire length of the heme substructure and true for a part of the heme substructure in the GSU_1996 test case. The problems with the GSU_1996 prediction were due to the presence of a hitherto unobserved heme pair motif in the experimental structure. As this prediction methodology is based on recognition of previously observed heme packings, as new heme pair motifs are observed and incorporated into the heme pair database, problems such as this will become less frequent. However, it should be noted that the HMM based prediction methodology developed in this thesis provides a significantly closer approximation of the GSU_1996 structure than any of the existing techniques tested during this work.

6.1.1 Future development of the structure prediction method

The automation of the hemoprotein structure prediction methodology described in this thesis will be an important next step. The successful automation of HMM searching has helped to allow more rapid identification of target proteins that have a higher likelihood of giving a successful prediction (see Chapter 3), but the automation of the model building itself will speed up the process even further. A useful addition to the methods described in Chapter 2 would be the incorporation of *de novo* prediction methods for the remaining unmodelled regions of hemoproteins (i.e. the N-terminal region before the first heme and the C-terminal region after the last heme). Another feature to add would be the ability to infer functional properties on the predicted structures by simulating heme redox potentials and residue pk_a values using computational methods such as Multi-Conformational Continuum Electrostatics (MCCE) **[Georgescu *et al* 2002]**.

The information discovered and techniques developed in this work may eventually lead to the design of new multiheme cytochrome structures with specific functions, such as artificial nano-wires. This could most likely involve the design of self-assembling proteins with structures similar to GSU_1996 and GSU_2210.

It may also be possible to adapt this methodology to predict the structures of other cofactor rich proteins, such as iron sulfur cluster containing proteins. The success of this would rely on the availability of sufficient structural data to identify conserved patterns in cofactor packing and corresponding amino acid sequences for HMM construction. It is that lack of such information that was the limiting factor in the quality of the predictions of the copper cluster containing proteins Mac1 and Ace1 in Chapter 4.

6.2 Copper chaperone studies

The work described on the copper chaperone CopZ in Chapter 4 has shown it is able to adjust its copper coordination stoichiometries to bind varying amounts of monovalent copper, with a four copper dimer and three copper timer observed in the crystal and one and two copper dimers observed in solution **[Kihlken *et al* 2002]**. This provides an insight into the versatility of the Atx1-like copper chaperones with respect to copper coordination. This versatility has been further highlighted by crystal structures published recently by Badarau *et al*, who solved a range of copper bound species of Atx1 from the cyanobacterium *Synechocystis* PCC6803 which used different monomer packing geometries and copper coordinating residues to bind varying number of copper ion **[Badarau *et al* 2010]**.

The structural homology of the N-terminal domain of CopA to CopZ and the existence of the trimeric species allowed putative models of copper exchange complexes to be constructed by substituting in turn each of the two N-terminal copper binding domains of CopA. The usual indicators of permanency for protein complexes (e.g. **[Ponsting *et al* 2000]**) could not be used for the assessment of this transient complex because a major driver for stability is presumably the formation of specific Cu(I) ion to cysteine thiolate co-ordinate bonds. However, the relatively low value of the solvent-accessible area lost on complex formation and the low number of inter-subunit hydrogen bonds are at least consistent with a tentative classification of the modelled complexes as transient.

6.2.1 Future work on the copper homeostasis pathway of *B.subtilis*

An area for future work for this project would need to be the development of a method to test the validity of the copper exchange complex. One method for accomplishing this would be to prepare protein solutions containing stable complexes of differing mixtures of CopA bound to CopZ and copper ions for crystallisation experiments, leading to X-ray data collection to ascertain a 3D structure of the complex. However the predicted transient nature of this complex is likely to make isolating a complex stable enough to form protein crystals quite challenging, if this is the case, surface plasmon resonance **[Van Der Merwe 2001]** or analytical ultracentrifugation are other potential techniques that could be analysed to assess the nature and strength of CopA-CopZ complex formation. At the very least a crystal structure for the N-termini domains of CopA would give a more accurate structure for the modelling of the complex.

6.3 The sulfur oxidation pathway of *Paracoccus pantotrophus*

The X-ray crystal structures of native and two product inhibited forms of the *sox* cycle flavoprotein SoxF have been solved and described in this thesis (Chapter 5). The structures are highly homologous with the main differences found at the active site cysteine residues (Figure 5.27).

SoxF is thought to interact with the di-heme cytochrome SoxE, although no structure for such or complex or SoxE itself is currently available. A model for SoxE was created based on a homologous structure (PDB ID: 1M70 **[Kadziola *et al* 1995]**) and a SoxEF complex modelled based on the homologous FccAB complex from *Allochromatium vinosum* (PDB ID: 1FCD **[Chen *et al* 1994]**).

The action of the active site residues and the size of the active site cavity led to the proposal of a mechanism for SoxF-mediated activation of the *sox* cycle sulfur transporting heterodimer SoxYZ via modification of the C-terminal of SoxY (Figure 5.29).

6.3.1 Future work on the structure-function relationships of SoxF

It will be necessary to assess whether or not the break in the active site trisulfide bridge observed in the structure of SoxF occurs naturally and is a requirement for function or simply arose as a result of photoreduction in the synchrotron X-ray beam. One method for ascertaining this would be to dissolve a number of crystals before and after X-ray exposure for analysis using mass spectrometry. It would be important to use dissolved crystals rather than protein solution to ensure the species being analysed is that same as that observed in crystal and not another form found only in solution. An alternative method could be to collect a dataset from a crystal using an “in house” radiation source, this would reduce the intensity of the X-rays and has been shown to prevent the reduction of disulfide bridge in flavoprotein structures **[Roberts *et al* 2005]**. For completeness, if the crystal survives it would also be good to collect a dataset at a synchrotron using the same crystal and examine any changes in the composition of the active site that could be due to photoreduction.

In order to validate the model of the SoxEF complex the ideal scenario would be to purify and crystallise a stable (trapped) form of the SoxEF complex, however if this is not possible then a crystal structure for SoxE alone would be enough to create a more accurate model of the complex and to characterise the unusual WXXCH presumed heme binding motif found in SoxE.

Appendix I - Pair clusters - 125 pairs (40% His-His set), 1.5A cutoff

Cluster 01 (28)
1SP3-A804-A803
1M1Q-A804-A803
2CY3-A121-A119
1OAH-A1522-A1521
1FGJ-A3-A2
1FGJ-A8-A7
1FT5-A216-A215
2OZY-A204-A205
2OT4-A1002-A1003
2OZY-A202-A203
2OT4-A1007-A1008
1OAH-A1523-A1524
3BNJ-A516-A517
1FGJ-A5-A6
1Y0P-A801-A802
1M1Q-A801-A802
3BNJ-A514-A515
2OT4-A1005-A1006
1RWJ-A90-A91
1OFW-A1299-A1301
1OFW-A1294-A1296
3CAO-A104-A106
3BXU-A72-A73
2BQ4-A1115-A1117
1GYO-A111-A113
1WAD-A112-A114
1AQE-A119-A121
1J0P-A1001-A1003

Cluster 02 (20)
1FGJ-A1-A2
2CZS-A500-A501
1SP3-A801-A803
1JNI-A111-A110
1OGY-B1128-B1129
1FGJ-A5-A3
1OAH-A1523-A1522
2OT4-A1003-A1005
1FGJ-A6-A7
2OT4-A1006-A1007
3BNJ-A515-A516
1SP3-A806-A807
1Y0P-A802-A803
1SP3-A804-A805
1M1Q-A802-A803
2OZY-A201-A202
2OZY-A203-A204
2OT4-A1001-A1002
1FT5-A213-A215
1H21-A1248-A1249

Cluster 03 (10)
2BQ4-A1115-A1118
1GYO-A111-A114
1WAD-A112-A115
1J0P-A1001-A1004
3BXU-A72-A74
2CY3-A119-A122
1AQE-A122-A119
3CAO-A107-A104
1OFW-A1294-A1298
1OFW-A1299-A1302

Cluster 04 (10)
1J0P-A1003-A1004
3BXU-A73-A74
2CY3-A121-A122
1AQE-A121-A122
1OFW-A1301-A1302
1GYO-A113-A114
2BQ4-A1117-A1118
1WAD-A115-A114
1OFW-A1298-A1296
3CAO-A107-A106

Cluster 05 (9)
1OFW-A1294-A1295
2BQ4-A1115-A1116
1J0P-A1001-A1002
2CY3-A119-A120
1AQE-A119-A120
1GYO-A111-A112
1WAD-A113-A112
1OFW-A1300-A1299
3CAO-A105-A104

Cluster 06 (9)
2CY3-A121-A120
1AQE-A121-A120
1OFW-A1295-A1296
1GYO-A112-A113
2BQ4-A1116-A1117
1WAD-A113-A114
1J0P-A1002-A1003
1OFW-A1300-A1301
3CAO-A105-A106

Cluster 07 (9)
1GYO-A114-A112
1OFW-A1302-A1300
1AQE-A122-A120
2CY3-A122-A120
1OFW-A1295-A1298
3CAO-A105-A107
2BQ4-A1116-A1118
1WAD-A113-A115
1J0P-A1002-A1004

Cluster 08 (7)
1SP3-A803-A805
2OZY-A202-A204
3BNJ-A514-A516
1M1Q-A801-A803
1Y0P-A801-A803
2OT4-A1005-A1007
2OT4-A1002-A1005

Cluster 09 (2)
1OAH-A1524-B1524
3BNJ-A517-B517

Cluster 10 (2)
1AQE-A119-B119
1GYO-A111-B111

Cluster 11 (2)
1SP3-A806-A805
1SP3-A808-A807

Cluster 12 (2)
1FGJ-A5-A2
1OAH-A1523-A1521

Cluster 13
2OT4-A1003-B1003

Cluster 14
2FWT-A803-A805

Cluster 15
1OFW-A1297-A1298

Cluster 16
1FGJ-A8-C2

Cluster 17
1OFW-A1297-A1296

Cluster 18
1OFW-A1297-A1299

Cluster 19
1KQF-809-810

Cluster 20
1Y5I-C806-C807

Cluster 21
1OFW-A1297-A1301

Cluster 22
1H21-A1249-B1249

Cluster 23
1Y0P-A803-A804

Cluster 24
2BS2-C1255-C1256

Cluster 25
1OFW-A1298-B1300

Cluster 26
1OFW-A1300-B1298

Cluster 27
1SP3-A805-A807

Appendix II - 115 Triplets from 40% His-His set

Cluster 01 (23) | 01-02
1FGJ-A3-A2-A1
1FGJ-A3-A2-A5
1FGJ-A5-A6-A3
1FGJ-A5-A6-A7
1FGJ-A8-A7-A6
1FT5-A216-A215-A213
1M1Q-A801-A802-A803
1M1Q-A804-A803-A802
1OAH-A1522-A1521-A1523
1OAH-A1523-A1524-A1522
1SP3-A804-A803-A801
1SP3-A804-A803-A805
1Y0P-A801-A802-A803
2OT4-A1002-A1003-A1001
2OT4-A1002-A1003-A1005
2OT4-A1005-A1006-A1003
2OT4-A1005-A1006-A1007
2OT4-A1007-A1008-A1006
2OZY-A202-A203-A201
2OZY-A202-A203-A204
2OZY-A204-A205-A203
3BNJ-A514-A515-A516
3BNJ-A516-A517-A515

