AN ABSTRACT OF THE DISSERTATION OF

Richard Bleier Cooley for the degree of Doctor of Philosophy in Biochemistry & Biophysics presented on September 8, 2011.
Title: Functional, Structural and Evolutionary Analyses of the Ferritin-like Superfamily of Proteins.

Abstract approved: _____
Daniel J. Arp

The ferritin-like superfamily (FLSF) of proteins is composed of a wide variety of functionally diverse proteins involved in oxygen dependent metal-mediated electron transfer reactions. Their biological importance is exemplified by the fact FLSF proteins are found in almost every organism from all three domains of life. Their functions range from protection against reactive oxygen species to iron detoxification and storage, the synthesis of DNA and lipid membrane building blocks, energy dissipation, the creation of antibiotics, the modification of tRNA nucleotides and the oxidation of hydrocarbons.

This dissertation presents studies aimed at the functional, structural and evolutionary characterization of two particular groups of FLSF proteins: bacterial multicomponent monooxygenases (BMMs) and rubrerythrins. BMMs are enzymes expressed by bacteria that allow them to grow on hydrocarbons, such as methane, butane and toluene, as their sole source of energy and carbon by inserting an oxygen atom into their rather unreactive C-H bond. Rubrerythrins, on the other hand, are peroxidases most commonly found in anaerobic bacteria; they provide protection against hydrogen peroxide damage and are believed to retain several ancestral features of the FLSF.

Five chapters of original research are presented in this dissertation, all of which are either published or accepted for publication. The first two chapters describe novel biochemical insights and physiological implications of a poorly studied BMM known as soluble butane monooxygenase. These results constitute the first *in vitro* biochemical characterization of an alkane oxidizing BMM from a non-methanotroph. They also provide the first description of how such bacteria can grow in natural gas even though they cannot grow on methane, the principle component of natural gas.

The third chapter uncovers a novel mechanism of protein evolution important not only in the functional diversification of BMMs like soluble butane monooxygenase, but nearly 15% of all proteins as well. In doing so, it reshapes our understanding of a poorly recognized protein secondary structure called the π-helix and helps bring this structural motif to the forefront of structural and evolutionary biology. Based on these findings, we predict the presence of certain novel features in a previously uncharacterized rubrerythrin-like protein found only in two "living fossil" oxygenic phototrophs. The fourth and fifth chapters that follow up this prediction describe the structural characterization of this rubrerythrin-like protein that we named symerythrin due to its unprecedented level of internal symmetry. The results confirm several of our predictions regarding the novelty of symerythrin's diiron metallocenter, and also unexpectedly show that it is capable of performing unprecedented chemistry: the formation of a carbon-carbon crosslink between two unfunctionalized amino acids. We also find unanticipated evidence that the single chain FLSF fold had multiple independent evolutionary origins.

Functional, Structural and Evolutionary Analyses of the Ferritin-like Superfamily of Proteins

by
Richard Bleier Cooley

A DISSERTATION
submitted to
Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented September 8, 2011
Commencement June 2012

Doctor of Philosophy dissertation of Richard Bleier Cooley presented on September 8, 2011.

APPROVED:

Major Professor, representing Biochemistry & Biophysics

Head of the Department of Biochemistry & Biophysics

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

Richard Bleier Cooley, Author

ACKNOWLEDGEMENTS

First I thank my advisor, Dan Arp, for his mentorship over the last five years. Your calm demeanor, enthusiasm and positive support create a fantastically positive learning environment for any student. Your ability to rationalize solutions to difficult problems is deserving of great respect. I thank you for giving me the freedom and encouraging me to follow up on any idea no matter how ridiculous you probably thought it was.

I also thank my co-advisor, Andy Karplus. Your excitement for science is both unmatched and contagious. Your relentless drive to teach, unparalleled patience and resilience to rejection are deserving of great admiration. I have grown both personally and scientifically because of this.

I thank our laboratory manager Luis Sayavedra-Soto for your incredible patience and help throughout these years. I am extremely appreciative of your making sure I had all the resources I needed and for fixing all the things I broke in the lab. I also would like express great gratitude toward your daughter for the countless hours she spent washing dishes in our laboratory.

I thank Dale Tronrud for providing me with your incredible wealth of crystallographic knowledge. You are a humbling person to be around and I am extremely appreciative of everything I have learned from you. I promise to never litter the PDB with any poorly refined structure.

I would also like to express my gratitude toward Peter Bottomley for your help, advice and encouragement with my projects as well as the laughter over the years. I thank the other members my committee, Joe Beckman and Jerry Heidel, for your interest in my development as a scientist. I thank Brad Dubbels for teaching me the ropes of the Arp laboratory during my first year, and Naraja Vajrla and Anne Taylor as well for all your help and advice throughout the years.

I thank my good friends and lab mates Justin and Andrea Hall, Camden Driggers, Russell Carpenter and Tim Rhoads for their family-like friendship and

support both in and out the laboratory. I have learned so much from you all and could not have asked to be around a better group of friends.

I thank my Mom, Dad and brother Mike for their support over the years. I will always attribute much of my success as a scientist to my high school chemistry teacher. So, thank you, Mom. I also thank my parents-in-law, Rich and Kari Brendtro, for their unwavering support, encouragement and continual supply of home-cooked food.

I would like to acknowledge the National Institutes of Health grant numbers 5RO1 GM56128-06 and GM R01-083136, the Oregon Agricultural Experiment Station and the Environmental Health Sciences Environmental Health Sciences Center grant ES00210 for their support. I also am extremely grateful for financial support from the Nellie Buck Yerex Graduate Fellowship.

Lastly and most importantly, I thank my wife Stacy. No words can express how grateful I am to have had you alongside me all these years. To say none of this dissertation would have been possible without you would be a monumental understatement. I love you.

CONTRIBUTION OF AUTHORS

Bradley Dubbels contributed to the purification and kinetic studies of soluble butane monooxygenase described in Chapter 2 as well as the writing of that text. Luis Sayavedra-Soto contributed to the experimental design and writing of Chapter 2. Peter Bottomley contributed text of Chapters 2 and 3 as well as the design and analysis of those experiments. Timothy Rhoads provided the mass spectrometry data collection and analysis in Chapter 5. Daniel J. Arp was involved in the design, analysis and writing of all experiments and chapters. P. Andrew Karplus was involved in the design, analysis and writing of Chapters 4, 5 and 6.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

Page

TABLE OF CONTENTS (Continued)

Page

TABLE OF CONTENTS (Continued)

Page

LIST OF FIGURES

LIST OF FIGURES (Continued)

LIST OF TABLES

LIST OF APPENDIX FIGURES

LIST OF APPENDIX TABLES

Dedicated to

My beautiful wife, Stacy.

Functional, Structural and Evolutionary Analyses of the Ferritin-like Superfamily of Proteins

Chapter 1

Thesis Overview

Introduction to the Study of Proteins

All organisms depend on proteins to survive. The importance of proteins was recognized in the earliest days of biochemistry and, until the 1940's, proteins were even thought to encode the genetic information of cells because they are more chemically and structurally diverse than DNA (O'Connor 2008). Today, we know proteins are the "workhorses" of the cell by acting, for instance, as catalysts (e.g. trypsin), transporters (e.g. myoglobin), molecular motors (e.g. dynein, ATP synthase) and as support for cellular structure (e.g. tubulin). As even small defects in proteins can have detrimental consequences for an organism, it is essential to understand how proteins interact with, respond to and organize themselves within the cellular environment.

Indeed, advancements in our understanding of protein function have led to the attenuation of life threatening diseases and ailments, and as such have contributed to the expansion of the multi-billion dollar pharmaceutical industry. In many cases, protein function is studied to find ways to specifically inactivate or inhibit a physiologically problematic process. Outside of the medicinal world, however, it is becoming increasingly clear that we can use proteins to catalyze complex molecular reactions with efficiencies beyond the reach of the capabilities of the most brilliant chemists and engineers. As the world population steadily increases along with our reliance on fossil fuels, the demand for biologically inspired, economically viable solutions to clean energy and pollution mitigation will surely increase.

Regardless of the end goal, there exist a multitude of approaches to elucidate details of protein functionality. When the first three-dimensional structure of a protein, that of myoglobin, was reported (Kendrew *et al.* 1958), it became clear that much could be learned about the function of a protein if its structure was known. Recent advancements in synchrotron-based data collection have allowed X-ray crystallography to become the most commonly used and successful tool for determining the three-dimensional shape of macromolecular structures. Currently,

there are almost 25,000 non-redundant structures deposited in the Protein Data Bank (Berman *et al.* 2000), of which 3800 were deposited just within the last 12 months alone. The importance of protein structure determination is evidenced by the fact that the 2003, 2006 and 2009 Nobel Prizes in Chemistry were awarded to investigators for their work in uncovering the three-dimensional structural details of biologically important complexes.

While much can be learned about a protein from knowing its structure, the use of other biochemical and biophysical techniques is essential to achieve a more complete understanding of a protein's function. The overall goal of this thesis therefore is to use a variety of structure- and function-based techniques to provide new insights into the function and evolution of a particular group of metal-containing enzymes called the ferritin-like superfamily. Many of these insights offer important advances in our understanding of this particular group of proteins while other new concepts have much broader implications. In the rest of Chapter 1, I will briefly introduce some fundamentals of protein structure and how they relate to the hierarchy of protein classification and protein evolution. This will lead into a review of the structural, functional and evolutionary framework of the ferritin-like superfamily that will provide the necessary background for understanding the questions that are addressed in each of the five research chapters. For convenience, I end this chapter with a brief summary of the topics of these chapters.

Protein structure and classification

Protein structure. Despite its potential to adopt a limitless set of structures, the peptide backbone is energetically restrained to a relatively small set of conformations. α-helices and β-sheets are the most commonly observed secondary structures. α-helices are formally defined as structures with repeating main-chain hydrogen bonds between residues four apart in sequence (Fig 1.1A). The π-helix, an important structural feature discussed in Chapters 4 and 6, is defined by repeating main-chain

hydrogen bonds between residues five apart in sequence (Fig. 1.1B) and is commonly embedded within longer α-helices (Fig. 1.1C). As multiple secondary structures along the peptide chain associate, supersecondary structures are formed, like αα-hairpins (two α-helices linked together) or ββ-hairpins (two β-strands linked together). These supersecondary structures then come together through a still poorly understood process involving entropic, hydrophobic, ionic and hydrogen bond interactions to create the tertiary structure. As a typical protein chain folds, a single globular structure forms to create active/ interactional sites essential for the function of the protein.

The hierarchy of protein classification. The recent surge in the number of proteins with experimentally determined structures demonstrates they not only adopt a relatively small set of secondary structures as mentioned above, but also a surprisingly limited set of autonomously folding three-dimensional structures, or domains. This allows proteins to be grouped into a hierarchy of protein families, superfamilies and folds (Soding and Lupas 2003). Proteins grouped into the same family share similar sequence, structure and often function. Groups of families that are homologous, i.e. share a common ancestor as evidenced by commonalities in structure and often function, are considered to form a superfamily. Superfamilies with analogous three-dimensional structures are grouped into folds, however these superfamilies having the same fold are not necessarily assumed to be evolutionarily related. For example, the methylenetetrahydrofolate reductase family of enzymes belongs to the FAD-linked oxidoreductase superfamily, which belongs to the TIM-β/α barrel fold (Murzin *et al.* 1995).

Recent evidence has suggested that the sub-domain building blocks (supersecondary structures) of superfamilies belonging to the same fold may in fact have a common origin (Soding and Lupas 2003; Alva *et al.* 2010). This observation is the basis for the "protein from peptides" hypothesis for the origin of proteins in which protein domains were formed by the concatenation of smaller peptides with supersecondary structures like αα- or ββ-hairpins. For example, two proteins having

a four-helix bundle fold could have been formed by the fusion of homologous αα-hairpin peptides even with the peptides having been fused in different directions relative to each other (Fig. 1.1D). In such an event, both folds would be four-helix bundles but the topology of the bundle would be different. These folds would not be homologous (since they had different origins), but the sub-domain building blocks (i.e. supersecondary structures) would be homologous. While this "proteins from peptides" hypothesis is intriguing, conclusive evidence in support of it has been lacking largely because steady genetic diversification over billions of years has masked the discernable signals needed to establish homology. As more genomes are sequenced, more structures solved and more sophisticated algorithms are developed to detect distant homology, we can begin to address the validity of this hypothesis and possibly gain new insights into the origin of proteins and life. As you will see, one of the most interesting applications of the studies described in Chapter 6 is novel evidence to support this hypothesis.

The Ferritin-like Superfamily of Proteins

Families of the ferritin-like superfamily. One well-characterized superfamily of proteins is called the ferritin-like superfamily (FLSF). As the name suggests, ferritins are a prominent member of this superfamily. Ferritins have been found in organisms from all three domains of life (Crichton and Declercq 2010) and functionally, these proteins serve two purposes (Andrews 2010). First, ferritin binds to ferrous iron and oxidizes it with molecular oxygen in a controlled manner preventing the formation of damaging reactive oxygen species that would otherwise form when free ferrous iron reacts with oxygen through Fenton-like reactions (Fenton 1894). The second purpose of ferritin is to form a protective protein cage surrounding the insoluble ferri-oxy hydroxides formed in the previous step. These two functions of ferritins (as well as bacterioferritins and DNA-binding proteins from starved cells, or

DPS proteins) serve to limit the iron-catalyzed formation of reactive oxygen species and keep cellular ferric iron biologically accessible.

In addition to the widely distributed ferritins and other related iron storage proteins, there are many other groups of proteins that fall within the FLSF. The two of most interest to this dissertation are the rubrerythrins and the bacterial multicomponent monooxygenases (BMMs), both of which will be discussed in greater detail later in this section.

Arguably the most well studied family of FLSF proteins is class I ribonucleotide reductases (RNR). Class I RNRs exist as a dimer of dimers ($\alpha_2\beta_2$). The α-subunit contains the active site where the reduction of ribonucleotides to deoxyribonucleotides occurs. The β-subunit (also referred to as the R2 subunit) serves to provide a radical generating cofactor used to initiate catalysis in the α-subunit. Historically, class I RNRs were divided into two groups, a and b, based on mechanisms of regulation and sequence similarity (Nordlund and Reichard 2006). In both of these classes, the radical generating cofactor is a tyrosyl radical, which is created by a nearby dinuclear center. Although both of these classes can generate the tyrosyl radical with a diiron center, new evidence shows that active class Ib RNRs *in vivo* have a dimanganese metallocenter (Cotruvo and Stubbe 2011). Recently, a third group of class I RNRs was discovered (Hogbom *et al.* 2004). This group, Ic, is of particular interest to the field because they do not generate a tyrosyl radical like class Ia/b RNRs. Instead, they utilize their dinuclear center directly to initiate ribonucleotide reduction. Research has shown that the active form of RNR Ic possess a heteronuclear Mn(IV)/Fe(III) cofactor (Jiang *et al.* 2007). Structures of all three groups of class I RNRs have been reported.

Many other FLSF members have been both structurally and biochemically characterized, including fatty acid acyl carrier protein desaturases, which are commonly found in the plastids of higher plants and catalyze dehydrogenation reactions that result in the introduction of a double bond in saturated fatty acids (Shanklin *et al.* 2009). Decarbonylases are enzymes that were only recently

discovered in cyanobacteria and catalyze the formation of alkanes from fatty aldehydes (Schirmer *et al.* 2010). Mn-catalases, as the name suggests, are dimanganese-dependent FLSF proteins found mostly in bacteria that catalyze the disproportionation of hydrogen peroxide (Andrews 2010). Lesser studied FLSF proteins with known structures include MiaE, which functions as a hydroxylase in the modification of an adenosine residue in transfer RNAs specific for recognizing codons beginning with uracil (Mathevon *et al.* 2007), and AurF, which is found in *Streptomyces thioluteus* and catalyzes the six electron oxidation of an aromatic amino group to an aromatic nitro group during the biosynthesis of the antibiotic aureothin (Choi *et al.* 2008).