Cluster 02 (10) | 01-04
1AQE-A119-A121-A122
1GYO-A111-A113-A114
1J0P-A1001-A1003-A1004
1OFW-A1294-A1296-A1298
1OFW-A1299-A1301-A1302
1WAD-A112-A114-A115
2BQ4-A1115-A1117-A1118
2CY3-A121-A119-A122
3BXU-A72-A73-A74
3CAO-A104-A106-A107

Cluster 03 (9) | 01-06
1AQE-A119-A121-A120
1GYO-A111-A113-A112
1J0P-A1001-A1003-A1002
1OFW-A1294-A1296-A1295
1OFW-A1299-A1301-A1300
1WAD-A112-A114-A113
2BQ4-A1115-A1117-A1116
2CY3-A121-A119-A120
3CAO-A104-A106-A105

Cluster 04 (9) | 04-07
1AQE-A121-A122-A120
1GYO-A113-A114-A112
1J0P-A1003-A1004-A1002
1OFW-A1298-A1296-A1295
1OFW-A1301-A1302-A1300
1WAD-A115-A114-A113
2BQ4-A1117-A1118-A1116
2CY3-A121-A122-A120
3CAO-A107-A106-A105

Cluster 05 (9) | 03-06
1GYO-A111-A114-A112
1J0P-A1001-A1004-A1002
1OFW-A1294-A1298-A1295
1OFW-A1299-A1302-A1300
1WAD-A112-A115-A113
2BQ4-A1115-A1118-A1116
2CY3-A119-A122-A120
1AQE-A119-A120-A122
3CAO-A105-A104-A107

Cluster 06 (5) | 01-05
1M1Q-A804-A803-A801
2OT4-A1005-A1006-A1002
2OT4-A1007-A1008-A1005
2OZY-A204-A205-A202
3BNJ-A516-A517-A514

Cluster 07 (4) | 02-05
1SP3-A801-A803-A805
2OT4-A1001-A1002-A1005
2OT4-A1003-A1005-A1007
2OZY-A201-A202-A204

Cluster 08 (3) | 02-12
1SP3-A804-A805-A806
1SP3-A806-A807-A805
1SP3-A806-A807-A808

Cluster 09 (2) | 01-09
1OAH-A1523-A1524-B1524
3BNJ-A516-A517-B517

Cluster 10 (2) | 01-10
1AQE-A119-A121-B119
1GYO-A111-A113-B111

Cluster 11 | 01-11
2OT4-A1002-A1003-B1003

Cluster 12 (2) | 01-13
1FGJ-A5-A6-A2
1OAH-A1523-A1524-A1521

Cluster 13 | 02-11
2OT4-A1003-A1005-B1003

Cluster 14 (2) | 03-10
1AQE-A122-A119-B119
1GYO-A111-A114-B111

Cluster 15 | 05-05
2OT4-A1005-A1007-A1002

Cluster 16 (2) | 06-10
1AQE-A119-A120-B119
1GYO-A111-A112-B111

Cluster 18 | 01-17
1FGJ-A8-A7-C2

Cluster 19 | 01-18
1OFW-A1294-A1296-A1297

Cluster 20 | 01-23
1OFW-A1299-A1301-A1297

Cluster 21 | 02-13
1FGJ-A1-A2-A5

Cluster 23 | 02-24
1H21-A1248-A1249-B1249

Cluster 24 | 02-25
1Y0P-A802-A803-A804

Cluster 25 | 02-29
1SP3-A804-A805-A807

Cluster 26 | 03-15
1OFW-A1294-A1298-A1297

Cluster 27 | 03-20
1OFW-A1299-A1302-A1297

Cluster 28 | 03-27
1OFW-A1294-A1298-B1300

Cluster 29 | 04-18
1OFW-A1298-A1296-A1297

Cluster 30 | 04-23
1OFW-A1301-A1302-A1297

Cluster 31 | 04-27
1OFW-A1298-A1296-B1300

Cluster 32 | 05-12
1SP3-A803-A805-A806

Cluster 35 | 05-25
1Y0P-A801-A803-A804

Cluster 36 | 05-29
1SP3-A803-A805-A807

Cluster 37 | 06-20
1OFW-A1300-A1299-A1297

Cluster 38 | 06-28
1OFW-A1300-A1299-B1298

Cluster 39 | 07-18
1OFW-A1295-A1296-A1297

Cluster 40 | 07-23
1OFW-A1300-A1301-A1297

Cluster 41 | 07-28
1OFW-A1300-A1301-B1298

Cluster 42 | 08-15
1OFW-A1295-A1298-A1297

Cluster 43 | 08-27
1OFW-A1295-A1298-B1300

Cluster 44 | 08-28
1OFW-A1302-A1300-B1298

Cluster 45 | 12-29
1SP3-A808-A807-A805

Cluster 46 | 15-20
1OFW-A1297-A1298-A1299

Cluster 47 | 15-23
1OFW-A1297-A1298-A1301

Cluster 48 | 15-27
1OFW-A1297-A1298-B1300

Cluster 49 | 18-20
1OFW-A1297-A1296-A1299

Cluster 50 | 18-23
1OFW-A1297-A1296-A1301

Appendix III - 99 Quartets from 40% His-His set

Cluster 1 (8) | 01-02-01
1FGJ-A5-A6-A3-A2
1FGJ-A8-A7-A6-A5
1M1Q-A801-A802-A803-A804
1OAH-A1523-A1524-A1522-A1521
2OT4-A1005-A1006-A1003-A1002
2OT4-A1005-A1006-A1007-A1008
2OZY-A202-A203-A204-A205
3BNJ-A514-A515-A516-A517

Cluster 2 (9) | 01-05-04
1AQE-A119-A121-A120-A122
1GYO-A111-A113-A112-A114
1J0P-A1001-A1003-A1002-A1004
1OFW-A1294-A1296-A1295-A1298
1OFW-A1299-A1301-A1300-A1302
1WAD-A112-A114-A113-A115
2BQ4-A1115-A1117-A1116-A1118
2CY3-A121-A119-A120-A122
3CAO-A104-A106-A107-A105

Cluster 3 (6) | 01-02-02
1FGJ-A3-A2-A1-A5
1FGJ-A5-A6-A3-A7
1SP3-A804-A803-A805-A801
2OT4-A1002-A1003-A1001-A1005
2OT4-A1005-A1006-A1007-A1003
2OZY-A202-A203-A204-A201

Cluster 4 (3) | 01-02-13
2OT4-A1002-A1003-A1001-B1003
2OT4-A1002-A1003-A1005-B1003
2OT4-A1005-A1006-A1003-B1003

Cluster 5 (2) | 01-02-08
2OT4-A1002-A1003-A1005-A1007
2OT4-A1005-A1006-A1007-A1002

Cluster 6 (2) | 01-08-02
2OT4-A1007-A1008-A1005-A1003
2OZY-A204-A205-A202-A201

Cluster 7 (2) | 02-08-01
2OT4-A1001-A1002-A1005-A1006
2OT4-A1003-A1005-A1007-A1002

Cluster 8 (2) | 01-05-10
1AQE-A119-A121-A122-B119
1GYO-A111-A113-A114-B111

Cluster 9 (2) | 01-04-10
1AQE-A119-A121-A120-B119
1GYO-A111-A113-A112-B111

Cluster 10 | 01-08-08
2OT4-A1007-A1008-A1005-A1002

Cluster 11 | 01-02-09
1OAH-A1523-A1524-A1522-B1524

Cluster 12 | 01-02-11
1SP3-A804-A803-A805-A806

Cluster 13 | 01-02-12
1FGJ-A5-A6-A7-A2

Cluster 14 | 01-02-16
1FGJ-A8-A7-A6-C2

Cluster 15 | 01-02-23
1Y0P-A801-A802-A803-A804

Cluster 16 | 01-02-27
1SP3-A804-A803-A805-A807

Cluster 17 | 01-04-25
1OFW-A1294-A1296-A1298-B1300

Cluster 18 | 01-08-09
3BNJ-A516-A517-A514-B517

Cluster 19 | 01-05-17
1OFW-A1294-A1296-A1295-A1297

Cluster 20 | 01-05-26
1OFW-A1299-A1301-A1300-B1298

Cluster 21 | 01-09-02
3BNJ-A516-A517-B517-A515

Cluster 22 | 01-09-12
1OAH-A1523-A1524-B1524-A1521

Cluster 23 | 01-12-02
1FGJ-A5-A6-A2-A1

Cluster 24 | 01-17-18
1OFW-A1294-A1296-A1297-A1299

Cluster 25 | 01-17-21
1OFW-A1294-A1296-A1297-A1301

Cluster 26 | 01-21-04
1OFW-A1299-A1301-A1297-A1302

Cluster 27 | 01-21-05
1OFW-A1299-A1301-A1297-A1300

Cluster 28 | 01-21-15
1OFW-A1299-A1301-A1297-A1298

Cluster 29 | 01-21-17
1OFW-A1299-A1301-A1297-A1296

Cluster 30 | 02-08-08
2OT4-A1001-A1002-A1005-A1007

Cluster 31 | 02-08-13
2OT4-A1003-A1005-A1007-B1003

Cluster 32 | 02-08-11
1SP3-A801-A803-A805-A806

Cluster 33 | 02-11-02
1SP3-A804-A805-A806-A807

Cluster 34 | 02-11-11
1SP3-A806-A807-A808-A805

Cluster 35 | 02-27-11
1SP3-A804-A805-A807-A808

Cluster 36 | 03-05-10
1GYO-A111-A114-A112-B111

Cluster 37 | 03-05-15
1OFW-A1294-A1298-A1295-A1297

Cluster 38 | 03-05-18
1OFW-A1299-A1302-A1300-A1297

Cluster 39 | 03-05-25
1OFW-A1294-A1298-A1295-B1300

Cluster 40 | 03-05-26
1OFW-A1299-A1302-A1300-B1298

Cluster 41 | 03-15-18
1OFW-A1294-A1298-A1297-A1299

Cluster 42 | 03-15-21
1OFW-A1294-A1298-A1297-A1301

Cluster 43 | 03-18-15
1OFW-A1299-A1302-A1297-A1298

Cluster 44 | 03-18-17
1OFW-A1299-A1302-A1297-A1296

Cluster 45 | 04-06-21
1OFW-A1301-A1302-A1300-A1297

Cluster 46 | 04-06-25
1OFW-A1298-A1296-A1295-B1300

Cluster 47 | 04-06-26
1OFW-A1301-A1302-A1300-B1298

Cluster 48 | 04-17-01
1OFW-A1298-A1296-A1297-A1294

Cluster 49 | 04-17-18
1OFW-A1298-A1296-A1297-A1299

Cluster 50 | 04-17-21
1OFW-A1298-A1296-A1297-A1301

Cluster 51 | 04-17-25
1OFW-A1298-A1296-A1297-B1300

Cluster 52 | 04-21-15
1OFW-A1301-A1302-A1297-A1298

Cluster 53 | 04-21-17
1OFW-A1301-A1302-A1297-A1296

Cluster 54 | 08-11-02
1SP3-A803-A805-A806-A807

Cluster 55 | 08-27-02
1SP3-A803-A805-A807-A801

Cluster 56 | 08-27-11
1SP3-A803-A805-A807-A808

Cluster 57 | 05-07-10
1AQE-A119-A120-A122-B119

Cluster 58 | 05-18-15
1OFW-A1300-A1299-A1297-A1298

Cluster 59 | 05-18-17
1OFW-A1300-A1299-A1297-A1296

Cluster 60 | 05-18-26
1OFW-A1300-A1299-A1297-B1298

Cluster 61 | 06-17-04
1OFW-A1295-A1296-A1297-A1298

Cluster 62 | 06-17-18
1OFW-A1295-A1296-A1297-A1299

Cluster 63 | 06-17-21
1OFW-A1295-A1296-A1297-A1301

Cluster 64 | 06-21-15
1OFW-A1300-A1301-A1297-A1298

Cluster 65 | 06-21-17
1OFW-A1300-A1301-A1297-A1296

Cluster 66 | 06-21-26
1OFW-A1300-A1301-A1297-B1298

Cluster 67 | 07-15-18
1OFW-A1295-A1298-A1297-A1299

Cluster 68 | 07-15-21
1OFW-A1295-A1298-A1297-A1301

Cluster 69 | 15-18-25
1OFW-A1297-A1298-A1299-B1300

Cluster 70 | 15-25-03
1OFW-A1297-A1298-B1300-A1294

Cluster 71 | 15-25-07
1OFW-A1297-A1298-B1300-A1295

Cluster 72 | 15-25-21
1OFW-A1297-A1298-B1300-A1301

Appendix IV - 59 Quintets from 40% His-His set

Cluster 1 (2) | 01-08-02-02
2OT4-A1007-A1008-A1005-A1003-A1006
2OZY-A204-A205-A202-A201-A203