There also exist several FLSF proteins that have been characterized structurally for which no physiologic function has been assigned. Many of these include proteins solved by structural bioinformatic groups (PDB codes 2oc5, 2oh3 and 2ib0) as well as a protein known to be upregulated under stress conditions in *E. coli* referred to as YciF (Hindupur *et al.* 2006).

Lastly, several FLSF protein families have been biochemically characterized but their structures have yet to be solved. Most notably are the alternative oxidases, which were originally found in plants but recently published genomic sequences suggest they also exist in animals and even bacteria (Moore and Albury 2008). Alternative oxidases function as the terminal component of a non-proton motive force-coupled electron transport pathway for the oxidation of ubiqionol, which results in the reduction of O_2 to water. Because these proteins are membrane associated, biochemical characterization has proven difficult. Nevertheless, a Japanese group has recently published preliminary reports for the crystallization of the alternative oxidase from *Trypanosoma brucei* (Kido *et al.* 2010), suggesting a crystal structure for this enzyme will soon be solved and published. Lastly, an enzyme known as COQ7 found in eukaryotes and bacteria but not archaea and involved in the biosynthesis of ubiquinone has recently been biochemically characterized as a member of the FLSF (Andrews 2010). No structure of this enzyme exists yet, however it seems to be

unusual among the FLSF in that NADH can directly reduce its diiron center without the need of a separate reductase component (Behan and Lippard 2010).

Defining Characteristics of the FLSF. Structurally, all of these families within the FLSF possess a conserved core four-helix bundle (individual helices referred to as A, B, C and D) with a down-up-up-down topology (Fig. 1.2A). This topological arrangement of the helices distinguishes it from other superfamilies within the four-helix bundle fold such as hemerythrin-like superfamily, which has a down-up-down-up arrangement.

A second feature common to FLSF members is the ability to bind two metal ions at the core of the four-helix bundle. These two metal ions are typically iron, but functional dimanganese (Cotruvo and Stubbe 2011) and heteronuclear iron/manganese (Jiang *et al.* 2007) centers have also been reported. Regardless of their identity, the metals are generally coordinated to the four-helix bundle via six positionally conserved metal-ligating residues: four carboxylate and two histidine residues (Fig. 1.2B). Structures have revealed that at least one of the carboxylate ligands bridges the two metal ions, giving rise to the term "carboxylate bridged" dinuclear centers.

A third common feature among FLSF proteins is that, although their physiologic functions differ quite substantially, they catalyze their respective reactions through analogous mechanisms. Prior to reaction with substrate, the dinuclear center is reduced (i.e. from the resting diferric state to the diferrous state), at which time the dinuclear center reacts with molecular oxygen or hydrogen peroxide. The ensuing reaction between the metals and oxygen species vary between enzymes. In some cases, like ferritin and alternative oxidases, the oxygen is simply reduced to water or hydrogen peroxide and the dinuclear center is oxidized back to the resting state in preparation for another round of catalysis. In other cases, high-valent metal-oxygen intermediates are formed, which are then used to oxidize a particular substrate. In the case of desaturases, a saturated fatty acid is oxidized to an unsaturated fatty acid. In class Ia and Ib ribonucleotide reductases, a nearby tyrosine residue is oxidized to a

tyrosyl radical. With soluble methane monooxygenase, a BBM enzyme discussed later in this chapter, methane is oxidized to methanol.

The wide range of functions of FLSF members from iron storage (ferritins, bacterioferritins) to enzymes that can perform some of the most energetically challenging oxidations in biology (methane monooxygenases) suggests there exist important structural differences in the regions surrounding the dinuclear active sites. One of the underlying goals of this dissertation is to provide insight into how these proteins have evolved to create such different functions, and how these changes have helped organisms to diversify and thrive in unusual environments. The following section of this introductory chapter will present what was previously understood about the structural differences between FLSF enzymes and then will provide more detailed descriptions of the two different enzyme systems (BMMs and rubrerythrins) that were used in this thesis work to address questions about FLSF evolution, structure and function.

Variations in the metallocenters of FLSF proteins. The simplest explanation for the remarkable variation in catalytic potential of FLSF proteins is differences in metal coordination. Prior to this dissertation, only three significant sequence level variations of metal-ligating motifs were known (Fig. 1.2B). The most common motif contains six metal-ligating residues and is characteristic of BMM enzymes, desaturases, bacterioferritins, class Ic RNRs, alternative oxidases and others. I refer to this as the canonical six-ligand motif because of its widespread occurrence and the fact there are six iron-ligating residues. A variation on the canonical six-ligand motif is found in RNR class Ia and Ib enzymes in which the glutamate ligating residue contributed by helix A is substituted by an aspartate. Not surprisingly, class Ia/b enzymes, which have this aspartate substitution, and class Ic enzymes, which have the canonical six-ligand motif, carry out ribonucleotide reduction through different mechanisms: class Ia/b use their diiron center to form a tyrosyl radical cofactor necessary for nucleotide reduction (Jordan and Reichard 1998) , while class Ic RNRs form a high-valent Mn(IV)-Fe(III) cofactor (Jiang *et al.* 2007).

The third motif, referred to as the seven-ligand motif, has only been found in rubrerythrin-like proteins and is considered unusual among the FLSF. This motif is identical to the six-ligand motif except that a seventh iron-ligating residue (a glutamate) is contributed by helix C. Our work in Chapters 4 and 6 demonstrate that this seventh ligand is an inserted residue (Fig. 1.2B) and it resides within a π-helical segment of helix C. With this constellation of metal ligating residues, rubrerythrins have the unique function among FLSF proteins in that they act as peroxidases to reduce potentially harmful hydrogen peroxide to water (Riebe *et al.* 2009).

Single amino acid substitutions change metallocenter chemistry. To address the effects these different motifs have on the chemical and catalytic properties of ferritin-like enzymes, several studies have utilized single amino acid substitutions designed to alter the primary ligation sphere of one enzyme to mimic that of a different enzyme within the superfamily. For example, a single substitution was made near the active site of desaturase by altering a threonine residue to a carboxylate-containing residue, effectively placing the seven-ligand motif characteristic of rubrerythrin into the active site of desaturase, which naturally contains the six-ligand motif. The results demonstrated the altered desaturase displayed catalytic properties and an active site structure similar to that of reduced rubrerythrin (Guy *et al.* 2006).

Separate studies (deMare *et al.* 1997; Coulter *et al.* 2000) attempted to do the opposite: to place a six-ligand motif into rubrerythrin by substituting the seventh ligating residue (Glu) in helix C in rubrerythrin for Ala. The resulting active site structure mimicked that of desaturase and RNR in the positioning of the iron atoms and new solvent ligands (deMare *et al.* 1997). Despite the observed similarities in active site coordination chemistry, the altered enzyme could not create the tyrosyl radical characteristic of RNR (deMare *et al.* 1997). However, the peroxidase activity of rubrerythrin was abolished by this alteration (Coulter *et al.* 2000).

Alterations in the metallocenter of class Ia/b RNRs have also been designed to make it more similar to RNR Ic and BMM enzymes. To do so, the Asp ligating residue in class Ia/b RNRs was altered to a Glu. These resulting mimics of BMMs

showed catalytic similarities to sMMO (a BMM enzyme) in the production of peroxo-diiron(III) intermediates (Moenne-Loccoz *et al.* 1998; Baldwin *et al.* 2003; Skulan *et al.* 2004) and hydrocarbon oxidation (Baldwin *et al.* 2001), as well as structural similarities (Voegtli *et al.* 2000). Despite these similarities between the MMO-like mutant of class Ia/b RNRs and MMO itself, differences in the coordination chemistry still existed and no accumulation of the oxidizing iron-oxygen intermediate characteristic of sMMO (called intermediate 'Q', discussed below) (Shu *et al.* 1997) was observed.

Variations in the peptide backbone surrounding FLSF metallocenters. Although the core four-helix bundle in the FLSF is universally conserved across its various families, the backbone geometry of each individual helix is not (Fig. 1.3). To understand the cause of these α-helical perturbations, detailed analyses described in Chapter 4 indicate these perturbations are in fact short, and sometimes overlapping, π-helices embedded within the α-helical core. One of the important findings outlined in that chapter is that each individual perturbation, or π-helix, results from a single amino acid insertion into the α-helix. This correlation suggests a novel mechanism of secondary structure evolution whereby α-helices are perturbed by the insertion of a single amino acid in order to create a π-helix. Although such insertions into an α-helix are destabilizing by ~3-6 kcal/mol (Heinz *et al.* 1993; Keefe *et al.* 1993; Heinz *et al.* 1994), experiments have shown that such alterations are surprisingly well tolerated (Sondek and Shortle 1992; Shortle and Sondek 1995). The insertion and deletion of multiple base-pairs into DNA during replication through slipped-strand mispairing is a common mechanism of genetic variability (~10^{-4} events per cellular division) that could account for the in-frame insertion or deletion of a single codon (Levinson and Gutman 1987; van Belkum *et al.* 1998). In the FLSF alone, rubrerythrin, eukaryotic ferritin, the R2 subunit of all class I RNRs, AurF and all BMM enzymes all contain at least one π-helix. In fact, soluble methane monooxygenase (a BMM enzyme) contains a total of 12 π-helices, the most of any single structure in the PDB. Given that π-helices and similar helical deformations (e.g. α-bulges, π-bulges) have been correlated

with function (Weaver 2000; Fodje and Al-Karadaghi 2002; Cartailler and Luecke 2004), it is not surprising that several of the π-helices in BMMs and RNRs are proposed to play important functions (Eriksson *et al.* 1998; Sazinsky and Lippard 2005; Sazinsky and Lippard 2006; Bailey *et al.* 2008).

Function and Catalysis of Bacterial Multicomponent Monooxygenases. There are three major groups of BMM enzymes: soluble methane monooxygenases, alkene/aromatic monooxygenases and phenol hydroxylases (Leahy *et al.* 2003). These enzyme complexes are expressed by bacteria and allow them to utilize otherwise biologically inaccessible hydrocarbons as growth substrates by oxidizing hydrocarbons to alcohols (or alcohols to diols, in the case of phenol hydroxylases) as the first step of metabolism. Once this oxidation process is accomplished, the hydrocarbon can be further oxidized by a variety of other enzymes, including alcohol and aldehyde dehydrogenases, to provide the energy needs of the cell.

It is important to note that while several chapters of this dissertation focus on alkane monooxygenases from the FLSF such as sMMO, other evolutionarily unrelated alkane monooxygenases that should not be confused with BMMs also exist. In fact, there are three other such families of alkane monooxygenases known. The first group are the copper-based particulate alkane monooxygenases; these are membrane-bound proteins most commonly found in methanotrophs where they are expressed when copper is in adequate supply. Only under copper limiting conditions is sMMO expressed. Biochemical characterization of particulate methane monooxygenase has proven challenging due to instability upon solubilization from membranes, which results in almost complete loss of activity. Nevertheless, crystal structures have been reported (Lieberman and Rosenzweig 2005), and recently the copper containing active site was identified (Balasubramanian *et al.* 2010). Particulate methane monooxygenases have a relatively narrow substrate range compared to sMMO, however a homolog referred to as particulate butane monooxygenase, with specificities for longer chain alkanes, has been identified recently in *Nocardioides sp.* strain CF8 (Sayavedra-Soto *et al.* 2011). Characterization of this particular butane

monooxygenase will be an important step in understanding the mechanisms of substrate specificity for these enzymes.

The second type of alternative alkane monooxygenase is known as the AlkB family. Like BMMs, AlkBs utilize a diiron metallocenter to carry out catalysis, however the metallocenter of AlkBs are distinct from BMMs in that they contain eight histidine residues (Shanklin *et al.* 1997). AlkBs are membrane-bound enzymes with a substrates range optimized for larger alkanes (C_5 - C_{16}) than the copper-containing particulate enzymes (van Beilen and Funhoff 2007). As for the particulate methane monooxygenase, detailed characterization of AlkB has proven challenging due to difficulties in obtaining purified enzyme in the fully folded and active state. However, a report published this year describes an optimized procedure for the isolation of AlkB (Xie *et al.* 2011), opening the door to future characterizations.

The last family of alternative alkane monooxygenases are the membrane bound cytochrome P450 enzymes. These monooxygenases are found in both bacterial and eukaryotic organisms and are known to oxidize alkanes with specificity ranges simlar to the AlkB monooxygenases (Funhoff *et al.* 2006; van Beilen and Funhoff 2007). Crystal structures of these P450 monooxygenases have been reported (Schlichting *et al.* 1997), and preliminary kinetic studies suggest that they have particularly low dissociation constants for their alkane substrates (Funhoff *et al.* 2006).

Because of the relative ease in isolating BMM enzymes compared to these three alternative families of alkane monooxygenases, BMMs have been characterized in much greater detail. The largest component (~250 kDa) is referred to as the hydroxylase and is a dimer of three unique subunits ($\alpha_2\beta_2\gamma_2$) (Fig. 1.4a). The diiron center resides within the α-subunit, which, like all FLSF proteins, possesses a core four-helix bundle. The β-subunit is structurally similar and homologous to the α-subunit but it does not bind metals and does not directly participate in catalysis. The function of the γ-subunit is not well understood and in fact there is no discernable sequence or structural similarity between the γ-subunits of the three major groups of BMMs. Representative structures of the hydroxylase components from all three

BMM families have been determined by X-ray crystallography (Rosenzweig *et al.* 1993; Sazinsky *et al.* 2004; Sazinsky *et al.* 2006).

The hydroxylase of BMMs, however, is inactive (or has very low activity) unless a small regulatory subunit is bound to it. This cofactorless protein (~15 kDa) binds to the α-subunit of the hydroxylase and alters the geometry of the active site in order for substrates to enter and oxygen to react with the diiron center (Mitic *et al.* 2008). The mechanism(s) by which the regulatory component activates the hydroxylase is not well understood, and is an important subject of the work presented in Chapters 2 and 4.

The third component common to all BMMs is a flavin- and [2Fe-2S] cluster-containing reductase. The purpose of the reductase is to shuttle electrons from NADH to the diiron site of the hydroxylase converting it from the diferric resting state to the oxygen-sensitive diferrous state. Alkene/aromatic hydroxylases, however, are unique among the BMMs in that they utilize a fourth, Rieske-type ferredoxin component to mediate electron transfer from the reductase to the hydroxylase (Leahy *et al.* 2003).

The exact interplay of the three (or four in the case of the alkene/aromatic monooxygenases) components of BMMs is not well understood, however they are all needed for continuous substrate turnover. Current hypotheses suggest that as the reductase binds to one α-subunit and reduces the diiron center, a regulatory subunit binds to the second α-subunit activating its already reduced diiron center for substrate turnover. Once substrate turnover is complete in the second α-subunit, the regulatory and reductase subunits dissociate from their respective α-subunits and bind to the opposite α-subunit, and the process repeats. Under this mechanism, only one active site is undergoing catalysis at a time while the other is being prepared for catalysis by the reductase (Murray and Lippard 2007).

sMMO, expressed by bacteria that grow on methane as their sole source of carbon and energy, is the most well studied BMM. The isolation of the hydroxylase component (MMOH) was first reported in 1984 (Woodland and Dalton 1984), and its structure was solved in 1993 (Rosenzweig *et al.* 1993). The crystal structure revealed

the diiron center resides deep within the α-subunit with no apparent pathway for substrate access. However, structures with substrate analogues soaked into the crystals of MMOH prior to data collection revealed a series of connecting hydrophobic pockets leading from the solvent to the active site, demonstrating unexpected flexibility at the core of the protein (Sazinsky and Lippard 2005). Structures of the reductase (MMOR) (Muller *et al.* 2002; Chatwood *et al.* 2004) and regulatory component (MMOB) (Chang *et al.* 1999; Walters *et al.* 1999) were also determined by NMR. No structure of MMOH in complex with MMOB has been reported, however the analogous complexes with toluene monooxygenase (an alkene/aromatic monooxygenase) (Bailey *et al.* 2008) and a phenol hydroxylase (Sazinsky *et al.* 2006) have been determined.

In addition to these structural studies, much effort has gone into understanding the activation of molecular oxygen by the reduced diiron center of methane monooxygenase. Through a series of spectroscopic methods including stopped-flow spectroscopy, rapid-freeze quenching, Mossbauer spectroscopy and electron paramagnetic resonance (EPR) spectroscopy, the currently proposed mechanism (Fig. 1.5) involves the stepwise formation of two peroxy-diiron(III) complexes (P* and P), followed by the formation of the strongly oxidizing "intermediate Q" that reacts with substrate. Intermediate Q is proposed to have a diamond core bis-μ-oxo-diiron (IV) structure and is the only characterized enzymatic intermediate capable of oxidizing the strong C-H bond of methane. As such, designing bio-inspired catalysts that mimic its reactivity toward otherwise unreactive substrates has been a top priority for inorganic chemists (Do and Lippard 2011). Recently, the structure of Q has been questioned based on incompatibilities with theoretical analyses, thus warranting more detailed investigations (Tinberg and Lippard 2011). Interestingly, alkene/aromatic monooxygenases do not generate a Q-like intermediate during catalysis (Bochevarov *et al.* 2011). Instead, P-like intermediates appear to be sufficient to break the weaker C-H bonds of alkenes and phenols.

Butane monooxygenase: a homolog of methane monooxygenase. Chapters 2 and 3 of this dissertation focus on an enzyme referred to as soluble butane monooxygenase (sBMO). This enzyme, expressed by the C_2-C_9 alkane-utilizing β-proteobacterium *Thauera butanivorans* (Takahashi *et al.* 1980; Dubbels *et al.* 2009), is closely related to sMMO (the α-, β-, and γ-subunits of BMOH, have 65, 42, and 38% sequence identity with the corresponding subunits of MMOH (Halsey *et al.* 2006)). The physiologic function of sBMO is to oxidize aqueous butane to 1- and 2-butanol (in an 85:15 ratio, respectively) in an NADH- and O_2-dependent manner (Dubbels *et al.* 2007). Although the metabolic pathway of secondary alcohols in *T. butanivorans* is not well understood, 1-butanol is thought to be oxidized to butyraldehyde and then to butanoic acid in energy-producing processes (Arp 1999). Catabolism of butanoic acid then proceeds via standard even-chained fatty acid metabolism that yields two acetyl-CoA molecules following β-oxidation. These two-carbon units are then oxidized via the TCA cycle to supply the cell with additional energy.

One of the advantages for studying sBMO is that the host organism, *T. butanivorans*, is more amenable to mutagenesis than methanotrophs because *T. butanivorans* can grow on many different substrates, including lactate, citrate, succinate and acetate, while methanotrophs can only grow on methane. Mutants with alterations to the sBMO operon, therefore, are not necessarily lethal even if sBMO is non-functional. This ability has allowed our laboratory to publish new insights into the regulation of alkane monooxygenase expression (Sayavedra-Soto *et al.* 2005; Doughty *et al.* 2006; Doughty *et al.* 2008). Furthermore, site-directed mutants of the sBMO hydroxylase have also been made and characterized (Halsey *et al.* 2006). These studies revealed that alterations in residues near the active site alter product regioselectivity (e.g. from 1-butanol to 2-butanol). A purification process for all three components of sBMO has been established (Dubbels *et al.* 2007), allowing us to address several important questions *in vitro* as well as *in vivo*.

One of the most intriguing questions about BMMs is how substrate specificity is determined. Although the only physiologic substrate of sMMO is methane and the

active site is buried deep within the core of the protein, over a hundred different hydrocarbons have been documented as substrates, including large polycyclic molecules like naphthalene (Borodina *et al.* 2007). This wide range of substrates utilized by sBMO and sMMO is of particular interest to the fields of bioremediation because it means that these enzymes, as powerful oxidants, can be used as catalysts to breakdown otherwise stable and carcinogenic pollutants commonly dumped into the environment. One of the goals of bioremediation research is to create variants of these enzymes that are more robust and produce breakdown products that are less toxic to the native organism. Toward this goal, understanding the mechanisms by which substrate specificity is determined in BMMs is crucial.

Despite the generally promiscuous natures of sMMO and sBMO, previous studies in our lab have suggested the two react differently with methane (Halsey *et al.* 2006). This result is particularly intriguing given that sequence alignments and homology modeling of BMOH shows it contains all the same residues lining the active site pocket as MMOH (Halsey *et al.* 2006). Studying these two closely related but notably different enzymes, therefore, provides a unique opportunity to address these questions regarding substrate specificity in order to advance the field of bioremediation and also to understand how *T. butanivorans* and methanotrophs thrive in their respective alkane rich environments.

The rubrerythrin family of the FLSF. Like BMMs, rubrerythrins are proteins with a carboxylate-bridged diiron center housed within a ferritin-like core four-helix bundle. The first isolation of rubrerythrin was reported in 1988 from the periplasmic fraction of the anaerobic sulfate-reducing bacterium *Desulfovibrio vulgaris* (LeGall *et al.* 1988). The name rubrerythrin was given to this protein because the sequence indicated it had a hemerythrin-like domain (i.e. a four-helix bundle) fused with a rubredoxin-like domain at its C-terminal end (rubredoxin + hemerythrin = rubrerythrin) (Fig. 1.4B). This naming convention is misleading, however, because the four-helix bundle of rubrerythrins actually has ferritin-like topology (down-up-up-down) rather than hemerythrin-like topology (down-up-down-up). Nevertheless, the

name rubrerythrin is still regularly used today and many others have expanded on it by giving names like nigerythrin (which was reported to be black in color and differs from normal rubrerythrins in that the rubredoxin domain is fused to the N-terminal end of the four-helix bundle (Iyer *et al.* 2005)) and sulerythrin (a rubrerythrin-like protein isolated from *Sulfolobus tokodaii* (Wakagi 2003)) to other rubrerythrin-like proteins.

With more genomes sequenced, Andrews (Andrews 1998) noted a gene in the cyanelluar genome of the primitive eukaryotic oxygenic phototroph *Cyanophora paradoxa* possessed similarity to rubrerythrin-like proteins, however this hypothetical protein lacked the rubredoxin domain. He coined the term "erythrin" to describe this protein (rubrerythrin - rubredoxin = erythrin). In Chapter 4 of this dissertation, I continue using the term erythrin in reference to this particular protein.

In Chapters 5 and 6, I describe follow up studies to the erythrin described in Chapter 4 but change the name of this protein to "symerythrin." The reason for this change is that shortly after the publication of Chapter 4, Andrews published a review of the FLSF (Andrews 2010) and used the name erythrin in a more general manner to describe essentially any uncharacterized ferritin-like protein possessing only a four-helix bundle whether or not there was any close relationship to the rubrerythrins or the two original erythrin proteins referred to in his 1998 publication. Thus, the "erythrin" described in Chapter 4 is the same protein described as "symerythrin" in Chapters 5 and 6, and is distinct from the proteins called "erythrins" in Chapter 6.

For many years, the physiologic function of classical rubrerythrins was not well understood. Initial *in vitro* studies reported pyrophosphatase (Van Beeumen *et al.* 1991), superoxide dismutase (Lehmann *et al.* 1996) and oxidase (Gomes *et al.* 2001) activities, however it is now reasonably well accepted that rubrerythrins act as the terminal component of an NADH-dependent peroxidase reaction (Zhao *et al.* 2007; Riebe *et al.* 2009). Interestingly, rubrerythrins are widespread in obligate anaerobes, microaerobes and in the oxygen-sensitive heterocysts of cyanobacteria (Gomes *et al.* 2001). Many of these organisms possess a superoxide reductase that reduces superoxide to hydrogen peroxide, which is then reduced to water by

rubrerythrin. As such, rubrerythrin appears to acts as part of an elaborate defense mechanism to detoxify reactive oxygen species in these oxygen-sensitive organisms.

The first crystal structure of a rubrerythrin was reported in 1996 (deMare *et al.* 1996) and showed, as mentioned previously in this introduction, that it possessed an unusual metallocenter compared to other FLSF proteins: the seven-ligand motif (Fig. 1.2C). This additional metal ligating residue causes the iron to which it is ligated (Fe1) to reside in a slightly different position compared to the six-ligand FLSF proteins. As a result, the conserved histidine contributed by core helix B is not ligating Fe1. The purpose of this alternate Fe1 position was not well understood until crystal structures of reduced (diferrous) rubrerythrin were reported, which showed that Fe1 moves 2.0 Å away from the unique seventh ligating residue to ligate the histidine from helix B and match the canonical position of Fe1 in other FLSF enzymes (Jin *et al.* 2002). Recent analyses have indicated that this toggling of Fe1 positions is important for rubrerythrin's peroxidase activity (Dillard *et al.* 2011), making it the most dynamic metallocenter of any characterized FLSF protein. Moreover, mutagenesis studies have revealed that rubrerythrin's unique seventh metal ligating residue is essential for maintaining its peroxidase activity (Coulter *et al.* 2000). Diferric and diferrous structures reported for nigerythrin support that this mechanism of peroxidase activity is relevant for all rubrerythrin-like proteins (Iyer *et al.* 2005).

Rubrerythrins and the origins of the FLSF. To understand how a particular group of proteins evolved, it is important to first understand which features are important for its function and then to determine how that feature was derived. In the FLSF, the two conserved features essential to the function of all families are the core four-helix bundle and the metal-ligating residues. Close inspection of the sequence alignment in Fig. 1.2B shows that the six-ligand motif is composed of two pairs of three metal-ligating residues where each pair consists of two glutamates and one histidine. Overlaying the two halves of the four helix bundle, that is helix pair A/B onto helix pair C/D, reveals that the two pairs of metal ligating residues are indeed structurally equivalent. This suggests that the four-helix bundle may have originated

from a dimerizing two-helix peptide that underwent a gene duplication and gene fusion event (Andrews 1998), and would support the previously mentioned "proteins from peptides" hypothesis in which the fundamental building blocks of the FLSF four-helix bundle were two αα-hairpin peptides.

If this scenario for the origin of the FLSF were true, then there might still exist proteins with residual sequence similarity between helix pairs A/B and C/D beyond just the residues involved in the metallocenter. Indeed, the four-helix bundle of rubrerythrin from *D. vulgaris* has ~29% internal sequence similarity (Kurtz and Prickril 1991), providing evidence that the two halves of rubrerythrin may have come from the same ancestral two-helix peptide. Interesting in the light of rubrerythrin having features resembling an ancestral state of the FLSF is that all rubrerythrins are found in anaerobic (or at least oxygen-sensitive) organisms, which are believed to be the earliest forms of life on Earth. It has therefore been proposed that rubrerythrin may resemble one of the earliest forms of defense against reactive oxygen species (Gomes *et al.* 2001).

The possibility that rubrerythrin may retain many ancestral features of the FLSF has its shortcomings though. First, rubrerythrin's metallocenter is asymmetric with seven metal ligating residues. So from where did that additional ligating residue originate? In Chapters 4 and 6, I use the fact that this seventh metal-ligating residue sits within a π-helical segment as critical information to address this question and also to further understand the role of rubrerythrin in the evolution of the FLSF. Another observation seemingly in conflict with the notion that the ancestral form of the FLSF looked much like modern day rubrerythrins is that other proteins with little sequence similarity to rubrerythrins that contain the canonical six metal-ligand motif also have an equivalent level of internal similarity as rubrerythrins. The discussion in Chapter 6 addresses this apparent conflict.

Contents of the Dissertation

This dissertation contains six remaining chapters. Chapters 2 through 5 have been published and Chapter 6 is *in press*. Chapters 2, 4 and 6 constitute full research articles. Chapter 3 is a *Report* published in the "Microbial Methane Cycle" special issue of *Environmental Microbiology Reports.* Chapter 5 is a short, 800 word Brevia published in *Science* detailing the discovery of a novel post-translation modification. Chapter 7 provides an overall summary of the research and an outlook for future directions. Below is a series of short introductions to Chapters 2 through 6 outlining the motivation behind each study, the important questions addressed and how it relates to the other chapters.

Chapter 2: "Kinetic characterization of the soluble butane monooxygenase from *Thauera butanivorans*, formerly *'Pseudomonas butanovora'* " Richard B. Cooley, Bradley L. Dubbels, Luis A. Sayavedra-Soto, Peter J. Bottomley, and Daniel J. Arp. Published in *Microbiology*, **155**, 2086-2096, 2009.

The chapter was a logical follow up to Dubbels et al. (2007), a study published immediately prior to my joining the Arp laboratory. In that study, the first purification and *in vitro* characterization of a soluble diiron alkane monooxygenase from an organism capable of growing on alkanes larger than methane was reported. In this chapter, we follow up with more detailed kinetic characterizations of sBMO and make comparisons with sMMO. During this time, I also attempted to crystallize the hydroxylase component of sBMO. After nearly three years and approximately 50,000 different trials, I was unable to produce crystals. Nevertheless, the biochemical data presented in this chapter provide insights into two important areas of BMMs: substrate specificity and activation of the hydroxylase by the regulatory subunit. These results provided also the framework for the work presented in Chapter 3.

Chapter 3. "Growth of a non-methanotroph on natural gas: ignoring the obvious to focus on the obscure." Richard B. Cooley, Peter J. Bottomley, and Daniel J. Arp. Published in *Environmental Microbiology Reports*, **1**(5), 408-413, 2009.

A question rarely asked about non-methanotrophic alkane utilizing organisms is, "What is their natural habitat?" For methanotrophs, the answer to this question is quite simple as natural gas seeps and areas with methanogens contain large quantities of methane. In contrast, the only natural source of alkanes larger than C_1 is also natural gas and in this context the larger alkanes are only present in low quantities. Given that *T. butanivorans* cannot grow on methane, we were curious to see if this organism was adapted to live in an environment with large quantities of methane and only small quantities of its growth substrates (ethane, propane and butane). The kinetic data outlined in Chapter 2 suggested to us that sBMO would be able to "filter out" the small quantities of growth substrates even in the presence of an overwhelming quantity of methane. The results of this chapter show that *T. butanivorans* has an uncanny ability to thrive in this type of environment and therefore could have evolved in such an unusual ecological niche.

Chapter 4. "Evolutionary origin of a secondary structure: π-helices as cryptic but widespread insertional variations of α-helices enhancing protein functionality". Richard B. Cooley, Daniel J. Arp, and P. Andrew Karplus. Published in *Journal of Molecular Biology*, **404**(2), 232-246, 2010.