Cluster 2 (2) | 01-02-02-13
2OT4-A1002-A1003-A1001-A1005-B1003
2OT4-A1005-A1006-A1007-A1003-B1003

Cluster 3 (2) | 01-05-04-10
1AQE-A119-A121-A120-A122-B119
1GYO-A111-A113-A112-A114-B111

Cluster 4 | 01-02-08-02
2OT4-A1002-A1003-A1005-A1007-A1006

Cluster 5 | 01-08-02-08
2OT4-A1007-A1008-A1005-A1003-A1002

Cluster 6 | 01-08-02-13
2OT4-A1007-A1008-A1005-A1003-B1003

Cluster 7 | 01-02-01-02
1FGJ-A8-A7-A6-A5-A3

Cluster 8 | 01-02-01-09
3BNJ-A514-A515-A516-A517-B517

Cluster 9 | 01-02-01-13
2OT4-A1005-A1006-A1003-A1002-B1003

Cluster 10 | 01-02-01-12
1FGJ-A8-A7-A6-A5-A2

Cluster 11 | 01-02-01-16
1FGJ-A8-A7-A6-A5-C2

Cluster 12 | 01-02-02-01
1FGJ-A5-A6-A3-A7-A2

Cluster 13 | 01-02-02-11
1SP3-A804-A803-A805-A801-A806

Cluster 14 | 01-02-02-27
1SP3-A804-A803-A805-A801-A807

Cluster 15 | 01-02-05-11
2OT4-A1002-A1003-A1005-A1007-B1003

Cluster 16 | 01-02-09-12
1OAH-A1523-A1524-A1522-B1524-A1521

Cluster 17 | 01-02-11-27
1SP3-A804-A803-A805-A806-A807

Cluster 18 | 01-05-04-17
1OFW-A1294-A1296-A1295-A1298-A1297

Cluster 19 | 01-05-04-21
1OFW-A1299-A1301-A1300-A1302-A1297

Cluster 20 | 01-05-04-25
1OFW-A1294-A1296-A1295-A1298-B1300

Cluster 21 | 01-05-04-26
1OFW-A1299-A1301-A1300-A1302-B1298

Cluster 22 | 01-12-02-02
1FGJ-A5-A6-A2-A1-A3

Cluster 23 | 01-21-05-26
1OFW-A1299-A1301-A1297-A1300-B1298

Cluster 24 | 01-21-15-25
1OFW-A1299-A1301-A1297-A1298-B1300

Cluster 25 | 02-06-01-01
2OT4-A1001-A1002-A1005-A1006-A1003

Cluster 26 | 02-06-01-02
2OT4-A1001-A1002-A1005-A1006-A1007

Cluster 27 | 02-06-01-13
2OT4-A1003-A1005-A1007-A1002-B1003

Cluster 28 | 02-27-11-02
1SP3-A804-A805-A807-A808-A806

Cluster 29 | 03-05-15-18
1OFW-A1294-A1298-A1295-A1297-A1299

Cluster 30 | 03-05-15-21
1OFW-A1294-A1298-A1295-A1297-A1301

Cluster 31 | 03-05-15-25
1OFW-A1294-A1298-A1295-A1297-B1300

Cluster 32 | 03-05-18-26
1OFW-A1299-A1302-A1300-A1297-B1298

Cluster 33 | 03-15-18-25
1OFW-A1294-A1298-A1297-A1299-B1300

Cluster 34 | 03-15-21-25
1OFW-A1294-A1298-A1297-A1301-B1300

Cluster 35 | 03-18-15-17
1OFW-A1299-A1302-A1297-A1298-A1296

Cluster 36 | 03-18-15-21
1OFW-A1299-A1302-A1297-A1298-A1301

Cluster 37 | 03-18-15-25
1OFW-A1299-A1302-A1297-A1298-B1300

Cluster 38 | 04-06-21-26
1OFW-A1301-A1302-A1300-A1297-B1298

Cluster 39 | 04-17-01-25
1OFW-A1298-A1296-A1297-A1294-B1300

Cluster 40 | 04-17-18-25
1OFW-A1298-A1296-A1297-A1299-B1300

Cluster 41 | 04-17-21-25
1OFW-A1298-A1296-A1297-A1301-B1300

Cluster 42 | 04-21-15-17
1OFW-A1301-A1302-A1297-A1298-A1296

Cluster 43 | 04-21-15-25
1OFW-A1301-A1302-A1297-A1298-B1300

Cluster 44 | 05-18-15-17
1OFW-A1300-A1299-A1297-A1298-A1296

Cluster 45 | 05-18-15-25
1OFW-A1300-A1299-A1297-A1298-B1300

Cluster 46 | 05-18-15-26
1OFW-A1300-A1299-A1297-A1298-B1298

Cluster 47 | 05-18-17-26
1OFW-A1300-A1299-A1297-A1296-B1298

Cluster 48 | 06-17-04-25
1OFW-A1295-A1296-A1297-A1298-B1300

Cluster 49 | 06-21-15-17
1OFW-A1300-A1301-A1297-A1298-A1296

Cluster 50 | 06-21-15-25
1OFW-A1300-A1301-A1297-A1298-B1300

Cluster 51 | 06-21-15-26
1OFW-A1300-A1301-A1297-A1298-B1298

Cluster 52 | 06-21-17-26
1OFW-A1300-A1301-A1297-A1296-B1298

Cluster 53 | 07-15-18-25
1OFW-A1295-A1298-A1297-A1299-B1300

Cluster 54 | 07-15-21-25
1OFW-A1295-A1298-A1297-A1301-B1300

Cluster 55 | 08-27-02-02
1SP3-A803-A805-A807-A801-A806

Cluster 56 | 08-27-11-02
1SP3-A803-A805-A807-A808-A806

References

Abramson, J., Riistama, S., Larsson, G., Jasaitis, A., Svensson-Ek, M., Laakkonen, L., Puustinen, A., Iwata, S. and Wikström, M. (2000) The structure of the ubiquinol oxidase from *Escherichia coli* and its ubiquinone binding site. *Nature Structural Biology,* **7**, 910 – 917

Achila, D., Banci, L., Bertini, I., Bunce, J., Ciofi-Baffoni, S. and Huffman, D.L. (2006) Structure of human Wilson protein domains 5 and 6 and their interplay with domain 4 and the copper chaperone HAH1 in copper uptake. *Proc.Natl.Acad.Sci.Usa* **103,** 5729-5734

Adams, P.D., Afonine, P.V., Bunkoczi, G., Chen, V.B., Davis, I.W., Echols, N., Headd, J.J., Hung, L.-W., Kapral, G.J., Grosse-Kunstleve, R.W., McCoy, A.J., Moriarty, N.W., Oeffner, R., Read, R.J., Richardson, D.C., Richardson, J.S., Terwilliger, T.C. and Zwart, P.H. (2010). PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Cryst*. **D66**, 213-221

Allen, F.H. (2002) The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallogr*., **B58**, 380-388,

Allen, J.W.A., Leach, N. and Ferguson, S.J. (2005). The histidine of the c-type cytochrome CXXCH haem-binding motif is essential for haem attachment by the *Escherichia coli* cytochrome c maturation (Ccm) apparatus. *Biochem. J.* **389**, 587–592

Alphey, M.S., Gabrielsen, M., Micossi, E., Leonard, G.A., McSweeney, S.M., Ravelli, R.B., Tetaud, E., Fairlamb, A.H., Bond, C.S., and Hunter, W.N. 2003. Tryparedoxins from *Crithidia fasciculata* and *Trypanosoma brucei*: Photoreduction of the redox disulfide using synchrotron radiation and evidence for a conformational switch implicated in function. *J. Biol. Chem.* **278**: 25919–25925.

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.

Anderson, K.E., Sassa, S., Bishop, D.F. & Desnick, R.J. (2001) Disorders of heme biosynthesis: X-linked sideroblastic anemia and the porphyrias. *The Metabolic & Molecular Bases of Inherited Disease*. C.R. Scriver, A.L. Beaudet, W.S. Sly & D. Valle, McGraw-Hill Medical Publishing Division, New York, pp. 2991-3062.

Arnesano, F., Banci, L., Bertini, I., Ciofi-Baffoni, S., Woodyear, T.L., Johnson, C.M. and Barker, P.D. (2000) Structural consequences of b- to c-type heme conversion in oxidized *Escherichia coli* cytochrome b562. *Biochemistry*, **39**, 1499-1514

Arnold K., Bordoli L., Kopp J., and Schwede T. (2006). The SWISS-MODEL Workspace: A web-based environment for protein structure homology modelling. *Bioinformatics*, **22**,195-201.

Askwith, C., Eide, D., Van Ho, A., Bernard, P.S., Li, L., Davis-Kaplan, S., Sipe, D.M. and Kaplan, J. (1994) The FET3 gene of *S.cerevisiae* encodes a multicopper oxidase required for ferrous iron uptake. *Cell*. **76**(2),403-410.

Aubert, C., Giudici-Orticoni, M.T., Czjzek, M., Haser, R., Bruschi, M. and Dolla, A. (1998) Structural and kinetic studies of the Y73E mutant of octaheme cytochrome c3 (Mr = 26 000) from *Desulfovibrio desulfuricans* Norway. *Biochemistry*. **37**, 2120-2130

Badarau, A., Firbank, S.J., McCarthy, A.A., Banfield, M.J. and Dennison, C. (2010) Visualizing the metal-binding versatility of copper trafficking sites. *Biochemistry*. **49**(36),7798-810.