The results from Chapter 3 suggested to us that the assimilation of a finely tuned sBMO into the genome of *T. butanivorans* was a critical step in the evolution of this organism. We therefore became interested in understanding the origins of sBMO and the whole BMM enzyme family. At the same time, I was continuing a project I started with Dr. P. Andrew Karplus in a first year rotation project regarding the evolutionary origin of π-helices. In an unexpected convergence of projects, it became

evident the evolution and function of BMMs was strongly shaped by the formation π-helices via the insertion of single amino acids into α-helices. Surveying the entire Protein Data Bank revealed that BMMs were an example of a more general mechanism of protein evolution, as the function of nearly 15% of all proteins have been influenced in such a way. This manuscript changes our understanding of what π-helices are, where they come from and how we can use them to gain insight into proteins structure, function and evolution. We use these insights to predict that a previously uncharacterized rubrerythrin-like protein (at that point named erythrin) has a novel diiron metallocenter. This work also forms the basis for Chapters 5 and 6.

Chapter 5. "A diiron protein autogenerates a Valine-Phenylalanine crosslink." Richard B. Cooley, Timothy W. Rhoads, Daniel J. Arp, and P. Andrew Karplus. Published in *Science*, **332**(6032), 929, 2011.

In this manuscript, we follow up the specific prediction presented in Chapter 4. After expressing, purifying, crystallizing and solving the structure of this previously uncharacterized rubrerythrin-like protein, we discovered that it had the ability to make an unprecedented post-translational modification so unusual and with interest to a broad audience that it was published in *Science*. Due to the length restrictions of a *Brevia*, we were not able to include many other interesting aspects of these studies. These additional aspects form the basis for Chapter 6.

Chapter 6. "Symerythrin structures at atomic resolution and the origins of rubrerythrins and the ferritin-like superfamily." Richard B. Cooley, Daniel J. Arp, and P. Andrew Karplus. Published in *Journal of Molecular Biology*, in press, 2011.

After describing the remarkable crosslink found in symerythrin, this manuscript provides a detailed analysis of the overall structure of this novel protein and its unusual metallocenter in a variety of states, including the diferric, diferrous and

azide-bound diferric states. These results of symerythrin are compared with its closest known homologs in the rubrerythrin family. Included in the manuscript is a new model for the evolutionary origins of the FLSF, suggesting that contrary to previous assumptions, there have been at least two independent origins of the single chain core four-helix bundle of the FLSF.

Finally, in Chapter 7, I place in context and discuss the impacts of these findings. While these results answer several questions regarding the structure, function and evolution of ferritin-like proteins, they also open the door to new and unexpected areas ripe for future research. Such areas of future research are outlined, particularly those involving *in vitro* follow up studies of symerythrin as well as *in vivo* analyses of it in its native organisms *Cyanophora paradoxa* and *Gloeobacter violaceus*.

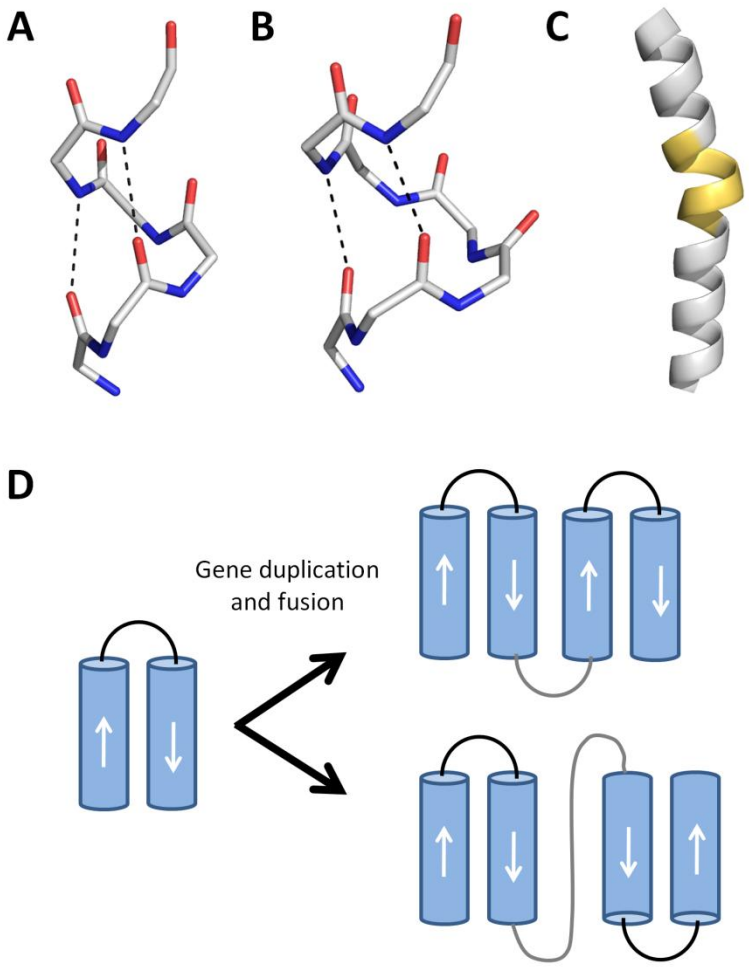


Figure 1.1. Helical structures in proteins and the "proteins from peptides" origin of proteins. (A) An α-helix showing main chain hydrogen bonds (black dashed lines) between residues four apart in sequence. Carbon, nitrogen and oxygen atoms are colored gray, blue and red, respectively. (B) A naturally occuring π-helix with main chain hydrogen bonds between residues five apart in sequence. (C) Cartoon representation of a π-helical segment (yellow) embedded within a longer α-helix (gray). (D) Formation of a four-helix bundle protein by the gene duplication and fusion of an αα-hairpin peptide (left, helices represented as cylinders with direction of the peptide indicated by the white arrows). Depending on the nature of the αα-hairpin and how it is genetically fused, the resulting core four-helix bundle can have different three dimensional topologies, giving rise to different superfamilies both having a four-helix bundle fold.

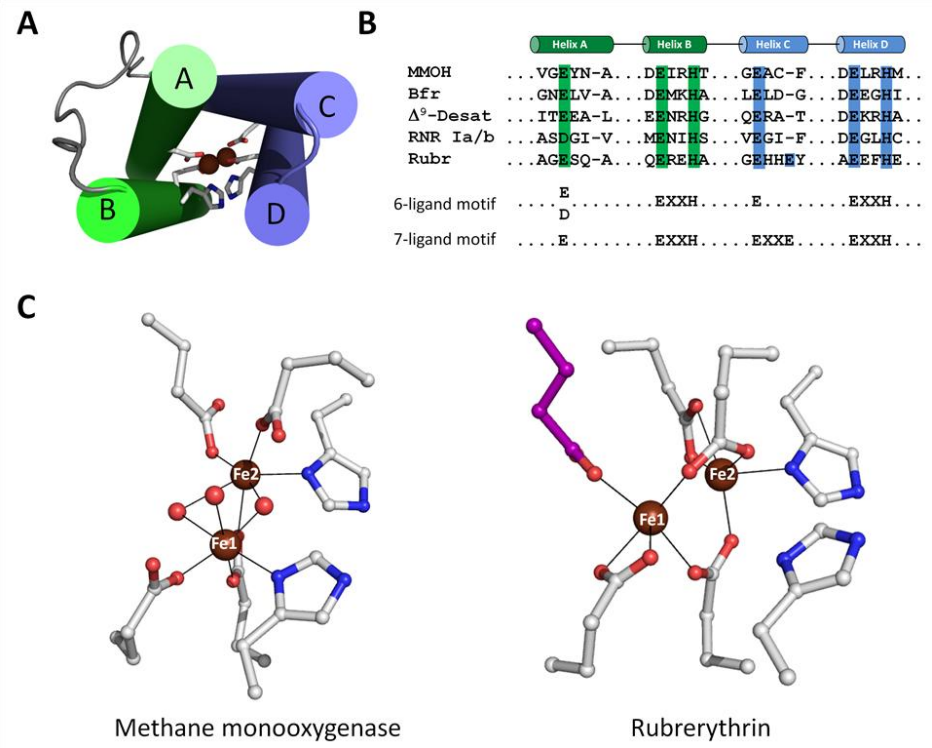


Figure 1.2. Overview of the ferritin-like superfamily. (A) The core four-helix bundle characteristic of the FLSF showing two metals (brown spheres) bound by six metal-ligating residues (coloring of side chains is the same as Fig. 1.1A). (B) Sequence alignment of representative FLSF proteins demonstrating the conservation of metal ligating residues contributed by helix pairs A/B (green) and C/D (blue). MMOH: methane monooxygenase hydroxylase, Bfr: bacterioferritin, Δ^9-Desat: Δ^9-desaturases, RNR: ribonucleotide reductase, Rubr: rubrerythrin. (C) Diiron site of methane monooxygenase with six metal-ligating residues (left, PDB 1mty) and of rubrerythrin with seven metal-ligating residues (right, PDB 1lkm). Carbon atoms of the additional seventh metal-ligating residue in rubrerythrin are colored purple. Metal-ligand interactions are indicated by solid black lines.

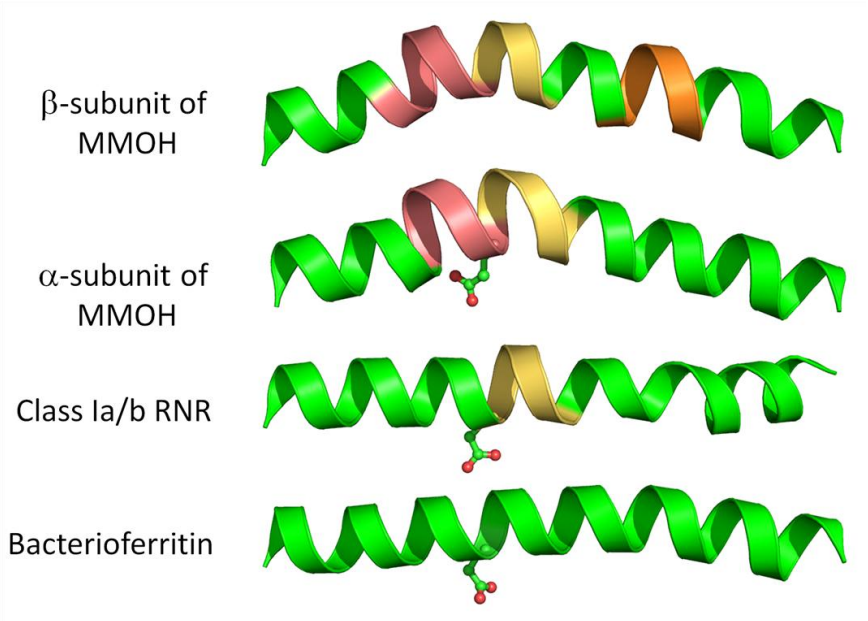


Figure 1.3. Helical perturbations in helix C of representative FLSF members. α-helices are colored green, while π-helices are colored yellow, salmon and orange.

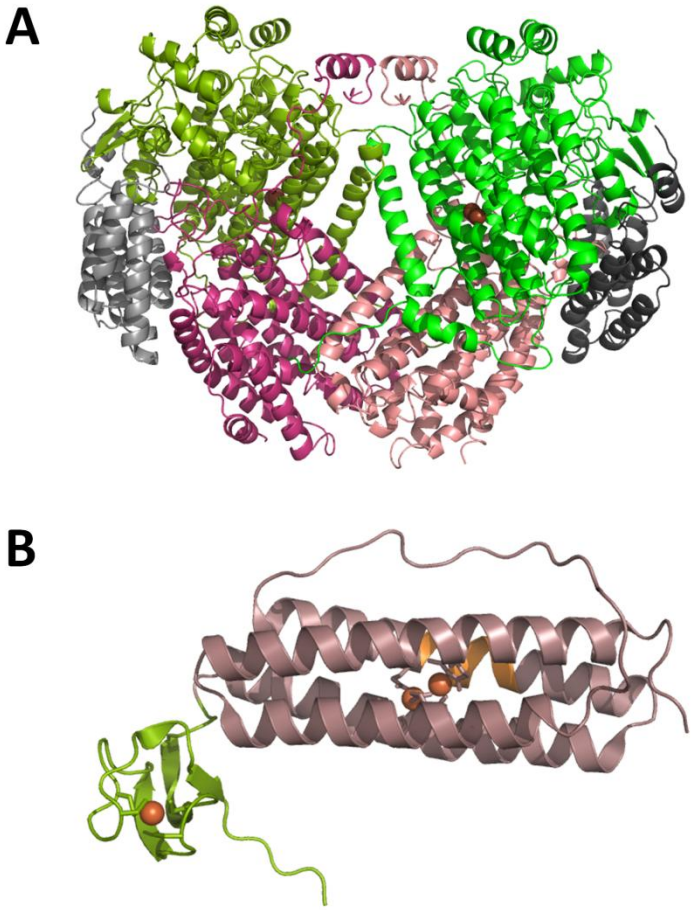


Figure 1.4. Structure of soluble methane monooxygenase hydroxylase (MMOH) and rubrerythrin. (A) α-, β- and γ-subunits of MMOH from *Methylococcus capsulatus* (Bath) are shown in shades of green, pink and gray, respectively. PDB 1mty. (B) One chain of the structure of rubrerythrin from *Desulfovibrio vulgaris* with the four-helix bundle and the C-terminal rubredoxin domain colored in light purple and green, respectively. The π-helical segment in the four-helix bundle of rubrerythrins is colored orange. PDB 1lkm. In both panels, iron atoms are shown as brown spheres.

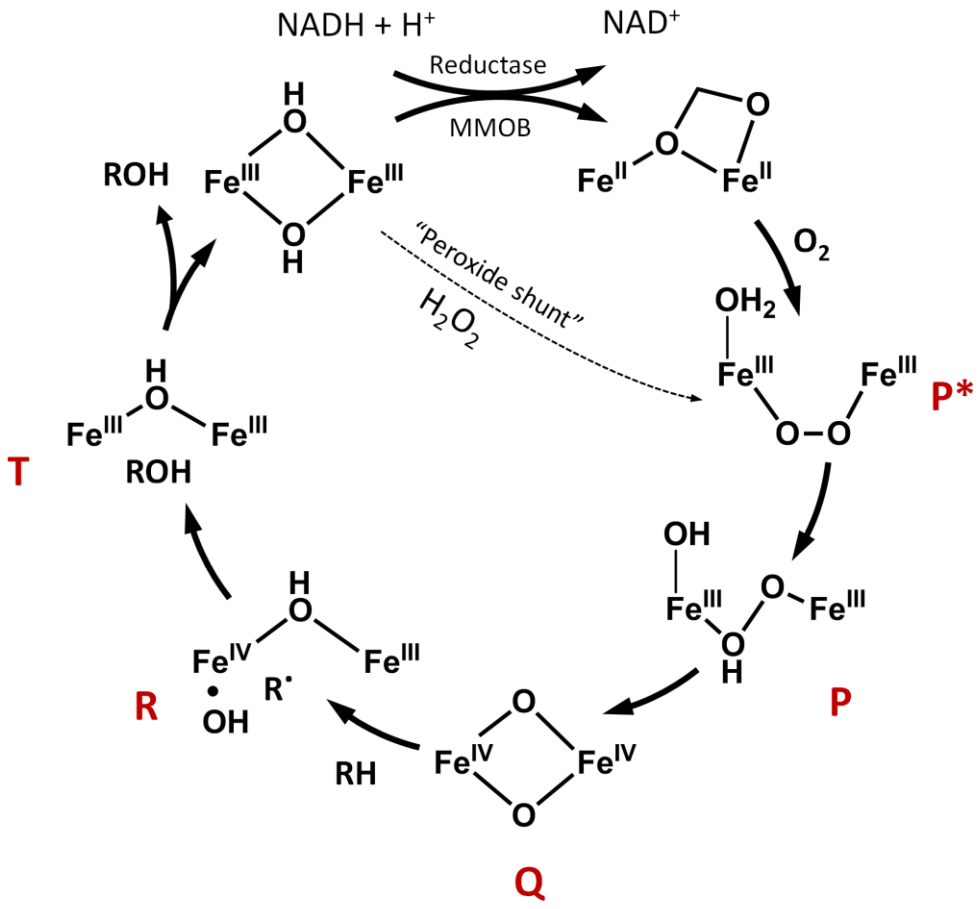


Figure 1.5. Proposed mechanism for the activation of molecular oxygen and hydroxylation of alkanes by methane monooxygenase. Adapted from (Brazeau and Lipscomb 2000).

Chapter 2

Kinetic characterization of the soluble butane monooxygenase from *Thauera butanivorans*, formerly *'Pseudomonas butanovora'*

Richard B. Cooley, Bradley L. Dubbels, Luis A. Sayavedra-Soto, Peter J. Bottomley, and Daniel J. Arp

Published in *Microbiology* (2009), 155, 2086-2096

Abstract

Soluble butane monooxygenase (sBMO), a three-component di-iron monooxygenase complex expressed by the C_2-C_9 alkane utilizing bacterium *Thauera butanivorans*, was kinetically characterized by measuring substrate specificities of alkanes C_1-C_5 and product inhibition profiles. sBMO has high sequence homology with soluble methane monooxygenase (sMMO) and shares a similar substrate range including gaseous and liquid alkanes, aromatics, alkenes and halogenated xenobiotics. Results indicated that butane was the preferred substrate (defined by k_{cat}/K_m ratios). Relative rates of oxidation for alkanes C_1-C_5 differed minimally; implying substrate specificity is heavily influenced by differences in substrate K_m's. The low micromolar K_m for linear alkanes C_2-C_5 and millimolar K_m for methane demonstrate sBMO is 2-3 orders of magnitude more specific for physiologically relevant substrates of *T. butanivorans*. Methanol, the product of methane oxidation and also a substrate itself, was found to have similar K_m and k_{cat} values as methane. This inability to kinetically discriminate between the C_1 alkane and C_1 alcohol is observed as a steady state concentration of methanol during the two step oxidation of methane to formaldehyde by sBMO. Unlike methanol, alcohols with chain length C_2-C_5 do not compete effectively with their respective alkane substrates. Results from product inhibition experiments suggest the geometry of the active site is optimized for linear molecules 4-5 carbons in length and is influenced by the regulatory protein component B (BMOB). The data suggest that alkane oxidation by sBMO is highly specialized for the turnover of C_3-C_5 alkanes and the release of their respective alcohol products. Additionally, sBMO is particularly efficient in preventing methane oxidation during growth on linear alkanes $\geq C_2$ despite its high sequence similarity with sMMO. These results represent the first kinetic *in vitro* characterization of the closest known homolog of sMMO.

Introduction

The use of aliphatic alkanes as a sole source of carbon and energy is carried out by a diverse range of aerobic prokaryotes (Shennan 2006). To initiate growth on alkanes C_2-C_9, the Gram-negative β-proteobacterium *Thauera butanivorans*, formerly called *'Pseudomonas butanovora'* (Dubbels *et al.* 2009), expresses a carboyxlate bridged non-heme di-iron monooxygenase, commonly referred to as soluble butane monooxygenase (sBMO) (Takahashi *et al.* 1980; Sluis *et al.* 2002). sBMO belongs to a family of bacterial multicomponent monooxygenases which includes soluble methane monooxygenases (sMMO), phenol hydroxylases (PH) and aromatic/alkene monooxygenases (TMO) (Sluis *et al.* 2002; Leahy *et al.* 2003). Due to their unusually large substrate ranges, these powerful oxidizers are of particular interest for their potential in bioremediation (Enzien *et al.* 1994; Parales *et al.* 2002; Smith and Dalton 2004; Halsey *et al.* 2007) and their ability to serve as industrial biocatalysts (Parales *et al.* 2002; Burton 2003).

The expression of sBMO is induced by the presence of 1-butanol and butyraldehyde, and is repressed by lactate and succinate (Sayavedra-Soto *et al.* 2005). When *T. butanivorans* is grown on butane, sBMO oxidizes butane primarily to 1-butanol (>85% terminal oxidation), which is then further oxidized to butyrate via butyraldehyde (Arp 1999). Further metabolism of butyrate probably proceeds through butyryl-CoA prior to β-oxidation (Arp 1999). The metabolic pathway of odd chain or subterminally oxidized alkanes has not yet been fully characterized in *T. butanivorans*; however it is known that the metabolism of odd and even chain alkane growth substrates is differentially controlled (Doughty *et al.* 2007). Although sBMO is capable of oxidizing methane to methanol (Halsey *et al.* 2006), *T. butanivorans* cannot assimilate C_1 compounds into biomass.

sBMO is a three component complex that was recently purified and found to contain a similar architecture to that of sMMO: (i) a 250 kDa hydroxylase component (BMOH) with $(\alpha\beta\gamma)_2$ subunit composition that, by analogy with sMMO, contains the di-iron active site in the α-subunit, (ii) a 40 kDa flavo-iron sulfur-containing reductase

(BMOR) that shuttles electrons from NADH to the active site of BMOH and (iii) a 15 kDa regulatory protein (BMOB) that enhances the overall activity of the complex (Dubbels *et al.* 2007). Sequence analysis showed BMOH to be most closely related to sMMO (65, 42 and 38% amino acid identity in the α, β and γ-subunits of BMOH, respectively) (Sluis *et al.* 2002). Additionally, sequence alignment and structural modeling show that all residues directly lining the active site cavity of BMOH are the same as MMOH. Despite the similarities, several biochemical observations indicate sBMO represents a class distinct from the sMMO family: (i) methanol accumulation ceases during methane oxidation once ~20-50 μM has been reached (Halsey *et al.* 2006), whereas sMMO can continue to accumulate methanol, (ii) sBMO is predominantly a terminal hydroxylator of intermediate chain length alkanes (Dubbels *et al.* 2007), whereas sMMO forms secondary alcohols (Froland *et al.* 1992), (iii) product regioselectivity is minimally altered by the presence of BMOB (Dubbels *et al.* 2007), unlike sMMO (Froland *et al.* 1992), (iv) catalase is required to maintain hydroxylase activity during steady state turnover (Dubbels *et al.* 2007) but not to maintain sMMO activity, and (v) the use of peroxide in the 'peroxide shunt' mechanism of substrate oxidation is 3 orders of magnitude less efficient in sBMO than sMMO (Dubbels *et al.* 2007).

The structural basis that defines the catalytic properties of these alkane monooxygenases is still poorly understood, most notably due to a lack of enzyme isolations and subsequent *in vitro* characterizations (van Beilen and Funhoff 2007). Even though crystal structures of apo-MMOH and product-bound MMOH have been reported (Rosenzweig *et al.* 1997; Sazinsky and Lippard 2005), the fully buried active site does not appear to be accessible or large enough to accommodate known substrates such as naphthalene or biphenyl. A few mutational studies of BMOH and MMOH have provided insight into this field, however. Substitution of residue Gly113 in the α-subunit of BMOH to the equivalent residue of MMOHα (Asn) results in a shift of properties that are more characteristic of sMMO, such as increased methanol accumulation and an alteration in product regioselectivity from primary to secondary

hydroxylation (Halsey *et al.* 2006). Alterations in the "Leucine Gate" residue of MMOHα (Leu110) have shown that it is critical in defining the regioselectivity of sMMO (Borodina *et al.* 2007).

Given the wide substrate range of alkane monooxygenases, a comprehensive kinetic characterization of sBMO would help provide insights into understanding the mechanism of substrate specificity. Here, we report a characterization of the substrate specificity for sBMO and show that, despite having the same active site residues as sMMO, sBMO is poorly suited for methane yet highly optimized for linear molecules 4-5 carbons in length. Since measurable substrate turnover is possible without BMOB, we were also able to directly observe a function for BMOB in catalysis using product inhibitors. Such types of experiments have not been possible with sMMO because substrate turnover is too slow without MMOB.

Methods

Chemicals. Gaseous alkanes were of reagent grade. Methane was purchased from Airco. Ethane was purchased from Matheson. Propane and butane were purchased from Airgas. Pentane was purchased as reagent grade from VWR. NADH was purchased from Research Organics and residual ethanol was removed by repeated lyophilization in 25 mM PIPES pH 7.2. Bovine liver catalase was purchased from Sigma. All other chemicals were obtained from Aldrich.

Bacterial cultivation and BMOH/BMOR purification. Wild type (WT) *T. butanivorans* was grown on butane as described previously (Dubbels *et al.* 2007) with the exception that 100 μM Fe^{3+}-EDTA was used as an iron source instead of 500 μM $FeSO_4 \cdot 7H_2O$. Mutant strain G113N, which has residue 113 of the α-subunit of BMOH altered from Gly to Asn (Halsey *et al.* 2006), was grown identically to that of WT except that 50 μM $MnCl_2$ and 10 μM Na_2CO_3 were added to the medium to facilitate the downstream metabolism of secondary alcohols. WT BMOH and BMOR components of sBMO were purified as previously described (Dubbels *et al.* 2007). The G113N BMOH was purified identically to that of the WT hydroxylase. The

alternative iron source yielded BMOH that exhibited much tighter elution profiles off the Sephacryl S-300 HR gel filtration column due to significantly reduced self-aggregation as determined by dynamic light scattering. Additionally, the iron content of BMOH, as measured by the ferrozine spectrophotometric assay (Percival 1991), contained 2.1 – 2.4 irons per active site rather than 1.4 – 1.8 reported previously (Dubbels *et al.* 2007). Typical preparations of the enzyme complex resulted in activities ranging from 400-700 nmoles min^{-1} mg^{-1}. Protein concentrations were determined by optical absorption at 280 nm using extinction coefficients of 2.2 and 0.56 ml mg^{-1} cm^{-1} for BMOH and BMOR, respectively.

Development of recombinant BMOB expression system. Purification of BMOH and BMOR from the native host yielded high quantities of highly active protein, and so no recombinant expression system was needed. However, purification of BMOB from the native host required several additional purification steps and the yields were not always satisfactory. We therefore chose to develop a recombinant expression system for BMOB that provided a simpler purification process with significantly increased yields.

To create the recombinant BMOB expression system, genomic and plasmid DNA were first isolated according established protocols (Ausubel *et al.* 2003). Primers for the amplification of BMOB were as follows: bmobfNdeI 5´-CAGGGGCAGAC**CATATG**TCAAACGT-3´ and bmobrBamHI 5´-CGCACCGGTGTGT**GGATCC**AAACCT-3´. Bases indicated in bold are the restriction sites included for subsequent cloning into the expression vector. A standard PCR reaction was carried out with the above primers with Taq (Promega). The PCR product was gel purified with QIAEX II (Qiagen) and restricted with NdeI and BamHI (Promega). Gel purification of the restriction digest was performed as stated above and the isolated PCR product was ligated into NdeI and BamHI digested pT7-7 (Tabor and Richardson 1985). The ligation reaction was transformed into chemically competent *Escherichia coli* JM109 (Sambrook *et al.* 1989). Plasmid DNA was isolated and the sequence of *bmob* was confirmed by DNA sequencing at the Center for Genome

Research and Biocomputing at Oregon State University. The resulting plasmid, pBD400, was transformed into chemically competent *E. coli* BL21(DE3) (Novagen) for subsequent expression and purification.

Expression and purification of recombinant BMOB. *E. coli* BL21(DE3) cells containing the pBD400 plasmid were grown in 3 l of LB medium in the presence of 100 µg ampicillin ml^{-1} at 37 ^{o}C to an $O.D._{600}$ = 0.8. Protein expression was stimulated by the addition of 1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG). After 3 h of further growth, cells were centrifuged at 5,000 ***g*** for 20 min and stored at -70 ^{o}C. Typically yields of 4-5 g of cell paste l^{-1} were obtained.

All purification steps of recombinant BMOB (rBMOB) were performed at 4 ^{o}C. Frozen *E. coli* BL21(DE3) cell paste (15 g) was resuspended in 25 mM PIPES pH 7.2 to a total volume of 40 ml containing 1000 units of DNase I (Sigma). Cells were lysed by two passes through a French pressure cell disrupter at 52000 Pa, and then centrifuged at 10,000 ***g*** for 20 min. The supernatant was carefully decanted, diluted to 60 ml and centrifuged at 150,000 ***g*** for 2 h. The resulting cell free lysate was decanted, pH adjusted to 7.2, and loaded onto a DEAE-Sepharose FF column (90 mm x 30 mm) pre-equilibrated with 25 mM PIPES pH 7.2 at 2 ml min^{-1}. The column was washed with 5 column volumes of the same buffer at a linear flow rate of 2 ml min^{-1}, after which a 0 – 0.4 M KCl linear gradient was applied over 4 column volumes. Fractions containing rBMOB eluted between 0.2 – 0.3 M KCl. These fractions were pooled together, repeatedly dialyzed against 25 mM PIPES pH 7.2 with 150 mM KCl, and concentrated to 3 ml total volume via ultrafiltration. The concentrate was then applied to Superdex 75 FF gel filtration column (550 mm x 25 mm) pre-equilibrated with the same buffer at a linear flow rate of 0.5 ml min^{-1}. Fractions containing purified rBMOB were pooled, dialyzed against 25 mM PIPES pH 7.2, concentrated to 2 mM, flash frozen in liquid nitrogen and stored at -70 ^{o}C. From 15 g of cells, 200-250 mg of purified rBMOB was obtained. rBMOB was found to display the same enhancement of sBMO activity as native BMOB, alter the product distribution similarly, have the same mobility on SDS-PAGE, and have identical molecular weight

as determined by MALDI-TOF mass spectrometry. rBMOB also displayed the same partial dimerization through intermolecular di-sulfide linkage as native BMOB (Dubbels *et al.* 2007). As such, all reactions in this study were performed with the recombinant form of BMOB. Concentrations were determined by optical absorption at 280 nm using an extinction coefficient of 1.2 ml mg^{-1} cm^{-1}.

Determination of methane K_m. Vials (7.7 ml) containing a stir bar and a 0.5 ml aliquot of 0.1 µM BMOH, 0.3 µM BMOB, 0.6 µM BMOR and 2400 units of catalase ml^{-1} in 25 mM PIPES pH 7.2 were crimp-sealed with butyl rubber septa. Methane was added directly as an overpressure. For higher concentrations, however, methane was used to refill the head space, to which additional methane and 20% (v/v) oxygen was added as an overpressure. This mixture was allowed to equilibrate at 25 ^{o}C for 5 min with gentle stirring. To initiate the reaction, NADH was added via a gas tight syringe to a final concentration of 1 mM. Samples (2 µl) were removed and injected into a Shimadzu GC-8A gas chromatograph (GC) equipped with a flame ionization detector and a stainless steel column packed with Porapak Q (Alltech) (80/100 mesh). Although methanol accumulation halts after the production of ~30 µM methanol (Halsey *et al.* 2006), linear rates could be obtained for the first 5-10 min prior to slowing of the reaction. Methane concentrations were based on calculations using a Henry's Constant of 0.0014 M atm^{-1} at 25 ^{o}C (Lide and Frederikse 1995). The K_m of methane was determined by fitting the initial rates of reactions as a function of substrate concentration using Origin Pro 7.5 (OriginLab, Northampton, MA) according to the following equation:

$$v_o = \frac{V \times [Methane]}{K_m + [Methane]} \qquad \text{(eq. 2.1)}$$

where v_o is the initial rate of reaction, V is the maximal rate of methanol accumulation under saturating concentrations of substrate, [Methane] is the aqueous concentration of methane and K_m is the Michaelis constant. Methanol concentrations were determined

based on a standard curve produced from authentic methanol in the same buffered conditions. All reported errors are standard deviations from three independent replicates.