(1) Banci, L., Bertini, I., Ciofi-Baffoni, S., Gonnelli L and Su, X.C. (2003) Structural basis for the function of the N-terminal domain of the ATPase CopA from Bacillus subtilis. *J Biol Chem*. **278**, 50506-505013

(2) Banci, L., Bertini, I. and Del Conte, R. (2003) Solution Structure of Apo CopZ from *Bacillus subtilis*: Further Analysis of the Changes Associated with the Presence of Copper Biochemistry **42,** 13422-13428

Banci, L., Bertini, I., Del Conte, R., Markey, J. and Ruiz-Duenas, F.J. (2001) Copper trafficking: the solution structure of *Bacillus subtilis* CopZ. Biochemistry **40,** 15660-15668

Bardischewsky, F., Quentmeier, A. and Friedrich, C.G. (2006) The flavoprotein SoxF functions in chemotrophic thiosulfate oxidation of Paracoccus pantotrophus in vivo and in vitro. *FEMS Microbiol Lett*. **258**(1,:121-126.

Barros, M.H., Nobrega, F.G. and Tzagoloff, A. (2002) Mitochondrial Ferredoxin Is Required for Heme A Synthesis in *Saccharomyces cerevisiae. J. Biol. Chem*. **277**, 9997-10002

Bento, I., Teixeira, V.H., Baptista, A.M., Soares, C.M., Matias, P.M. and Carrondo, M.A. (2003) Redox-Bohr and other cooperativity effects in the nine-heme cytochrome C from Desulfovibrio desulfuricans ATCC 27774: crystallographic and modeling studies. *J.Biol.Chem*. **278,** 36455-36469

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov L.N. and Bourne, P.E. (2000) The Protein Data Bank. *Nucleic Acids Research*, **28**, 235-242

Borjigin, M., Li, H., Lanz, N.D., Kerby, R.L., Roberts, G.P., Poulos, T.L. (2007) Structure-based hypothesis on the activation of the CO-sensing transcription factor CooA. *Acta Crystallogr.*,Sect.D **63,** 282-287

Bradford MM. (1976) A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal Biochem*. **72**,248-254.

Brito, J.A., Sousa, F.L., Stelter, M., Bandeiras, T.M., Vonrhein, C., Teixeira, M., Pereira, M.M. and Archer, M. (2009) Structural and functional insights into sulfide:quinone oxidoreductase. *Biochemistry*. **48**, 5613-5622

Brown, K.R., Keller, G.L., Pickering, I.J., Harris, H.H., George, G.N. and Winge, D.R. (2002) Structures of the Cuprous-Thiolate Clusters of the Mac1 and Ace1 Transcriptional Activators. *Biochemistry*. **41**, 6469-6476

Brylinski, M. and Skolnick, J. (2008) A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proc Natl Acad Sci USA*.**105**(1),129-134

Bryson, K., McGuffin, L.J., Marsden, R.L., Ward, J.J., Sodhi, J.S. and Jones, D.T. (2005) Protein structure prediction servers at University College London. *Nucleic Acids Res*. **33**(Web Server issue),W36-38.

Buchanan, S.K., Smith, B.S., Venkatramani, L., Xia, D., Esser, L., Palnitkar, M., Chakraborty, R., van der Helm, D. and Deisenhofer, J. (1999) Crystal structure of the outer membrane active transporter FepA from Escherichia coli. *Nat.Struct.Biol*. **6**, 56-63

Bull, P.C. and Cox, D.W. (1994) Wilson disease and Menkes disease: New handles on heavy-metal transport. *Trends Genet*. **10**, 246–252.

Burge, C. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol*. **268**, 78–94.

Bushnell, G.W., Louie, G.V. and Brayer, G.D. (1990) High-resolution three-dimensional structure of horse heart cytochrome c. *J Mol Biol*. **214**, 585-95.

Calderone, V., Dolderer, B., Hartmann, H.J., Echner, H., Luchinat, C., Del Bianco, C., Mangani, S. and Weser, U. (2005) The crystal structure of yeast copper thionein: the solution of a long-lasting enigma. *Proc Natl Acad Sci USA*. **102**(1),51-56.

Chen, Z.W., Koh, M., Van Driessche, G., Van Beeumen, J.J., Bartsch, R.G., Meyer, T.E., Cusanovich, M.A. and Mathews, F.S. (1994) The structure of flavocytochrome c sulfide dehydrogenase from a purple phototrophic bacterium. *Science.* **266**, 430-432

Cherney, M.M., Zhang, Y., Solomonson, M., Weiner, J.H. and James, M.N. (2010) Crystal structure of sulfide:quinone oxidoreductase from Acidithiobacillus ferrooxidans: insights into sulfidotrophic respiration and detoxification. *J.Mol.Biol*. **398**, 292-305

Chillappagari, S., Miethke, M., Trip, H., Kuipers, O.P. and Marahiel, M.A. (2009) Copper acquisition is mediated by YcnJ and regulated by YcnK and CsoR in *Bacillus subtilis*. *J Bacteriol.* **191**(7),2362-2370

Chiu, T. and Goldstein, R. (2000) How to generate improved potentials for protein tertiary structure prediction: A lattice model study. *Proteins,* **41**, 157–163.

Clarke, T.A., Cole, J.A., Richardson, D.J. and Hemmings, A.M. (2007) The crystal structure of the pentahaem c-type cytochrome NrfB and characterization of its solution-state interaction with the pentahaem nitrite reductase NrfA. *Biochem.J.* **406,** 19-30

Coates, J.D., Phillips, E.J., Lonergan, D.J., Jenter, H. and Lovley, D.R. (1966) Isolation of *Geobacter* species from diverse sedimentary environments. *Appl Environ Microbiol*. **62**(5),1531-1536.

Cobine, P., Wickramasinghe, W.A., Harrison, M.D., Weber, T., Solioz, M. and Dameron, C.T. (1999) The *Enterococcus hirae* copper chaperone CopZ delivers copper(I) to the CopY repressor. *FEBS Lett*. **445**(1),27-30.

Cobine, P.A., George, G.N., Jones, C.E., Wickramasinghe, W.A., Solioz, M. and Dameron, C.T. (2002) Copper transfer from the Cu(I) chaperone, CopZ, to the repressor, Zn(II)CopY: metal coordination environments and protein interactions. *Biochemistry*. **41**(18),5822-5829.

Collaborative Computational Project, Number 4. (1994). The CCP4 Suite: Programs for Protein Crystallography. *Acta Cryst*. **D50**, 760-763

Crofts, A.R., Lhee, S., Crofts, S.B., Cheng, J. and Rose, S. (2006). Proton pumping in the bc1 complex: a new gating mechanism that prevents short circuits. *Biochim Biophys Acta*., **1757**, 1019-10134

Cudney, R., Patel, S., Weisgraber, K., Newhouse, Y. and Macpherson, A. (1994) Screening and optimization strategies for macromolecular crystal growth. *Acta Crystallogr.* **D50**, 414-423

Cukier, R.I. (2004) A molecular dynamics study of water chain formation in the proton-conducting K channel of cytochrome *c* oxidase. *Biochimica et Biophysica Acta (BBA) – Bioenergetics*, **1706(1-2)**, 134-146

Czjzek, M., Elantak, L., Zamboni, V., Morelli, X., Dolla, A., Guerlesquin, F. and Bruschi, M. (2002) The crystal structure of the hexadeca-heme cytochrome Hmc and a structural model of its complex with cytochrome c(3). *Structure.* **10**, 1677-1686

Czjzek, M., Payan, F., Guerlesquin, F., Bruschi, M. and Haser, R. (1994) Crystal structure of cytochrome *c*3 from Desulfovibrio desulfuricans Norway at 1.7 A resolution. *J Mol Biol*. **243**, 653-67

Das, A., Trammell, S.A. and Hecht, M.H. (2006) Electrochemical and ligand binding studies of a de novo heme protein. *Biophys. Chemist*, **123**, 102-112

DeLano, W.L. (2002) The PyMOL Molecular Graphics System. DeLano Scientific, Palo Alto, CA, USA. http://www.pymol.org (Last accessed 15th September 2008)

Delcher,A.L., Harmon,D., Kasif,S., White,O. and Salzberg,S.L. (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res*. **27**, 4636–4641.

DeSilva, T.M., Veglia, G. and Opella, S.J. (2005) Solution structures of the reduced and Cu(I) bound forms of the first metal binding sequence of ATP7A associated with Menkes disease. *Proteins* **61,** 1038-1049

Dill, K.A., Ozkan, S.B., Weikl, T.R., Chodera, J.D. and Voelz, V.A. (2007) The protein folding problem: when will it be solved? *Curr Opin Struct Biol*. **17**, 342-346

Ding, Y.H., Hixson, K.K., Giometti, C.S., Stanley, A., Esteve-Núñez, A., Khare, T., Tollaksen, S.L., Zhu, W., Adkins, J.N., Lipton, M.S., Smith, R.D., Mester, T. and Lovley, D.R. (2006) The proteome of dissimilatory metal-reducing microorganism *Geobacter sulfurreducens* under various growth conditions. *Biochim. Biophys. Acta* **1764**(7), 1198–1206.

Drescher, D.F., Follmann, H. and Häberlein, .I (1998) Sulfitolysis and thioredoxin-dependent reduction reveal the presence of a structural disulfide bridge in spinach chloroplast fructose-1,6-bisphosphatase. *FEBS Letters*, **424**(1-2), 109-112

Dunbrack, R.L. Jr and Karplus, M. (1993) Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J Mol Biol*. **230**, 543-574

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

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*. **5**,113.

Einsle, O., Andrade, S.L., Dobbek, H., Meyer, J. and Rees, D.C. (2007) Assignment of individual metal redox states in a metalloprotein by crystallographic refinement at multiple X-ray wavelengths. *J Am Chem Soc.* **129**, 2210-2211

Emsley, P. and Cowtan, K. (2004) Coot: Model-Building Tools for Molecular Graphics *Acta Cryst*. **D60,** 2126-2132

Environmental Literacy Council. (2006). Sulfur Cycle. Available: http://www.enviroliteracy.org/article.php/1348.html (Last accessed 16th Dec 2010)

Esteve-Núñez, A., Sosnik, J., Visconti, P. and Lovley, D.R. (2008) Fluorescent properties of c-type cytochromes reveal their potential role as an extracytoplasmic electron sink in *Geobacter sulfurreducens*. *Environ Microbiol*. **10**(2),497-505.

Eswar, N., Webb, B., Marti-Renom, M.A., Madhusudhan, M.S., Eramian, D., Shen, M.Y., Pieper, U. and Sali, A. (2006) Comparative protein structure modeling using Modeller. Curr Protoc Bioinformatics. Chapter 5:Unit 5.6.

Evans, P.(2005) Scaling and assessment of data quality. *Acta Crystallogr D Biol Crystallogr.* **62**(1),72-82

Fermi, G., Perutz, M.F., Shaanan, B. and Fourme, R. (1984). The crystal structure of human deoxyhaemoglobin at 1.74 Å resolution. *J Mol Biol*. **175**, 159-174.

Fidelis K, Stern PS, Bacon D, Moult J. (1994) Comparison of systematic search and database methods for constructing segments of protein structure. *Protein Eng.,* **7**, 953-960

Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunesekaran, P., Ceric, G., Forslund, K.,Holm, L., Sonnhammer, E.L., Eddy, S.R. and Bateman, A.(2010) The Pfam protein families database. *Nucleic Acids Research*. Database Issue **38**,D211-222

Fiser A, Do RK, Sali A. (2000) Modelling of loops in protein structures. *Protein Sci.*, **9**, 1753-1773

Friedrich, C.G., Bardischewsky, F., Rother, D., Quentmeier, A. and Fischer, J. (2005) Prokaryotic sulfur oxidation. *Curr Opin Microbiol*. **8**(3),253-259.