Determination of the K_m for alkanes C_2-C_5. K_m measurements for longer chain alkanes could not be performed due to slow diffusion rates and limited GC detection at sub-micromolar concentrations of product. Alternatively, the K_m's for alkanes were measured indirectly by competition with nitrobenzene. In a sealed quartz cuvette, 0.5 ml of a mixture containing 0.06 µM BMOH, 0.18 µM BMOB, 0.36 µM BMOR, 2400 units of catalase ml^{-1} and 1 mM nitrobenzene in 25 mM PIPES pH 7.2 was incubated for 5 min with varying amounts of gaseous alkane added to the head space. Reactions were initiated by the addition of 1 mM NADH, and monitored by the formation of *p*-nitrophenol (ε_{404nm} = 15 mM^{-1} cm^{-1}) at 404 nm in a Beckman DU-640 spectrophotometer. The initial linear portion of the reaction curve (~1-2 min) was taken as the initial reaction rate of nitrobenzene formation when in competition with the alkane. The K_m of the alkane was determined by fitting the following equation:

$$v_o = \frac{V \times [S]}{K_{m2}(1 + [a]/K_m) + [S]} \qquad \text{(eq. 2.2)}$$

where v_o is the initial rate of nitrobenzene formation, V is the maximal reaction rate, [S] is the initial concentration of nitrobenzene, K_{m2} is the Michaelis constant for nitrobenzene (40 µM), [a] is the aqueous concentration of the alkane, and K_m is the Michaelis constant for the alkane. Aqueous concentrations of ethane, propane, butane and pentane were determined using Henry's constants of 0.0019, 0.0014, 0.0011 and 0.0008 M atm^{-1}, respectively. To ensure the accuracy of this method, a similar analysis of methane competition with nitrobenzene was performed in an identical manner as that of ethane, propane, butane and pentane except that only 50 µM nitrobenzene was used instead of 1 mM.

Determination of alcohol inhibition constants. A sealed quartz cuvette containing 0.25 µM BMOH, 0.75 µM BMOB, 1.5 µM BMOR, 2400 units ml^{-1} catalase and nitrobenzene in 25 mM PIPES pH 7.2 was incubated with varying amounts of primary and secondary alcohols for 5 min prior to reaction initiation with 1 mM NADH. Three titrations per alcohol were performed, each with different concentrations of nitrobenzene (50, 100 or 200 µM) in order to determine the type of inhibition. Linear rates of *p*-nitrophenol formation were monitored by the increase in absorption at 404 nm for 2 min. For experiments without BMOB, reactions were monitored for 5 min due to slower rates of product formation. Competitive and uncompetitive inhibition constants were modeled using the mixed inhibitory equation (Cornish-Bowden 1995):

$$v_o = \frac{V \times [S]}{K_m(1 + [i]/K_{ic}) + [S](1 + [i]/K_{iu})} \quad \text{(eq. 2.3)}$$

where [i] is the concentration of inhibitor and K_{ic} and K_{iu} are the competitive and uncompetitive inhibition constants, respectively. All observed inhibition data was fit to equation 3 with an $R^2 > 0.97$. Stock nitrobenzene concentrations were determined using an extinction coefficient of 7800 M^{-1} cm^{-1} at 268 nm (Zhu *et al.* 2007).

Formaldehyde analysis. Quantification of formaldehyde was performed by derivatization with acetoacetanilide and subsequent fluorescence detection as previously described (Li *et al.* 2007). Concentrations of formaldehyde were based on a standard curve made from a stock solution with a known concentration. Formaldehyde stock concentrations were determined by titration of a diluted sample with a molar excess of iodine (I_2) in the presence of 0.2 M NaOH. After 15 min, the excess I_2 was acidified with H_2SO_4 and back-titrated with 0.1 M sodium thiosulfate in the presence of 0.02% (w/v) starch indicator. The same procedure was repeated without formaldehyde. The difference in thiosulfate needed to titrate the two solutions was used to determine the original concentration of formaldehyde.

Results

Oxidation of alkanes by sBMO. In order to further understand the substrate specificity of sBMO, kinetic parameters K_m and k_{cat} were determined for linear alkanes C_1-C_5. Although methane is known to be a substrate of sBMO, the accumulation of its product, methanol, was previously shown to cease once methanol concentrations reached ~20-30 µM (Halsey *et al.* 2006). However, at concentrations ranging from 0-15 µM, the rates of methanol formation were both constant and dependent on the concentration of aqueous methane. These rates became saturated above 5 mM methane, allowing for a direct measurement of the K_m for methane (Fig. 2.1a). Doubling the enzyme concentration along the first-order region of the titration curve resulted in a doubling of the reaction rate, indicating the reactions were not diffusion limited. In contrast to the low micromolar K_m of sMMO for methane (3 to 13 µM) (Green and Dalton 1986; Nesheim and Lipscomb 1996), the K_m of sBMO for methane was 1.10 ± 0.14 mM (Table 2.1). Determination of K_m's for the growth substrates C_2-C_5 could not be performed by direct measurement due to limited diffusion rates at low concentrations. Alternatively, K_m's were derived by competition experiments between nitrobenzene and the alkane substrates according to equation 2 (Fig. 2.1b). The K_m for methane was also measured in this manner, giving similar results as the direct measurement (1.25 ± 0.12 mM, data not shown).

The K_m values listed in Table 2.1 indicate a sharp transition of large magnitude from methane to ethane, but less of a difference between propane, butane and pentane. Although preferential binding of substrates is likely to be heavily influenced by the 'hydrophobic effect', this cannot explain the drop of nearly 3 orders of magnitude in K_m from methane to ethane (if K_m constants are assumed to correlate with binding affinity). Alternatively, given the Bi Uni Uni Bi Ping Pong kinetic mechanism proposed for sMMO (Green and Dalton 1986), the observed K_m for the alkane should be heavily influenced by the rate of alcohol release (Leskovac 2003). As alcohol release becomes rate limiting, which has been reported for sMMO (Lee *et al.* 1993;

Wallar and Lipscomb 2001), the K_m of the substrate drops below the K_d. Such a scenario would explain the large drop in the K_m for ethane. Similar arguments have been proposed for protein kinase A (Werner *et al.* 1996). The observed maximal turnover rates for methane and ethane listed in Table 2.1 also support this hypothesis. If the rate limiting step of the reaction was turnover, the cleavage of a C-H bond for ethane should be 3 orders of magnitude faster since its dissociation energy is about 4 kcal mol^{-1} weaker than that of methane (Korth and Sicking 1997; Zheng and Lipscomb 2006). Such a dramatic increase in ethane turnover was not observed in sBMO, nor observed in earlier studies with sMMO (Green and Dalton 1986).

Total turnover rates for substrates C_1-C_5 were also determined (Table 2.1). The individual rates of the different alcohol isomers from the oxidation of alkanes C_3-C_5 were also proportional to previously reported product distributions (Dubbels *et al.* 2007). The ratio k_{cat}/K_m indicated sBMO to be most specific for butane; however propane and pentane were not substantially different. sBMO was nearly 280 fold more specific for ethane than methane. To confirm the calculated substrate specificities, competition experiments were performed with a mixture of equal concentrations of alkanes $C_1 – C_5$. Although methanol formation was not detected, the relative amounts of each alcohol $C_2 – C_5$ produced was similar to those predicted by the measured k_{cat}/K_m ratio (Table 2.1), providing further evidence that the sBMO enzyme is optimized for butane as a substrate while minimizing C_1 oxidation.

Product inhibition of sBMO. Halogenated alcohols were shown by x-ray crystallography to bind directly in the active site of sMMO with the oxygen atom of the alcohol group bridging the di-iron center (Sazinsky and Lippard 2005). The binding of primary and secondary alcohols to sBMO was characterized in order to gain additional insight into the approximate geometry of the active site. Methanol inhibition of nitrobenzene oxidation was modeled with a mixed inhibitory scenario (equation 3), which indicated the competition was purely competitive and relatively weak (K_{ic} = 1.25 ± 0.06 mM). Previously, our laboratory hypothesized that the plateau of methanol accumulation at ~30 µM during methane oxidation was the result

of strong product inhibition (Halsey *et al.* 2006). However, the millimolar K_{ic} observed for methanol and similar K_m for methane does not support the conclusion that methanol would effectively inhibit methane oxidation at such low concentrations. More insights into the special case of methane oxidation are discussed later in this section.

Primary alcohols $C_1 - C_6$ displayed pure competitive inhibition with nitrobenzene. An average increase of 0.69 ± 0.1 kcal mol^{-1} per methylene group was observed for primary alcohols $C_1 - C_5$ (Fig. 2.2, shaded bars). Mutational analyses of hydrophobic residues in the interior of proteins have suggested that each methylene group contributes approximately 1.1 ± 0.5 kcal of stability mol^{-1} methylene $group^{-1}$ (Pace *et al.* 1996). It would appear, therefore, that the binding of alcohols C_1-C_5 to the fully buried, hydrophobic active site of BMOH is heavily influenced by this 'hydrophobic effect'. After C_5 however, binding affinity drops until C_7, after which the affinity increases. Interestingly, the tighter inhibition constants for these larger alcohols are accompanied by a distinct change in inhibition from competitive to mixed, suggesting the alcohol is able to bind to both free enzyme (E) and the enzyme-substrate (ES) complex. Incubation of the enzyme complex with these longer chain alcohols with and without NADH for 1 h did not yield any significant loss in activity compared to the control after dialysis, indicating that the change in inhibition type is not a result of enzyme inactivation. Given that product bound structures of MMOH have demonstrated alternative small molecule binding cavities (Sazinsky and Lippard 2005), it is possible that the uncompetitive nature of primary alcohols larger than C_6 is derived from the preferential binding to alternative hydrophobic pockets rather than the active site.

Inhibition constants for secondary alcohols $C_1 - C_9$ were also measured (Fig. 2.2, white bars). Three distinct observations were made in comparison with primary alcohols. First, all secondary alcohols bound with less affinity than their primary counterparts, indicating that the active site of sBMO is optimized for linear molecules. Second, the break in decreasing K_{ic}'s is found between C_4 and C_5, rather than C_5 and

C_6 as for primary alcohols. Third, the break from competitive to mixed inhibition occurs at C_8 rather than C_7, indicating secondary alcohols do not bind to the proposed alternative binding pockets as well as primary alcohols. Lastly, branched alcohols, such as 2-methyl-2-butanol (K_{ic} = 1.45 ± 0.08 mM), bind poorly, providing further evidence that the active site of sBMO is structurally optimized for linear molecules.

Effect of BMOB. Component B of the sMMO system (MMOB) has profound effects on both the rate of MMOH catalysis and product distribution (Wallar and Lipscomb 1996). Our previous characterization of purified sBMO revealed that BMOB had little effect on both rate enhancement and product distribution (Dubbels *et al.* 2007), however improvements in culturing *T. butanivorans* has yielded BMOH whose most notable difference is the effect BMOB has on the rate of catalysis (Fig. 2.3a). At a ratio of 3:1 BMOB:BMOH, the turnover rate of nitrobenzene increased 14-fold. Similar effects were seen for ethylene and butane oxidation. Although not as dramatic as the sMMO system, the stimulation of activity by BMOB is more consistent with the role of regulatory components in other di-iron monooxygenases. Unlike MMOB, addition of BMOB had little effect on the product distribution of butane oxidation, generating approximately 80:20 and 85:15 1-butanol:2-butanol without and with BMOB, respectively.

Because substrate turnover by sMMO without MMOB is slow, elucidating its effects on MMOH is difficult and continues to be an active area of research (Mitic *et al.* 2008). However, because the BMOH/BMOR complex oxidizes substrates at rates ~40 nmoles min^{-1} mg^{-1} without BMOB, we were able to measure the influence of the latter on the inhibition of alcohols during steady-state turnover. Only alcohols $C_3 – C_6$ were examined because this was the chain length range where breaks in binding affinities were observed. All alcohols bound with less affinity to BMOH when BMOB was not present (Table 2.2 and Fig. 2.3b). As a consequence, the break in primary alcohol binding affinity was observed between C_4 and C_5 rather than C_5 and C_6. Secondary alcohols $C_3 – C_6$ displayed a much less pronounced break between C_4 and C_5. 2-Methyl-2-butanol could not effectively inhibit sBMO without BMOB

present. Lastly, 1-hexanol displayed weak uncompetitive binding characteristics (K_{iu}>400 μM) without BMOB, implying a possible shift away from active site binding. The data suggest that BMOB helps to open the active site of BMOH for more efficient substrate access. Similarly, the dissociation of BMOB after substrate turnover may facilitate the release of products. Similar conclusions have been reached in detailed mutational and kinetic analyses in the sMMO system where MMOB plays a role in regulating substrate access and product release (Wallar and Lipscomb 2001; Zheng and Lipscomb 2006). Whether BMOB enhances the activation of oxygen as MMOB does for MMOH remains to be determined.

Special case of methane oxidation. Methanol is not an efficient inhibitor of BMOH even though methanol accumulation during methane oxidation ceases once low micromolar concentrations have been reached both *in vivo* (Halsey *et al.* 2006) and *in vitro*. Even though all the physiologically relevant alcohols were more effective inhibitors than methanol, no stoppage in product accumulation was observed during oxidation of alkanes $\geq C_2$, presumably due to effective competition between substrate and product. Moreover, *in vitro* characterization of the 1-butanol dehydrogenases BDH and BOH from *T. butanivorans* show that the K_m for 1-butanol is well below the K_{ic} for sBMO, which further emphasizes that sBMO inhibition by alcohols is unlikely to play an important physiological role in the metabolism of longer chain alkanes (Vangnai and Arp 2001; Vangnai *et al.* 2002). As such, methane oxidation clearly represents a unique case from other alkanes.

One possibility for the apparent steady level of methanol during methane oxidation is that it is the result of equilibrium between methane conversion to methanol and methanol conversion to formaldehyde, in which case methane and methanol consumption would be kinetically indistinguishable. Methanol has also been reported as a substrate for sMMO, but the low micromolar K_m for methane and approximately 1 mM K_m for methanol made methanol turnover by sMMO physiologically irrelevant (Colby *et al.* 1977). Kinetic analysis of methane and methanol as substrates of sBMO showed they have nearly identical K_m's and k_{cat}'s

(Table 2.3). As a result, reactions that initially contain only methane generate formaldehyde once methanol begins to accumulate. As the methanol concentration increases, the rate of formaldehyde formation increases until it equals the rate of methanol formation and a steady state level of methanol is reached (Fig. 2.4). This indiscrimination of C_1 substrates by BMOH is partially alleviated by an alteration in residue 113 from Gly to the corresponding MMOH amino acid, Asn. Kinetic characterization of the G113N variant of BMOH (Table 3) showed that this alteration makes sBMO more specific for methane than methanol, primarily by lowering the K_m for methane 3.3 fold. This MMOH conserved Asn residue appears to be critical in maintaining sMMO's specificity for methane over methanol, thereby eliminating effective substrate competition between the two.

Discussion

The first step in metabolizing an alkane requires the input of energy to cleave the highly stable C-H bond to form an alcohol, which can then be metabolized further in order to provide both the energy and carbon needs for the cell. Attempts to purify alkane monooxygenases from a variety of organisms have proven difficult (Shennan 2006), limiting in-depth kinetic and structural characterizations. In this study, we have characterized a soluble butane monooxygenase complex, which is the closest relative to the well-studied soluble methane monooxygenase enzymes that has been purified to homogeneity with high activity.