Friedrich, C.G., Rother, D., Bardischewsky, F., Quentmeier, A. and Fischer, J. (2001) Oxidation of reduced inorganic sulfur compounds by bacteria: emergence of a common mechanism? *Appl Environ Microbiol*. **67**(7),2873-2882.

Friedrich, C.G.,Rotsaert, F.A., Covian, R. and Trumpower, B.L. (2008) Mutations in cytochrome b that affect kinetics of the electron transfer reactions at center N in the yeast cytochrome bc1 complex. *Biochim Biophys Acta*. **1777**(3),239-249

Frishman, D., Argos, P. (1995) Knowledge-Based Protein Secondary Structure Assignment. *Proteins: Structure, Function, and Genetics,* **23,** 566-579

Furuyama, K., Kaneko, K. and Vargas, P.D. V (2007) Heme as a magnificent molecule with multiple missions: heme determines its own fate and governs cellular homeostasis. *Tohoku J Exp Med* **213,** 1-16.

Georgescu R.E., Alexov E.G., Gunner M.R.(2002). Combining conformational flexibility and continuum electrostatics for calculating pKa's in proteins. *Biophys J.* **83**, 1731-1748.

Gilis, D. and Rooman, M. (1996) Stability changes upon mutation of solvent-accessible residues in proteins evaluated by database-derived potentials. *J. Mol. Biol.* **257**, 1112–1126.

Ginalski, K. (2006) Comparative modeling for protein structure prediction. *Curr Opin Struct Biol.,* **16**, 172-177

Giometti, C.S. (2006) Tale of two metal reducers: comparative proteome analysis of *Geobacter sulferreducens* PCA and *Shewanella oneidensis* MR-1.*Methods Biochem Anal*. **49**, 97-111.

Gitschier, J., Moffat, B., Reilly, D., Wood, W.I. and Fairbrother, W.J. (1998) Solution structure of the fourth metal-binding domain from the Menkes copper-transporting ATPase. *Nat.Struct.Biol.* **5,** 47-54

Glaser, F., Pupko, T., Paz, I., Bell, R.E., Bechor-Shental, D., Martz E. and Ben-Tal N. (2003) ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics*. **9**(1),163-164.

Griesbeck, C., Schütz, M., Schödl, T., Bathe, S., Nausch, L., Mederer, N., Vielreicher, M. and Hauska, G. (2002) Mechanism of sulfide-quinone reductase investigated using site-directed mutagenesis and sulfur analysis. *Biochemistry*. **41**(39),11552-11565.

Häberlein, .I (1994) Structure requirements for disulfide bridge sulfitolysis of oxidized *Escherichia coli* thioredoxin studied by fluorescence spectroscopy. European Journal of *Biochemistry*, **223**, 473–479.

Halliwell, B. and Gutteridge, J. M. (1990) Role of free radicals and catalytic metal ions in human disease: an overview. *Methods Enzymol*. **186**, 1–85

Harada, S., Sasaki, T., Shindo, M., Kido, Y., Inaoka, D.K., Omori, J., Osanai, A., Sakamoto, K., Mao, J., Matsuoka, S., Inoue, M., Honma, T., Tanaka, A. and Kita, K. Crystal structure of porcine heart mitochondrial complex II bound with N-Biphenyl-3-yl-2-trifluoromethyl-benzamide. PDB ID: 3ABV

Harrison, M.D., Jones, C.E., Solioz, M. and Dameron, C.T. (2000) Intracellular copper routing: the role of copper chaperones. *Trends Biochem Sci*. **25**, 29-32.

Hartshorne, R.S., Reardon, C.L., Ross, D., Nuester, J., Clarke, T.A., Gates, A.J., Mills, P.C., Fredrickson, J.K., Zachara, J.M., Shi, L., Beliaev, A.S., Marshall, M.J., Tien, M., Brantley, S., Butt, J.N. and Richardson, D.J. (2009) Characterization of an electron conduit between bacteria and the extracellular environment. *PNAS*, **106**(52), 22169-22174

Hayward, S. (2006): Lecture 5 – Fold recognition. Lecture presented as part of the MSc course “BIO-M532 Protein structure, prediction & modelling” at the University of East Anglia.

Hearnshaw, S., West, C., Singleton, C., Zhou, L., Kihlken, M.A., Strange, R.W., Le Brun, N.E. and Hemmings, A.M. (2009) A tetranuclear Cu(I) cluster in the metallochaperone protein CopZ. *Biochemistry*. **48**(40),9324-9326.

Heidelberg, J.F., Paulsen, I.T., Nelson, K.E., Gaidos, E.J., Nelson, W.C., Read, T.D., Eisen, J.A., Seshadri, R., Ward, N., Methe, B., Clayton, R.A., Meyer, T., Tsapin, A., Scott, J., Beanan, M., Brinkac, L., Daugherty, S., DeBoy, R.T., Dodson, R.J., Durkin, A.S., Haft, D.H., Kolonay, J.F., Madupu, R., Peterson, J.D., Umayam, L.A., White, O., Wolf, A.M., Vamathevan, J., Weidman, J., Impraim, M., Lee, K., Berry, K., Lee, C., Mueller, J., Khouri, H., Gill, J., Utterback, T.R., McDonald, L.A., Feldblyum, T.V., Smith, H.O., Venter, J.C., Nealson, K.H. and Fraser, C.M. (2002) Genome sequence of the dissimilatory metal ion-reducing bacterium Shewanella oneidensis. *Nat Biotechnol*. **20**(11),1118-11123

Henriksen, A., Smith, A.T. and Gajhede, M. (1999). The structures of the horseradish peroxidase C-ferulic acid complex and the ternary complex with cyanide suggest how peroxidases oxidize small phenolic substrates. *J Biol Chem*., **274**, 35005-35011

Hersleth, H.P., Ryde, U., Rydberg, P., Görbitz, C.H. and Andersson, K.K. (2006) Structures of the high-valent metal-ion haem-oxygen intermediates in peroxidases, oxygenases and catalases. *J Inorg Biochem*. **100**, 460-476

Holm L, Rosenström P. (2010) Dali server: conservation mapping in 3D. *Nucleic Acids Res*. **38**(Web Server issue): W545–W549.

Horrigan, F.T., Heinemann, S.H. and Hoshi, T. (2005). Heme regulates allosteric activation of the Slo1 BK channel. *J Gen Physiol*, **126**, 7-21

Hu, V.W., Chan, S.I. and Brown, G.S. (1977) X-ray absorption edge studies on oxidized and reduced cytochrome *c* oxidase. *Proc Natl Acad Sci USA*. **74**, 3821-3825

Hughey R, Krogh A. Hidden Markov models for sequence analysis: extension and analysis of the basic method. Comput Appl Biosci.**12**,95–107

Hung, I.H., Casareno, R.L., Labesse, G., Mathews, F.S. and Gitlin, J.D. (1998) HAH1 is a copper-binding protein with distinct amino acid residues mediating copper homeostasis and antioxidant defense. *J. Biol. Chem*. **273**, 1749–1754.

Ihaka, R., and Gentleman, R. (1996) R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics.* **5 (3)**, 299–314

Inês A. C. Pereira, António V. Xavier (2006) Multi-Heme Cytochromes & Enzymes. *Encyclopaedia of Inorganic Chemistry*, 1-17 (Online Book)

Iwata, S., Lee, J.W., Okada, K., Lee, J.K., Iwata, M., Rasmussen, B., Link, T.A., Ramaswamy, S. and Jap, B.K. (1998) Complete structure of the 11-subunit bovine mitochondrial cytochrome *bc*1 complex. *Science*, **281**, 64-71

J.D. Peterson, L.A. Umayam, T.M. Dickinson, E.K. Hickey and O. White. 2001 The Comprehensive Microbial Resource. *Nucleic Acids Research*, **29**(1), 123-125.

Jacobson, M,P., Pincus, D.L., Rapp, C.S., Day, T.J., Honig, B., Shaw, D.E. and Friesner, R.A. (2004) A hierarchical approach to all-atom protein loop prediction. *Proteins*. **55**, 351-367

Jancarik, J. and Kim, S-H.J. (1991) Sparse matrix sampling: a screening method for crystallization of proteins. *J. Appl. Cryst.* **24**, 409-411.

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S, and Madden, T.L. (2008) NCBI BLAST: a better web interface. *Nucleic Acids Res*. **36**(Web Server issue):W5-9

Jones, D.T. (1994) THREADER Info Page. http://bioinf.cs.ucl.ac.uk/threader/ (Last accessed 23rd October 2007)

Jones, D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol*. **292**(2),195-202

Jones, D.T., Taylor, W.R. and Thornton, J.M. (1992) A new approach to protein fold recognition. *Nature*. **358**, 86-89.

Kabsch, W. (1988). Evaluation of single-crystal X-ray diffraction data from a position-sensitive detector. *J.Appl.Cryst*. **21**, 916-924

Kadziola, A., Larsen, S., Christensen, H.M., Karlsson, J.J. and Ulstrup, J. (1995) Crystallization and preliminary crystallographic investigations of cytochrome c4 from *Pseudomonas stutzeri. Acta Crystallogr D Biol Crystallogr*. **51**(6), 1071-1073.

Karplus, K., Barrett, C. and Hughey, R. (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics*. **14**, 846-856

Kau, L.S., Spira-Solomon, D.J., Penner-Hahn, J.E., Hodgson, K.O. and Solomon, E.I. (1987) X-ray absorption edge determination of the oxidation state and coordination number of copper. Application to the type 3 site in *Rhus vernicifera* laccase and its reaction with oxygen. *J. Am. Chem. Soc*. **109**, 6433–6442

Keller, G., Bird, A. and Winge, D.R. (2000) Independent metalloregulation of Ace1 and Mac1 in Saccharomyces cerevisiae. *Eukaryot Cell*. **4**, 1863-1871

Kelley, L.A. and Sternberg, M.J. (2009) Protein structure prediction on the Web: a case study using the Phyre server. *Nat Protoc*. **4**(3),363-371.

Klimmek, O., Kreis, V., Klein, C., Simon, J., Wittershagen, A. and Kroger, A. (1998) The function of the periplasmic Sud protein in polysulfide respiration of Wolinella succinogenes. *Eur. J. Biochem*. **253**,263-269

Klimmek, O., Stein, T., Pisa, R., Simon, J. and Kroger, A. (1999) The single cysteine residue of the Sud protein is required for its function as a polysulfide-sulfur transferase in Wolinella succinogenes. *Eur. J. Biochem*. **263**,79-84

Klomp, L.W., Lin, S.J., Yuan, D.S., Klausner, R.D., Culotta, V.C. and Gitlin, J.D. (1997) Identification and functional expression of HAH1, a novel human gene involved in copper homeostasis. *J.Biol.Chem*. **272**, 9221–9226.

Komori, H., Inagaki, S., Yoshioka, S., Aono, S. and Higuchi, Y. (2007). Crystal structure of CO-sensing transcription activator CooA bound to exogenous ligand imidazole. *J Mol Biol.*, **367**, 864-871

Krissinel, E. and Henrick, K. (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Cryst*. **D60**, 2256-2268

Kuhlman, B., Dantas, G., Ireton, G.C., Varani, G., Stoddard, B.L. and Baker, D. (2003) Design of a novel globular protein fold with atomic-level accuracy. *Science*, **302**, 1364-1368.