Substrate specificity. With the exception of sMMO, substrate specificity characterizations of alkane monooxygenases are mostly limited to *in vivo* analyses and only compare relative rates of substrate turnover using saturating concentrations of substrate, thereby neglecting differences in K_m's. Our analysis of sBMO shows that even though methane has the highest turnover rate of alkanes tested, it clearly is a poor substrate due to the high K_m relative to alkanes C_2-C_5. As such, sBMO is very effective at discriminating between methane and longer alkane substrates. While sMMO is more specific for methane over alkanes C_2-C_5, the apparent discrimination

for its physiologic substrate over longer alkanes is much less striking than sBMO. In sMMO, the methane V_{max}/K_m ratio reported for *M. capsulatus* sMMO was only 13 fold higher than that of ethane, and only 7 fold higher than that of propane (Green and Dalton 1986). For sBMO these same values for ethane and propane are 290 and 580 fold higher than methane, respectively. While it is difficult to compare substrate specificities with other alkane monooxygenases due to limited *in vitro* characterization, *in vivo* studies with soluble di-iron containing propane monooxygenases from *Gordonia* sp. TY-5, *Mycobacterium* sp. TY-6 and *Pseudonocardia* sp. TY-7 have suggested they are similar to sBMO in that they are specific for short chain alkanes C_2-C_6 but poorly suited for methane oxidation (Kotani *et al.* 2003; Kotani *et al.* 2006).

While particulate methane monooxygenase (pMMO) is unusual in that it has a narrow alkane substrate range that includes only linear C_1-C_5 alkanes (Elliott *et al.* 1997), wide substrate ranges are not uncommon in membrane bound alkane monooxygenases. The alkB hydroxylase from *Pseudomonas putida* GPo1 is capable of oxidizing large ($>C_{12}$) linear alkanes and substituted cyclic alkanes at comparable rates (van Beilen *et al.* 1994), while also oxidizing the smaller growth substrates propane and butane with high affinity (K_s = 66 and 13 µM, respectively) (Johnson and Hyman 2006). The alkane hydroxylase from propane utilizing *Mycobacterium vaccae* JOB5 was also recently characterized as an alkB hydroxylase (Lopes Ferreira *et al.* 2007), and has a low reported K_s for propane of 3.3 - 4.4 µM but high K_s's for branched substrates *tert*-butyl alcohol and methyl *tert*-butyl ether (1.36 and 1.18 mM, respectively) (Smith *et al.* 2003). The soluble heme-containing CYP153A6 alkane hydroxylase from *Mycobacterium* sp. HXN-1500 was reported to have low K_d values (20 nM) for growth substrates C_9-C_{11}, however K_d's for cyclic hydrocarbons were nearly 200 fold higher (Funhoff *et al.* 2006). Despite the large substrate ranges for these alkane monooxygenases, these data, together with our kinetic analyses of sBMO, emphasize the importance for these alkane monooxygenases to maintain high

specificities for physiologic substrates in order to out compete oxidation of molecules that would not provide the carbon and energy needs to sustain cell growth.

Studies have characterized the different iron-oxygen intermediates generated by the hydrocarbon oxidizers Toluene/*o*-xylene monooxygenase (ToMO) and sMMO. While sMMO generates a diamond core bis-μ-oxo-$(Fe^{IV})_2$ intermediate (Shu *et al.* 1997), ToMO generates a weaker oxidizing peroxo-bridged-$(Fe^{III})_2$ intermediate (Murray *et al.* 2007), suggesting that this may be a means of substrate selection. This type of selection mechanism, whereby sBMO generates a weaker oxidizing intermediate in order to take advantage of the weaker C-H bonds of longer chain alkanes over methane, is unlikely given that it oxidizes methane at faster rates than longer chain alkanes and that it hydroxylates the primary carbon. Therefore, the basis for substrate discrimination is more likely to be based on structural differences in the binding pockets than to be based on different chemical mechanisms. Unfortunately, these structural mechanisms are still poorly understood and are complicated by the observation that naphthalene and methane are both substrates of sMMO and sBMO despite their fully buried, identical active sites (Rosenzweig *et al.* 1997; Halsey *et al.* 2006). However, we have identified one particular residue in the α-subunit of BMOH, Gly-113, which contributes toward defining these specificities. As such, further coordinated mutational analyses of these homologs will certainly provide additional insight into the mechanism of substrate selection.

Component B. Several roles for the small regulatory component have been observed for soluble di-iron containing multicomponent monooxygenases. Recent studies demonstrated that in order to enhance the activity of MMOH, MMOB must induce both (i) a geometric rearrangement of a single Fe atom (Fe2) to enhance its reactivity with oxygen, and (ii) cause a more global conformational change within the active site to allow for efficient O_2 access (Mitic *et al.* 2008; Schwartz *et al.* 2008). The data shown here demonstrate that BMOB also has a large effect on substrate turnover rates (14-fold increase in activity), although this effect is modest compared to sMMO. While our data continue to emphasize the importance of these regulatory

components in di-iron monooxygenases for maintaining proper substrate selection and efficient product release, one significant difference remains between the sBMO system and both sMMO and other toluene oxidizers: the regulatory component in sBMO does not significantly alter the product regioselectivity for either alkanes or nitrobenzene. The change in regioselectivity caused by MMOB in sMMO was considered a consequence of the requirement that a global conformation change within the active site must occur for proper O_2 access (Mitic *et al.* 2008). While it appears that a conformational change within the active site of BMOH must be induced by BMOB, the mechanism must be different from that of sMMO where product distribution does change substantially. As such, sBMO provides a unique system in which to uncover details about the nature of these active site conformation changes necessary to enhance O_2 binding and activation without altering the position of substrate hydroxylation.

The kinetic studies of purified sBMO demonstrate that it is highly specific for linear alkanes $\geq C_2$ while effectively filtering out methane oxidation even though sBMO and sMMO share identical residues in the active site. While several physiological and enzymatic implications can be made, it is clear that sBMO can be a useful tool in helping elucidate specific factors that influence substrate specificity and activity in bacterial alkane monooxygenases. Efforts to crystallize BMOH are currently underway in our laboratory to address such comparisons from a structural point of view.

Acknowledgements

We thank Dr. Michael Schimerlik for his helpful comments and suggestions during the course of this study, and Christine Lastovica and Lisa Roberton for culturing and harvesting of cells. We are grateful for the research support from the National Institutes of Health, grant no. 5RO1 GM56128-06.

Abbreviations

BMOB, butane monooxygenase regulatory component; BMOH, butane monooxygenase hydroxylase; BMOR, butane monooxygenase reductase; MMOB, methane monooxygenase regulatory component; MMOH, methane monooxygenase hydroxylase; MMOR, methane monooxygenase reductase; sBMO, soluble butane monooxygenase; sMMO, soluble methane monooxygenase.

Table 2.1. Kinetic parameters for substrates C_1-C_5 with wild-type sBMO

Substrate	K_m (μM)	k_{cat} (sec^{-1}) [#]				k_{cat}/K_m (μM^{-1} sec^{-1})	Measured Specificity[¶]
		Total	1 - OH	2 - OH	3 - OH		
Methane	1100 ± 140*	1.3 ± 0.1[§]	-	-	-	0.0012 ± 0.0002	N/D
Ethane	2.2 ± 0.1[†]	0.76 ± 0.05*	0.76 ± 0.05	-	-	0.35 ± 0.03	1 ± 0.06
Propane	0.94 ± 0.05[†]	0.65 ± 0.06*	0.55 ± 0.05	0.098 ± 0.01	-	0.69 ± 0.07	2.4 ± 0.2
Butane	0.24 ± 0.02[†]	0.60 ± 0.03*	0.48 ± 0.02	0.12 ± 0.01	-	2.5 ± 0.2	5.7 ± 0.2
Pentane	0.34 ± 0.03[†]	0.39 ± 0.03*	0.34 ± 0.02	0.039 ± 0.01	0.008 ± 0.001	1.2 ± 0.1	4.4 ± 0.1

*Direct measurement
[†]Calculated from competition with nitrobenzene
[§]Extrapolated from Michaelis plot
[#] 1-OH column represents the rate of hydroxylation on the primary carbon, 2-OH for the secondary carbon at position 2, 3 – OH for the secondary carbon at position 3 and total represents the sum of all pr oducts.
[¶] Direct measurement by competition using an equal concentration of all alkanes in a single reaction, relative to ethane.
N/D: Not detected
Errors represent standard deviations from 3 replicates.

Table 2.2. Alcohol inhibition constants for wild-type BMOH

Inhibitor	Alcohol position	With BMOB		Without BMOB	
		K_{ic} (μM)	K_{iu} (μM)	K_{ic} (μM)	K_{iu} (μM)
Methanol	1	1250 ± 60	N/A[§]	-	-
Ethanol	1	150 ± 21	N/A	-	-
Propanol	1	69 ± 11	N/A	78 ± 5	N/A
	2	349 ± 25	N/A	550 ± 35	N/A
Butanol	1	23 ± 3	N/A	29 ± 5	N/A
	2	69 ± 9	N/A	475 ± 27	N/A
Pentanol	1	8.8 ± 1.0	N/A	49 ± 3	N/A
	2	180 ± 20	N/A	549 ± 47	N/A
Hexanol	1	13 ± 1.4	N/A	89 ± 14	414 ± 61
	2	580 ± 50	N/A	1290 ± 75	N/A
Heptanol	1	127 ± 14	102 ± 7	-	-
	2	1830 ± 90	N/A	-	-
Octanol	1	86 ± 16	139 ± 24	-	-
	2	521 ± 65	1200 ± 175	-	-
Nonanol	1	28 ± 8	29 ± 4	-	-
	2	142 ± 16	233 ± 21	-	-
2-methyl-1-butanol	1	515 ± 37	N/A	-	-
2-methyl-2-butanol	2	1450 ± 80	N/A	N/A	N/A

[§] N/A: Not observed. Inhibition constants greater than 10 mM were not considered. Errors represent standard deviations from 3 replicates.

Table 2.3: Kinetic parameters for wild-type and G113N BMOH oxidation of methane and methanol

Substrate	K_m (μM)		k_{cat} (sec^{-1})		k_{cat}/K_m (10^{-4} μM^{-1} sec^{-1})	
	WT	G113N	WT	G113N	WT	G113N
Methane	1100 ± 140	340 ± 20	1.3 ± 0.1	0.13 ± 0.01	11.8	3.8
Methanol	1250 ± 60	750 ± 40	1.6 ± 0.1	0.21 ± 0.02	12.8	2.8

Errors represent standard deviations from 3 replicates.

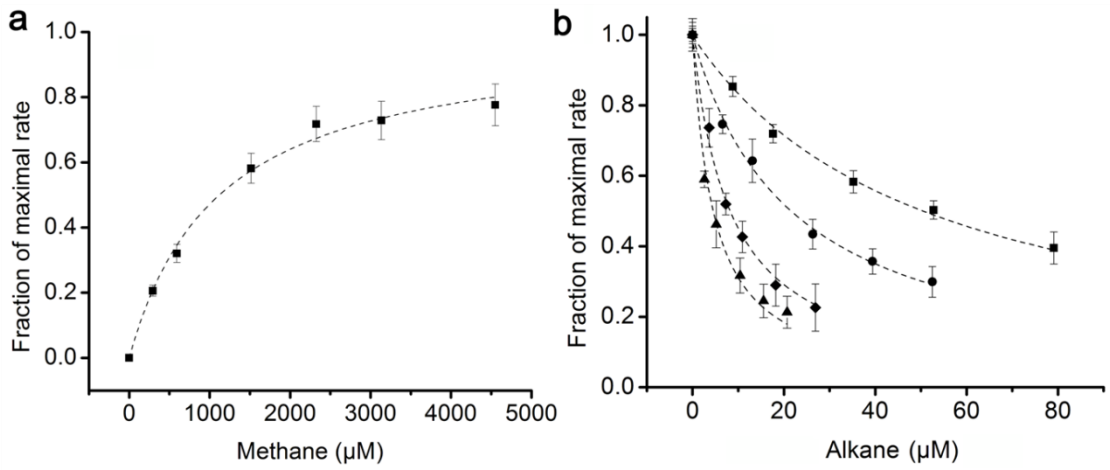


Figure 2.1. Determination of K_m's for alkanes C_1-C_5. (a) Dependence of the rate of methanol formation on the aqueous concentration of methane. Solid line is the best fit line according to eq. 2.1. Fractional rate of 1.0 is equal to 590 nmoles min^{-1} mg^{-1}. (b) Rates of nitrobenzene oxidation with ethane (■), propane (●), butane (▲) and pentane (♦) present as a competing substrate. Best fit lines are modeled to the data according to eq. 2.2. Error bars represent standard deviations from 3 replicates. Fractional rate of 1.0 is equal to 450 nmoles min^{-1} mg^{-1}.

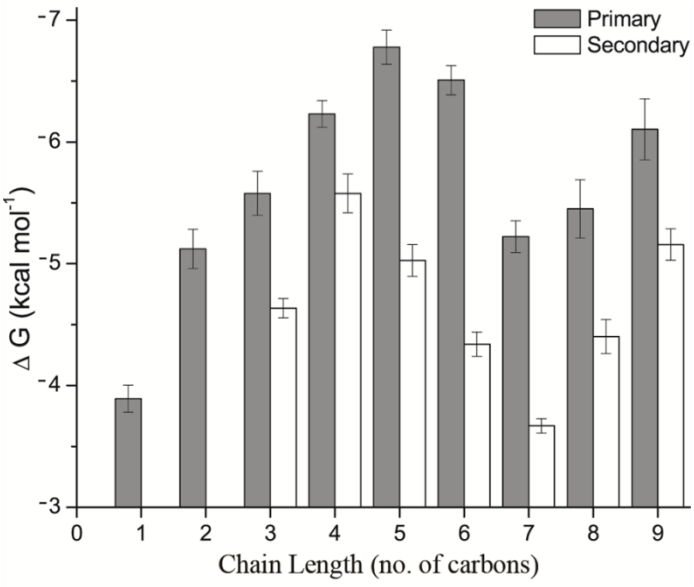


Figure 2.2. Alcohol binding energies to BMOH. Binding energies of primary alcohols (shaded bars) and secondary alcohols with the hydroxyl group at the second carbon position (white bars) were calculated from the observed competitive inhibition constants using the equation $\Delta G = -RT \text{ Ln } (K_{ic}^{-1})$ where T is the temperature in Kelvin (298 K) and R is the universal gas constant.