Laan, W., van der Horst, M.A., van Stokkum, I.H.M. and Hellingwerf, K.J. (2003) Initial characterization of the primary photochemistry of AppA, a blue-light-using flavin adenine dinucleotide-domain containing transcriptional antirepressor protein from Rhodobacter sphaeroides: a key role for reversible intramolecular proton transfer from the flavin adenine dinucleotide chromophore to a conserved tyrosine? *Photochem. Photobiol*. **78**, 290–297.

Lacapère, J.J., Pebay-Peyroula, E., Neumann, J.M. and Etchebest, C. (2007) Determining membrane protein structures: still a challenge! *Trends Biochem Sci*. **32**, 259-270

Lanzilotta, W.N., Schuller, D.J., Thorsteinsson, M.V., Kerby, R.L., Roberts, G.P. and Poulos, T.L. (2000) Structure of the CO sensing transcription activator CooA. *Nat Struct Biol*., **7**, 876-880

Laskowski, R.A., Macarthur, M.W., Moss, D.S. and Thornton J.M. (1993) PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst*. **26**, 283-291.

Lathrop, R.H. and Smith, T.F. (1996) Global optimum protein threading with gapped alignment and empirical pair score functions. *J Mol Biol*. **255**, 641-65.

Lemer, C.M., Rooman, M.J. and Wodak, S.J. (1995) Protein structure prediction by threading methods: evaluation of current techniques. *Proteins*. **23**, 337-355

Lesk, A.M. and Chothia, C. (1986) The response of protein structures to amino-acid sequence changes. *Philos. Trans. R. Soc. Lond B Biol. Sci.,* **317**, 345–356.

Levinthal, C. (1968). Are there pathways for protein folding? *J. Chim. Phys*. **65**, 44-45.

Leys, D., Meyer, T.E., Tsapin, A.I., Nealson, K.H., Cusanovich, M.A. and Van Beeumen, J.J. (2002) Crystal structures at atomic resolution reveal the novel concept of "electron-harvesting" as a role for the small tetraheme cytochrome c. *Biol.Chem*. **277**, 35703-35711

Lin, S.J. and Culotta, V.C. (1995) The ATX1 gene of *Saccharomyces cerevisiae* encodes a small metal homeostasis factor that protects cells against reactive oxygen toxicity. *Proc Natl Acad Sci USA.* **92**(9),3784-3788.

Lin, S.J., Pufahl, R.A., Dancis, A., O'Halloran, T.V. and Culotta, V.C. (1997) A role for the *Saccharomyces cerevisiae* ATX1 gene in copper trafficking and iron transport. *J Biol Chem*. **272**(14),9215-9220.

Liu, T., Ramesh, A., Ma, Z., Ward, S.K., Zhang, L., George, G.N., Talaat, A.M., Sacchettini, J.C. and Giedroc, D.P.(2007) CsoR is a novel Mycobacterium tuberculosis copper-sensing transcriptional regulator. *Nat.Chem.Biol*. **3**, 60-68

López-Barneo, J. and Castellano, A. (2005) Multiple facets of maxi-k+ channels: the heme connection. *J Gen Physiol*., **126**, 1-5

Lovell, S.C., Word, J.M., Richardson, J.S. and Richardson, D.C. (2000) The penultimate rotamer library. *Proteins*. **40**, 389-408

Lukashin, A.V. and Borodovsky, M. 1998. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res*. **26**, 1107– 1115.

Lutsenko, S., Barnes, N.L., Bartee, M.Y. and Dmitriev, O.Y. (2007) Function and regulation of human copper-transporting ATPases. *Physiol. Rev*. **87**, 1011–104

M.A. Kihlken, A.P. Leech and N.E. Le Brun (2002) Copper-mediated dimerization of CopZ, a predicted copper chaperone from *Bacillus subtilis. Biochem. J*. **368**, 729–739

MacKerell, A.D. Jr, Bashford, D., Bellott, M., Dunbrack, R. Jr, Evanseck, J., Field, M., Fischer, S., Gao, J., Guo, H., and Ha, S. (1998) All-atom empirical potential for molecular modelling and dynamics studies of proteins. *J Phys Chem B,* **102**, 3586–3616.

Madera, M. and Gough, J. (2002) A comparison of profile hidden Markov model procedures for remote homology detection. *Nucleic Acids Res.* **30**(19), 4321-4328.

Magnani, D. and Solioz, M. (2005) Copper chaperone cycling and degradation in the regulation of the cop operon of *Enterococcus hirae*. *BioMetals* **18**, 407–412

Maines, M.D. (1997) THE HEME OXYGENASE SYSTEM: A Regulator of Second Messenger Gases *Annu. Rev. Pharmacol. Toxicol.,* **37**, 517–554

Maines, M.D., Trakshel, G.M. and Kutty, R.K. (1986) Characterization of two constitutive forms of rat liver microsomal heme oxygenase. Only one molecular species of the enzyme is inducible. *J. Biol. Chem*., **261**, 411-419

Marcia M, Ermler U, Peng G, Michel H. (2010) A new structure-based classification of sulfide:quinone oxidoreductases. *Proteins* **78**(5), 1073-1083

Marcia, M., Ermler, U., Peng, G.H. and Michel, H. (2009) The structure of *Aquifex aeolicus* sulfide:quinone oxidoreductase, a basis to understand sulfide detoxification and respiration. *Proc.Natl.Acad.Sci.USA* **106**, 9625-9630

Mccoubrey Jr, W.K., Huang, T.J. and Maines, M.D. (1997) Isolation and Characterization of a cDNA from the Rat Brain that Encodes Hemoprotein Heme Oxygenase-3. *Eur J Biochem*, **247**, 725-732

McDowell, S.E., Spacková, N., Sponer, J. and Walter, N.G. (2007) Molecular dynamics simulations of RNA: an in silico single molecule approach. *Biopolymers*. **85**, 169-184

Melo, F. and Feytmans, E. (1997) Novel knowledge-based mean force potential at atomic level. *J. Mol. Biol.* **267**, 207–222.

Melo, F., Sánchez, R. and Sali, A. (2002) Statistical potentials for fold assessment. *Protein Sci*., **11**, 430-448.

Meyer,F., Paarmann,D., D'Souza,M., Olson,R., Glass,E.M., Kubal,M., Paczian,T., Rodriguez,A., Stevens,R., Wilke,A. et al. (2008) The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*. **9**, 386.

Mirny, L.A. and Shakhnovich, E.I. (1998) Protein structure prediction by threading. Why it works and why it does not. *J Mol Biol*., **283**, 507-526

Mirny, L.A., Finkelstein, A.V. and Shakhnovich, E.I. (2000) Statistical significance of protein structure prediction by threading. *Proc Natl Acad Sci USA.,* **97**, 9978-9983

Misura, K.M. and Baker, D. (2005) Progress and challenges in high resolution refinement of protein structure models. *Proteins*, **59**, 15-29.

Monod, J., Wyman, J. and Changeux, J.-P. (1965) On the nature of allosteric transitions: a plausible model. *J. Mol. Biol.* **12**, 88–118.

Moraes, C.T., Diaz, F. and Barrientos, A. (2004) Defects in the biosynthesis of mitochondrial heme *c* and heme *a* in yeast and mammals. *Biochimica et Biophysica Acta*, **1659**, 153– 159

Morgado, L., Fernandes, A.P., Londer, Y.Y., Pokkuluri, P.R., Schiffer, M. and Salgueiro, C.A. (2009) Thermodynamic characterization of the redox centres in a representative domain of a novel c-type multihaem cytochrome. *Biochem J*. **420**(3), 485-492.

Morgado, L., Bruix, M., Orshonsky, V., Londer, Y.Y., Duke, N.E., Yang, X., Pokkuluri, P.R., Schiffer, M. and Salgueiro, C.A. (2008) Structural insights into the modulation of the redox properties of two Geobacter sulfurreducens homologous triheme cytochromes. *Biochim.Biophys.Acta* **1777,** 1157-1165

Mortuza, G.B., Haire, L.F., Stevens, A., Smerdon, S.J., Stoye, J.P. and Taylor, I.A. (2004) High-resolution structure of a retroviral capsid hexameric amino-terminal domain *Nature*. **431**, 481-485.

Moser, C.C., Page, C.C. and Dutton, P.L. (2006) Darwin at the molecular scale: selection and variance in electron tunnelling proteins including cytochrome c oxidase. *Philos Trans R Soc Lond B Biol Sci*. **361**, 1295-1305

Moult, J. (1997) Comparison of database potentials and molecular mechanics force fields. *Curr Opin Struct Biol*., **7**, 194-199.

Moult, J., Fidelis, K., Kryshtafovych, Rost, B. and Tramontano, A. (2009) Critical assessment of methods of protein structure prediction—Round VII. *Proteins: Structure, Function, and Bioinformatics.* **77**, 1-4

Muller, F.L., Song, W., Liu, Y., Chaudhuri, A., Pieke-Dahl, S., Strong, R., Huang, T.T., Epstein, C.J., Roberts, L.J. 2nd, Csete, M., Faulkner, J.A. and Van Remmen, H. (2006) Absence of CuZn superoxide dismutase leads to elevated oxidative stress and acceleration of age-dependent skeletal muscle atrophy. *Free Radical Biology & Medicine*. **40**, 1993-2004

Multhaup, G., Strausak, D., Bissig, K.D. and Solioz, M. (2001) Interaction of the CopZ Copper Chaperone with the CopA Copper ATPase of *Enterococcus hirae* Assessed by Surface Plasmon Resonance. *Biochem Biophys Res Commun*. **288**, 172-177

Murshudov, G.N., Vagin, A.A. and Dodson, E.J. (1997) Refinement of Macromolecular Structures by the Maximum-Likelihood Method. *Acta Cryst*. **D53**, 240-255.

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

Myers, C.R. and Nealson, K.H. (1988) Bacterial manganese reduction and growth with manganese oxide as the sole electron acceptor. *Science.* **240**, 1319–1321.

Needleman, S.B. and Wunsch, C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*. **48**, 443–453

Nienhaus, K., Lamb, D.C., Deng, P. and Nienhaus, G.U. (2002) The effect of ligand dynamics on heme electronic transition band III in myoglobin. *Biophys J*. **82**, 1059-1067

Noguchi,H., Park,J. and Takagi,T. (2006) MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res*. **34**, 5623–5630.

Notredame, C., Higgins, D.G. and Heringa, J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*. **302**(1), 205-217.

O'Brian, M.R. (1996) Heme synthesis in the rhizobium-legume symbiosis: a palette for bacterial and eukaryotic pigments. *J Bacteriol*. **178**, 2471–2478.

Omura, T. (2005) Heme-thiolate proteins. *Biochem Biophys Res Commun*. **338**, 404-9

Oudgenoeg, G., Dirksen, E., Ingemann, S., Hilhorst, R., Gruppen, H., Boeriu, C.G., Piersma, S.R., van Berkel, W.J., Laane, C. and Voragen, A.G. (2002) Horseradish peroxidase-catalyzed oligomerization of ferulic acid on a template of a tyrosine-containing tripeptide. *J Biol Chem.*, **277**, 21332-21340

Oyedotun, K.S., Sit, C.S. and Lemire, B.D. (2007) The Saccharomyces cerevisiae succinate dehydrogenase does not require heme for ubiquinone reduction. *Biochim Biophys Acta*. **1767**(12), 1436-1445

P.A. Van Der Merwe (2001) Protein-Ligand Interactions: A Practical Approach – Chapter 6 - Surface plasmon resonance, *Oxford University Press*

Page, C.C., Moser, C.C. and Dutton, P.L. (2003) Mechanism for electron transfer within and between proteins. *Curr Opin Chem Biol*., **7**, 551-556

Page, C.C., Moser, C.C., Chen, X. and Dutton, P.L. (1999). Natural engineering principles of electron tunnelling in biological oxidation-reduction. *Nature,* **402,** 47-52.

Paixao, V.B., Salgueiro, C.A., Brennan, L., Reid, G.A., Chapman, S.K. and Turner, D.L. (2008) The solution structure of a tetraheme cytochrome from Shewanella frigidimarina reveals a novel family structural motif. *Biochemistry*. **47**: 11973-11980

Panchenko, A., Marchler-Bauer, A. and Bryant, S. (2000) Combination of threading potentials and sequence profiles improves fold recognition. *J. Mol. Biol.* **296**, 1319–1331.

Pande, V. and Stanford University (2000): Folding@home – Main. http://folding.stanford.edu/ (Last accessed 23rd October 2007)

Papa, S., Capitanio, N. and Villani, G. (1998) A cooperative model for proton motive heme-copper oxidases. The role of heme a in the proton pump of cytochrome c oxidase. *FEBS Letters*, **439**, 1-8

Pattarkine, M.V., Tanner, J.J., Bottoms, C.A., Lee, Y.H., Wall, J.D. 2006 Desulfovibrio desulfuricans G20 tetraheme cytochrome structure at 1.5 Angstrom and cytochrome interaction with metal complexes. *J.Mol.Biol*. **358**, 1314-1327

Perrakis, A., Sixma, T.K., Wilson, K.S. and Lamzin, V.S. (1997) wARP: improvement and extension of crystallographic phases by weighted averaging of multiple-refined dummy atomic models. *Acta Cryst. Sect. D Biol, Cryst*. **53**, 240-255

Perutz, M. F. (1972). Haem–haem interaction. *Nature*, **237**, 495–499.

Peterson, C.W., Narula, S.S. and Armitage, I.M. (1996) 3D solution structure of copper and silver-substituted yeast metallothioneins. *FEBS Lett*. **379**(1):85-93.

Petr, M., Petr, B. and Jirí, S. (2008) Multicriteria Tunnel Computation. Computer Graphics and Imaging, Innsbruck, Austria, 2008.

Pitera, J.W. and Swope, W. (2003) Understanding folding and design: replica-exchange simulations of ‘‘Trp-cage’’ miniproteins. *Proc Natl Acad Sci USA,* **100**, 7587-7592.

Pokkuluri, P.R., Londer, Y.Y., Duke, N.E.C., Erickson, J., Pessanha, M., Salgueiro, C.A. and Schiffer, M. (2004) Structure of a novel c7-type three-heme cytochrome domain from a multidomain cytochrome c polymer. *Protein Sci.* **13,** 1684-1692

Polyakov, K.M., Boyko, K.M., Tikhonova, T.V., Slutsky, A., Antipov, A.N., Zvyagilskaya, R.A., Popov, A.N., Bourenkov, G.P., Lamzin,V.S. and Popov, V.O. (2009) High-resolution structural of a novel octaheme cytochrome c nitrite reductase from the haloalkaliphilic bacterium *Thiolkalivibrio nitratireducens J.Mol.Biol*. **389**(5), 846-862

Ponstingl, H., Henrick, K. and Thornton, J. M. (2000) Discriminating between homodimeric and monomeric proteins in the crystalline state. *Prot. Struct. Funct. Genet*. **41**, 47–57

Poulos, T.L. (2005) Structural and functional diversity in heme mono-oxygenases. *Drug Metab Dispos*, **33**, 10-18

Poulos, T.L. (2006) The Janus nature of heme. *Nat. Prod. Rep.*, **24**, 504 – 510

Pufahl, R.A., Singer, C.P., Peariso, K.L., Lin, S.J., Schmidt, P.J., Fahrni, C.J., Culotta, V.C., Penner-Hahn, J.E. and O'Halloran, T.V. (1997) Metal ion chaperone function of the soluble Cu(I) receptor Atx1. *Science*. **278**, 853-856

Puranik, M., Nielsen, S.B., Youn, H., Hvitved, A.N., Bourassa, J.L., Case, M.A., Tengroth, C., Balakrishnan, G., Thorsteinsson, M.V., Groves, J.T., McLendon, G.L., Roberts, G.P., Olson, J.S. and Spiro, T.G. (2004) Dynamics of carbon monoxide binding to CooA. *J Biol Chem*., **279**, 21096-21108

Qian, B., Raman, S., Das, R., Bradley, P., McCoy, A.J., Read, R.J. and Baker, D. (2007) High-resolution structure prediction and the crystallographic phase problem. *Nature*. [Epub ahead of print]

Quentmeier, A., Hellwig, P., Bardischewsky, F., Wichmann, R. and Friedrich, C.G. (2004) Sulfide dehydrogenase activity of the monomeric flavoprotein SoxF of Paracoccus pantotrophus. *Biochemistry*. **43**(46), 14696-14703.

R Development Core Team (2010) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Avaliable at: http://www.R-Project.org (Last accessed 16th Dec 2010)

Rabiner, L.R. (1989) A tutorial on hidden Markov models and selected applications in speech Recognition. *Proc. IEEE*, **77**(2), 257-285.

Radford, D.S., Kihlken, M.A., Borrelly, G.P., Harwood, C.R., Le Brun, N.E. and Cavet, J.S. (2003) CopZ from Bacillus subtilis interacts in vivo with a copper exporting CPx-type ATPase CopA. *FEMS Microbiol Lett*. **220**(1), 105-112.

Rae, T.D., Schmidt, P.J., Pufahl, R.A., Culotta, V.C. and O'Halloran, T.V. (1999) Undetectable intracellular free copper: the requirement of a copper chaperone for superoxide dismutase. *Science*. **284**, 805-808.

Rawlings, D.E. (2002) Heavy metal mining using microbes. *Annu Rev Microbiol*. **56**, 65-91

Reynolds, C., Damerell, D. and Jones, S. (2009) ProtorP: a protein-protein interaction analysis server. *Bioinformatics*. **25**(3), 413-414.

Rhee, Y.M. and Pande, V.S. (2003) Multiplexed-replica exchange molecular dynamics method for protein folding simulation. *Biophys J.* **84**, 775-86

Rho, M., Tang, H. and Ye, Y. (2010) FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res*. **38**(20), e191.

Rich, P.R. (2004). The quinone chemistry of the *bc* complexes. *Biochim Biophys Acta.* **1658**, 165-171

Roberts, B.R., Wood, Z.A., Jonsson, T.J., Poole, L.B. and Karplus, P.A. (2005) Oxidized and synchrotron cleaved structures of the disulfide redox center in the N-terminal domain of *Salmonella typhimurium* AhpF *Protein Sci*. **14**(9), 2414–2420.

Rodrigues, M.L., Oliveira, T.F., Pereira, I.A.C. and Archer, M. (2006) X-ray structure of the membrane-bound cytochrome c quinol dehydrogenase NrfH reveals novel haem coordination. *Embo J*. **25**, 5951-5960

Rohl, C.A., Strauss, C.E., Chivian, D. and Baker, D. (2004) Modeling structurally variable regions in homologous proteins with rosetta. *Proteins*., **55**, 656-677

Rosenzweig, A.C., Huffman, D.L., Hou, M.Y., Wernimont, A.K., Pufahl, R.A. and O`Halloran, T.V. (1999) Crystal structure of the Atx1 metallochaperone protein at 1.02 A resolution. *Structure Fold.Des.* **7,** 605-617

Ross, D.E., Ruebush, S.S., Brantley, S.L., Hartshorne, R.S., Clarke, T.A., Richardson DJ and Tien, M.. (2007) Characterization of protein-protein interactions involved in iron reduction by *Shewanella oneidensis* MR-1. *Appl Environ Microbiol*. **73**(18), 5797-5808.

Rost, B., Schneider, R. and Sander, C. (1997) Protein fold recognition by prediction-based threading. *J Mol Biol*. **270**, 471-480

Rother, D., Henrich, H.J., Quentmeier, A., Bardischewsky, F. and Friedrich, C.G. (2001) Novel genes of the sox gene cluster, mutagenesis of the flavoprotein SoxF, and evidence for a general sulfur-oxidizing system in *Paracoccus pantotrophus* GB17. *J Bacteriol.* **183**(15), 4499-4508.

Rother, D., Orawski, G., Bardischewsky, F. and Friedrich, C.G. (2005) SoxRS-mediated regulation of chemotrophic sulfur oxidation in *Paracoccus pantotrophus*. *Microbiology*. **151**(5), 1707-1716.

Rutter, J., Winge, D.R. and Schiffman, J.D. (2010) Succinate dehydrogenase - Assembly, regulation and role in human disease. *Mitochondrion*. **10**(4), 393-401

Ryter, S.W. and Tyrrell, R.M. (2000) THE HEME SYNTHESIS AND DEGRADATION PATHWAYS: ROLE IN OXIDANT SENSITIVITY Heme Oxygenase has Both Pro- and Antioxidant Properties. *Free Radical Biology & Medicine*, **28**, 289–309

Saiki, K., Mogi, T., Ogura, K., and Anraku, Y. (1993) In vitro heme O synthesis by the cyoE gene product from Escherichia coli *J. Biol. Chem.* **268,** 26041–26044

Salamov, A.A. and Solovyev, V.V. 2000. Ab initio gene finding in Drosophila genomic DNA. *Genome Res*. **10**, 516–522.

Salzberg, S.L., Pertea, M., Delcher, A.L., Gardner, M.J. and Tettelin, H. 1999. Interpolated Markov models for eukaryotic gene finding. *Genomics* **59**, 24–31.

Santos-Silva, T., Dias, J.M., Dolla, A., Durand, M.C., Gonçalves, L.L., Lampreia, J., Moura, I. and Romão, M.J.(2007) Crystal structure of the 16 heme cytochrome from *Desulfovibrio gigas*: a glycosylated protein in a sulphate-reducing bacterium. *J Mol Biol*. **370**(4), 659-673.

Sauvé, V., Bruno, S., Berks, B.C. and Hemmings, A.M. (2007) The SoxYZ complex carries sulfur cycle intermediates on a peptide swinging arm. *J Biol Chem*. **282**(32), 23194-23204.

Scott, E.E., White, M.A., He, Y.A., Johnson, E.F., Stout, C.D. and Halpert, J.R. (2004). Structure of mammalian cytochrome P450 2B4 complexed with 4-(4-chlorophenyl)imidazole at 1.9-A resolution: insight into the range of P450 conformations and the coordination of redox partner binding. *J Biol Chem.* **279**, 27294-27301

Segal, D.J., Crotty, J.W., Bhakta, M.S., Barbas, C.F. and Horton, N.C. (2006) Structure of Aart, a designed six-finger zinc finger peptide, bound to DNA. J.Mol.Biol. 363: 405-421

Serre, L., Rossy, E., Pebay-Peyroula, E., Cohen-Addad, C. and Coves, J. (2004) Crystal Structure of the Oxidized Form of the Periplasmic Mercury-binding Protein MerP from Ralstonia metallidurans CH34 *J.Mol.Biol.* **339,** 161-171

Sharma, S., Cavallaro, G. and Rosato, A. (2010) A systematic investigation of multiheme c-type cytochromes in prokaryotes. *J Biol Inorg Chem*. **15**(4), 559-571.

Sheldrick, G. (1991). XPREP. Space Group Determination and Reciprocal Space Plots. Siemens Analytical X-ray Instruments, Madison, Wisconsin, USA

Sheldrick, G.M. and Schneider, T.R. (1997) SHELXL: high-resolution refinement. *Methods Enzymol*. **277**, 319–343

Shi, L., Squier, T.C., Zachara, J.M. and Fredrickson, J.K. (2007) Respiration of metal (hydr)oxides by *Shewanella* and *Geobacter*: a key role for multihaem c-type cytochromes. *Mol Microbiol*. **65**(1), 12-20.

Shibata, N., Suto, K., Ichimura, E., Yoshimura, K., Muneo, K., Tomigami, S., Morimoto, Y., Ogata, M., Yagi, T., Higuchi, Y. and Yasuoka N. (2004) Crystallization and MAD data collection of high-molecular weight cytochrome c from *Desulfovibrio vulgaris* Miyazaki F. *Protein Pept Lett*. **11**(1), 93-96.

Singleton, C., Hearnshaw, S., Zhou, L., Le Brun, N.E. and Hemmings, A.M. (2009) Mechanistic insights into Cu(I) cluster transfer between the chaperone CopZ and its cognate Cu(I)-transporting P-type ATPase, CopA. *Biochem J*. **424**(3), 347-56.

Smaldone, G.T. and Helmann, J.D. (2007) CsoR regulates the copper efflux operon copZA in *Bacillus subtilis. Microbiology*. **153**(12), 4123-4128.

Solioz, M. and Stoyanov, J.V. (2003) Copper homeostasis in Enterococcus hirae. FEMS *Microbiol Rev*. **27**(2-3), 183-195.

Sonnhammer, E.L., Eddy, S.R., Birney, E., Bateman, A. and Durbin, R. (1998) Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res*. **26**(1), 320-322.

Srinivasan, C., Posewitz, M.C., George, G.N. and Winge, D.R. (1998) Characterization of the copper chaperone Cox17 of *Saccharomyces cerevisiae*. *Biochemistry*. **37**(20), 7572-7577

Steele, R.A. and Opella, S.J. (1997) Structures of the reduced and mercury-bound forms of MerP, the periplasmic protein from the bacterial mercury detoxification system. *Biochemistry* **36,** 6885-6895

Stiburek, L., Hansikova, H., Tesarova, M., Cerna, L. and Zeman, J. (2006) Biogenesis of eukaryotic cytochrome c oxidase. *Physiol Res*. **55**, S27-S41.

Stillman, M.J. (1995) Metallothioneins. *Coord.Chem.Rev*. **144**, 461–511.

Stout, J., Van Driessche, G., Savvides, S.N. and Van Beeumen, J. (2007) X-ray crystallographic analysis of the sulfur carrier protein SoxY from Chlorobium limicola f. thiosulfatophilum reveals a tetrameric structure. *Protein Sci*. **16**, 589-601

Strausak, D. and Solioz, M. (1997) CopY is a copper-inducible repressor of the *Enterococcus hirae* copper ATPases. *J Biol Chem*. **272**(14), 8932-8936.

Sugita, Y., and Okamoto, T. (1999) Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett*. **314**, 141–151.

Sun, J., Hoshino, H., Takaku, K., Nakajima, O., Muto, A., Suzuki, H., Tashiro, S., Takahashi, S., Shibahara, S., Alam, J., Taketo, M.M., Yamamoto, M. and Igarashi, K. (2002) Hemoprotein Bach1 regulates enhancer availability of heme oxygenase-1 gene. *EMBO J.* **21**, 5216–5224.

Sun, F., Huo, X., Zhai, Y., Wang, A., Xu, J., Su, D., Bartlam, M. and Rao Z. (2005) Crystal structure of mitochondrial respiratory membrane protein complex II. *Cell*. **121**(7), 1043-1057.

Suzuki, H., Tashiro, S., Hira, S., Sun, J., Yamazaki, C., Zenke, Y., Ikeda-Saito, M., Yoshida, M. and Igarashi, K. (2004) Heme regulates gene expression by triggering Crm1-dependent nuclear export of Bach1. *EMBO J.,* **23**, 2544–2553.

Svensson-Ek, M., Abramson, J., Larsson, G., Törnroth, S., Brzezinski, P. and Iwata, S. (2002) The X-ray crystal structures of wild-type and EQ(I-286) mutant cytochrome c oxidases from *Rhodobacter sphaeroides*. *J Mol Biol*. **321**, 329-339

Tang, X.D., Xu, R., Reynolds, M.F., Garcia, M.L., Heinemann, S.H. and Hoshi, T. (2003). Haem can bind to and inhibit mammalian calcium-dependent Slo1 BK channels. *Nature*, **425**, 531-535

Tollin, G., Hanson, L.K., Caffrey, M., Meyer, T.E. and Cusanovich, M.A. (1986) Redox pathways in electron-transfer proteins: correlations between reactivities, solvent exposure, and unpaired-spin-density distributions. *Proc Natl Acad Sci U S A*. **83**, 3693-3697

Trumpower, B.L. (1990). Cytochrome *bc*1 complexes of microorganisms. *Microbiol Rev*., **54**, 101-129.

Tsuneshige, A., Park, S. and Yonetani, T. (2002). Heterotropic effectors control the hemoglobin function by interacting with its T and R states—a new view on the principle of allostery. *Biophysical Chemistry*,**98**,49-63

Umhau, S., Fritz, G., Diederichs, K., Breed, J., Welte, W. and Kroneck, P.M. (2001) Three-dimensional structure of the nonaheme cytochrome c from *Desulfovibrio desulfuricans* Essex in the Fe(III) state at 1.89 A resolution. *Biochemistry.* **40**, 1308-1316

Unno, M., Matsui, T. and Ikeda-Saito, M. (2007) Structure and catalytic mechanism of heme oxygenase. *Nat. Prod. Rep.*, **24**, 553 – 570

Vendruscolo, M., Najmanovich, R. and Domany, E. (2000) Can a pairwise contact potential stabilize native protein folds against decoys obtained by threading? *Proteins,* **38**, 134–148.

Walsh, I., Martin, A.J., Mooney, C., Rubagotti, E., Vullo, A. and Pollastri, G. (2009) Ab initio and homology based prediction of protein domains by recursive neural networks. *BMC Bioinformatics*. **10**, 195.

Wang, G. and Dunbrack, Jr. R.L. (2003) PISCES: a protein sequence culling server. *Bioinformatics.*, **19**, 1589-1591

Wang, N., Zhao, X. and Lu, Y. (2005) Role of Heme Types in Heme-Copper Oxidases: Effects of Replacing a Heme b with a Heme o Mimic in an Engineered Heme-Copper Center in Myoglobin. *J. Am. Chem. Soc.,* **127**, 16541 -16547

Weik, M., Ravelli, R.B., Kryger, G., McSweeney, S., Raves, M.L., Harel, M., Gros, P., Silman, I., Kroon, J., and Sussman, J.L. (2000) Specific chemical and structural damage to proteins produced by synchrotron radiation. *Proc. Natl. Acad. Sci*. **97**, 623–628.

Wernimont, A.K., Huffman, D.L., Lamb, A.L., O`Halloran, T.V. and Rosenzweig, A.C. (2000) Structural basis for copper transfer by the metallochaperone for the Menkes/Wilson disease proteins. *Nat.Struct.Biol*. **7,** 766-771

Wimmer, R., Herrmann, T., Solioz, M. and Wüthrich, K. (1999) NMR structure and metal interactions of the CopZ copper chaperone. *J.Biol.Chem*. **274,** 22597-22603

Wistrand, M. and Sonnhammer, E.L. (2005) Improved profile HMM performance by assessment of critical algorithmic features in SAM and HMMER. *BMC Bioinformatics*. **6**, 99.

Wodara, C., Bardischewsky, F. and Friedrich, C.G. (1997) Cloning and characterization of sulfite dehydrogenase, two c-type cytochromes, and a flavoprotein of *Paracoccus denitrificans* GB17: essential role of sulfite dehydrogenase in lithotrophic sulfur oxidation. *J. Bacteriol*. **179**, 5014-5023

Wood, Z.A., Poole, L.B., and Karplus, P.A. 2001. Structure of intact AhpF Reveals a mirrored thioredoxin-like active site and implies large domain rotations during catalysis. *Biochemistry.* **40**, 3900–3911.

Wyman, J. (1967) Allosteric linkage. *J. Am. Chem. Soc*., **89**, 2202–2218.

Xiang Z. (2006) Advances in homology protein structure modeling. Curr *Protein Pept Sci*., **7**, 217-227

Xiang, Z. and Honig, B. (2001) Extending the accuracy limits of prediction for side-chain conformations. *J Mol Biol*. **311**, 421-430

Xu, Y. and Uberbacher, E.C. 1997. Automated gene identification in large-scale genomic sequences. *J. Comput. Biol*. **4**, 325–338.

Xu, Y. and Xu, D. (2000) Protein threading using PROSPECT: design and evaluation. *Proteins*. **40**, 343-354

Yankovskaya, V., Horsefield, R., Törnroth, S., Luna-Chavez, C., Miyoshi, H., Léger, C., Byrne, B., Cecchini, G. and Iwata, S. (2003) Architecture of succinate dehydrogenase and reactive oxygen species generation. *Science*. **299**(5607), 700-7004.

Yao, H., Guo, L., Fu, Y., Borsuk, L.A., Wen, T.J., Skibbe, D.S., Cui, X., Scheffler, B.E., Cao, J., Emrich, S.J., Ashlock, D.A. and Schnable, P.S. (2005) Evaluation of five ab initio gene prediction programs for the discovery of maize genes. *Plant Mol Biol*. **57**(3), 445-460.

Zagrovic, B., Snow, C.D., Shirts, M.R. and Pande, V.S. (2002) Simulation of folding of a small alpha-helical protein in atomistic detail using worldwide-distributed computing. *J Mol Biol*, **323**, 927-937.

Zemla, A. (2003). LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Res*., **31**, 3370-3374.

Zemla, A., (ver. 09/2006) LGA - Protein Structure Comparison Facility. http://predictioncenter.llnl.gov/ (Last accessed 23rd October 2007)

Zhang, W.,Chunhai, F., Yuting, S and Li, G. (2003) An electrochemical investigation of ligand-binding abilities of biomimetic membrane-entrapped myoglobin. *Biochim. Biophys. Acta Gen. Subj.* **1623**, 29–32.

Zhang, Y. (2008) I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*. **9**, 40.

Zhao Y., White, M.A., Muralidhara, B.K., Sun, L., Halpert, J.R. and Stout, C.D. (2006). Structure of microsomal cytochrome P450 2B4 complexed with the antifungal drug bifonazole: insight into P450 conformational plasticity and membrane interaction. *J Biol Chem*. **281**, 5973-5981