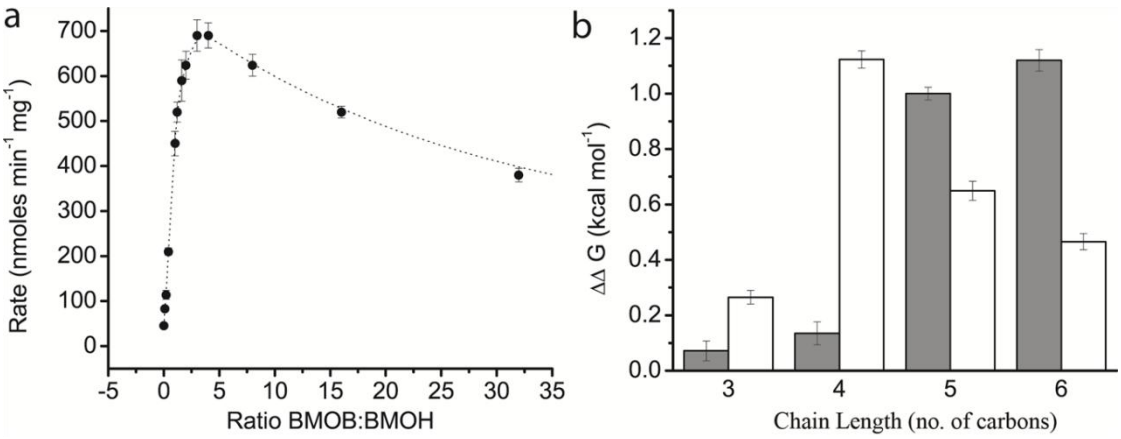


Figure 2.3. Influence of BMOB on the sBMO complex. (a) Titration of BMOB into a fixed concentration of BMOH during steady state nitrobenzene oxidation. (b) The energetic difference (in kcal mol^{-1}) between alcohol binding to the BMOH:BMOB complex and to BMOH alone for primary alcohols (shaded bars) and for secondary alcohols with the hydroxyl group at second carbon position (white bars). Positive values indicate weaker binding without BMOB. Error bars represent standard deviations from 3 replicates.

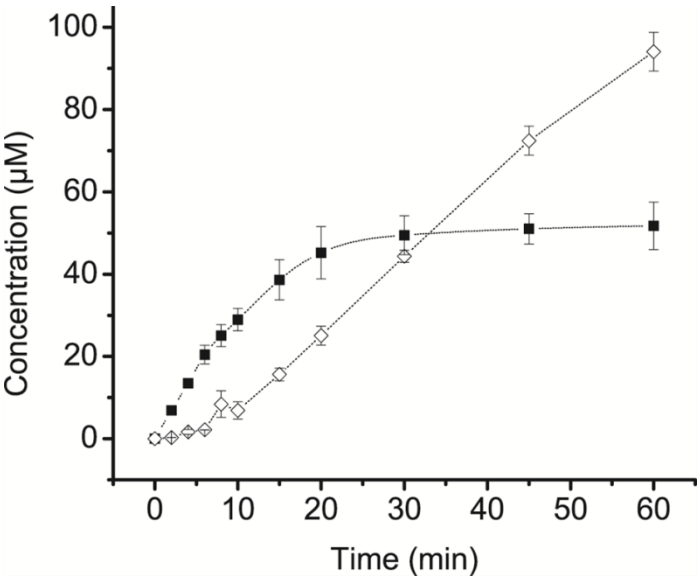


Figure 2.4. Oxidation of methane by sBMO. Concentrations of both methanol (■) and formaldehyde (◊) were measured in a reaction initially containing only methane as a substrate. Error bars represent standard deviations from 3 replicates.

Chapter 3

Growth of a non-methanotroph on natural gas: ignoring the obvious to focus on the obscure

Richard B. Cooley, Peter J. Bottomley, and Daniel J. Arp

Published in *Environmental Microbiology Reports* (2009), 1(5), 408-413.

Abstract

Methanotrophs are well known for their ability to grow on methane in natural gas environments; however these environments also contain low concentrations of longer chain length gaseous alkanes. This mixture of alkanes poses a problem for organisms that might otherwise grow on alkanes $\geq C_2$ because methane could inhibit oxidation of growth substrates and lead to an accumulation of toxic C_1 metabolites. Here, we have characterized the growth of a C_2-C_9 alkane utilizing bacterium, *Thauera butanivorans*, in conditions containing high concentrations of methane and small amounts (<3% of total alkane) of C_2-C_4. During such growth, methanol accumulates transiently before being consumed in an O_2-dependent process that leads to the formation of a proton gradient and subsequent ATP generation. In contrast, formaldehyde-dependent O_2 consumption is insensitive to uncouplers and does not lead to significant ATP production. This efficient C_1 oxidation process that regains much of the energy loss inflicted by oxidizing methane, coupled with an alkane monooxygenase effective at limiting methane oxidation, allows *T. butanivorans* to grow uninhibited in natural gas environments. Although longer chain length gaseous alkane utilizing organisms have been previously identified to grow in natural gas seepages, the data presented here represent the first detailed characterization of the physiologic effects associated with inadvertent methane oxidation by a non-methanotroph, and suggest the presence of a well evolved series of biochemical processes that allow them to grow in natural gas deposits without the need for developing the unique metabolic machinery characteristic of methanotrophs.

Introduction

Thauera butanivorans (ATCC 43655), formerly named *'Pseudomonas butanovora'* (Dubbels *et al.* 2009), is a gram-negative β-proteobacterium capable of growth on linear alkanes C_2-C_9, but not methane (Takahashi *et al.* 1980). In contrast, methanotrophic bacteria are capable of growth on methane, but none have been documented to grow on longer chain length alkanes. This sharp distinction between C_1 and $\geq C_2$ alkane utilizing organisms has led to a wealth of literature that characterizes their physiology under their respective growth substrates. Although this makes sense when considering CH_4 as a terminal product of anaerobic carbon processing, natural gas emitted from sources such as vents or seeps from petroleum deposits consist not only of methane (80-99%), but also low concentrations of longer chain length gaseous alkanes such as ethane, propane and butane (Hunt 1979; Hobson and Tiratsoo 1981). This geochemical mixture of alkanes clearly bridges the distinct C_1/C_2 biological boundary, yet there is little characterization of longer chain length alkane oxidizers growing in an environment containing significant quantities of non-growth alkanes (e.g. methane) even when their alkane monooxygenase is known to bridge such barriers.

The ability of $\geq C_2$ alkane utilizing bacteria to grow under high concentrations of methane and low concentrations of longer chain alkanes poses several metabolic and physiologic difficulties due to the inherent toxicity of methane metabolites (methanol, formaldehyde and formic acid). To grow in such conditions, these alkane utilizers would require the expression of an alkane monooxygenase capable of excluding methane oxidation even at low concentrations of physiologic substrates, the ability to continually and efficiently detoxify the metabolites of methane oxidation, or both. Nevertheless, growth on small, but continuous sources of C_2-C_4 alkanes would provide a high energy source of carbon without the need to adopt the genetic machinery of methanotrophs necessary to assimilate the toxic C_1 products of methane oxidation.

Such ecological niches have been industrially exploited, most notably for geomicrobiological oil prospecting efforts whereby oil companies would collect soil samples in areas suspect of natural gas deposits and measure C_2-C_4 alkane uptake by these soil samples (Shennan 2006). Because C_2-C_4 alkanes are created through geochemical rather than biological processes, the presence of such alkane utilizing organisms in the soil would identify seepages of exploitable natural gas beneath the surface. Unfortunately, little has been published on these commercially protected prospecting techniques, and their contribution toward oil exploration has been minimal, presumably due to cost and experimental limitations (Shennan 2006). The ability to grow in the presence of high concentrations of methane may not be a universal feature of alkane utilizers growing on C_2 and longer alkanes. Therefore, characterization of a bacterium capable of uninhibited and rapid growth on alkanes larger than methane while in the presence of high concentrations of methane would be of ecological and industrial interest.

The alkane monooxygenase expressed by *T. butanivorans* is a soluble, di-iron containing monooxygenase complex closely related to soluble methane monooxygenase (sMMO) (Sluis *et al.* 2002). This complex was isolated with high activity (Dubbels *et al.* 2007) and substrate specificities for alkanes C_1-C_5 have been recently reported (Cooley *et al.* 2009). Despite high similarity to sMMO, sBMO was 300-fold more specific for ethane and 1800-fold more specific for butane than methane even though methane turnover was fastest. This incredible discrimination by sBMO, which falls into a class of enzymes known to have over 100 different substrates (Borodina *et al.* 2007), led us to hypothesize that *T. butanivorans* might be capable of uninhibited growth on C_2-C_4 alkanes even under high concentrations of methane. The metabolic fate and the energetics of C_1 oxidation were also studied to understand the consequences of unintended methane oxidation. The results indicate that even though sBMO is very efficient at excluding methane oxidation in the presence of trace amounts of alkanes C_2-C_4, *T. butanivorans* is also highly effective at detoxifying C_1 metabolites with little energy loss.

Results and Discussion

Effects of methane on cell growth. To determine the effect of methane on the growth of *Thauera butanivorans*, a series of cultures were prepared by injecting a mixture of alkanes designed to mimic typical natural gas compositions (96.5% methane, 2% ethane, 1% propane and 0.5% butane). Cultures containing the same quantities of ethane, propane and butane but without methane grew at the same rate (doubling time ~6 h) and to the same final cell density as cultures containing methane (Fig. 3.1). This growth rate is not substantially different from that of cells grown with only butane. Methanol was detected in the medium of cultures exposed to methane and accumulated transiently during cell growth before being completely consumed. Formaldehyde was not detected in these cultures. However, formaldehyde was detected in small quantities (15 ± 5 µM) in cultures grown in the presence of 10 mM methanol (which also did not affect growth) and without methane (data not shown). Lastly, formate accumulated up to concentrations of 120 µM in cultures grown in the presence of methane. Cultures grown in the presence of 10 mM methanol and no methane accumulated 10 mM formate in the medium, indicating a stoichiometric conversion of methanol to formate and no further metabolism of formate by *T. butanivorans*. After this transformation of methanol to formate, the pH of the medium dropped by 0.8 units, suggesting formic acid rather than formate is exported to the media (cultures that were not grown in the presence of methanol did not change pH significantly). Based on the amount of formic acid produced in the presence of a mixture of methane and C_2-C_4 alkanes, it was calculated that methane consumption accounted for approximately 1-2% of the alkane flux through sBMO. This amount is not surprising given the relative substrate specificities of sBMO for alkanes C_1-C_4 (Cooley *et al.* 2009). Growth stopped after 30 h due to complete consumption of alkanes C_2-C_4 rather than being due to toxic side effects of C_1 metabolites or O_2 limitation. It is also worth noting that SDS-PAGE protein gels of whole cells grown

on alkanes in the presence of methane showed no identifiable difference in protein expression from cells grown without methane (data not shown).

The data indicate that *T. butanivorans* is capable of utilizing small amounts of alkane growth substrates even in the presence of a large excess of methane. This ability is largely due to the highly efficient filter sBMO provides for larger alkane oxidation over methane despite its high sequence similarity to sMMO. Additionally, high concentrations of methanol (10 mM) did not affect growth despite its stoichiometric conversion to formic acid during growth on longer chain alkanes. *T. butanivorans* therefore provides two mechanisms to minimize the toxicities of C_1 compounds: (i) an optimized alkane oxidizing enzyme capable of filtering out methane oxidation and (ii) a highly efficient process of oxidizing C_1 metabolites. The presence of methane, methanol, formaldehyde or formic acid had no effect on either sBMO activity or the expression of the sBMO operon as determined by the ethylene oxide activity assay (Hamamura *et al.* 1999) and *lac-z* reporter assay (Sayavedra-Soto *et al.* 2005), respectively. It appears, therefore, that *T. butanivorans* is well suited to grow in an environment containing large amounts of methane in order to pick out the trace amounts of longer chain alkanes, which, to the best of our knowledge, is the first characterization detailing how a non-methanotroph grows uninhibited in the presence of high quantities of methane and its metabolites.

Metabolism of methane oxidation products. Methanotrophs have developed efficient means to oxidize methane metabolites through energy yielding processes (Hanson and Hanson 1996). The accumulation and subsequent consumption of methanol by *T. butanivorans* led us to investigate whether the oxidation of methanol to formic acid via formaldehyde was energy yielding. Upon addition of 10 mM methanol to cells that had sBMO inactivated with acetylene (to prevent methanol oxidation and O_2 consumption by sBMO) and their growth substrate removed, the rate of O_2 consumption increased approximately 4-fold indicating that O_2 consumption is coupled to methanol oxidation. Respiration inhibitors Antimycin A, Oligomycin, *n*-propyl gallate and salicylhydroxamic acid (SHAM) were used to determine whether

the pathway of electron flow occurred through an energy coupled cytochrome oxidase pathway or through an uncoupled alternative oxidase pathway. Cyanide was not used as an inhibitor because it is known to inhibit PQQ-dependent alcohol dehydrogenases (Harris and Davidson 1993). Methanol oxidation was sensitive to inhibitors of both pathways (Table 3.1), as previously observed for 1-butanol-dependent O_2 uptake (Vangnai *et al.* 2002). Methanol-dependent alcohol dehydrogenase activity corresponded via native PAGE to the previously identified PQQ-dependent 1-butanol dehydrogenases BOH and BDH (Vangnai and Arp 2001) (Fig. 3.2). No NAD^+- or $NADP^+$-dependent methanol dehydrogenase activity was identified by native PAGE, nor was any methanol consumption detected in resuspended membranes. These observations were surprising given that the rate of methanol oxidation by BOH and BDH *in vitro* was only 2% that of 1-butanol activity (Vangnai and Arp 2001; Vangnai *et al.* 2002), while *in vivo* it was about 50% that of 1-butanol. The presence of the uncoupler carbonyl cyanide *m*-chloro phenyl hydrazone (CCCP, 50 µM) led to a 2-fold increase in the rate of methanol-dependent O_2 uptake indicating that the oxidation of methanol is coupled to the generation of a proton gradient. To determine if methanol oxidation was coupled to ATP production, cellular ATP levels were measured after mixing anaerobically incubated cells in an O_2 saturated buffer containing 10 mM methanol. The presence of methanol led to significantly increased cellular ATP concentrations compared to that of no substrate (Fig. 3.3). Interestingly, the relative rates of methanol- and butanol-dependent O_2 uptake correspond with the relative rates of methanol- and 1-butanol-dependent ATP production, which indicate the efficiencies of ATP formation are likely similar for the two substrates. This hypothesis is also supported by our inability to identify an enzyme specific for methanol oxidation. This scheme of methanol oxidation by *T. butanivorans* whereby methanol is oxidized via PQQ-dependent alcohol dehydrogenases and ATP is generated as a result of proton translocation across the periplasmic membrane is similar to that of C_1 utilizing organisms (Anthony 1993). However, methanol can also be oxidized to formaldehyde by sBMO in a reductant consuming process when alkane

growth substrates have been fully consumed (Cooley *et al.* 2009), at which point methanol can be consumed by both BOH/BDH and sBMO. To determine the flux of methanol consumed by these two pathways, cells with active sBMO and with acetylene-inactivated sBMO were incubated with either 10 mM or 50 µM methanol. At both concentrations, methanol was consumed by sBMO inactivated cells at 2/3 the rate of cells with active sBMO, suggesting that the majority of the methanol flux is sent through a reductant producing, energy-coupled manner rather than a reductant consuming process (data not shown). These data indicate that the majority of methanol, whether during growth or stationary phase, is oxidized to formaldehyde via BOH and BDH in an energy coupled mechanism, leading to increased intracellular ATP levels.

Methanotrophs and methylotrophs utilize formaldehyde for both assimilation into biomass and energy production. Because *T. butanivorans* cannot assimilate methane metabolites into biomass, the consumption of formaldehyde would be restricted only to the latter pathway. No formaldehyde was detected in cultures of *T. butanivorans* grown in the presence of methane; however, when grown in the presence of 10 mM methanol, low quantities (15 ± 5 µM) were observed. This result would indicate that formaldehyde consumption occurs at a rate nearly as fast as its production even in the presence of growth substrates. Indeed, the addition of 10 mM formaldehyde (prepared from paraformadehyde) to washed cell suspensions increased O_2 uptake by 8 fold (a rate twice as fast as methanol-dependent O_2 uptake). Formaldehyde-dependent O_2 uptake was significantly less sensitive to inhibition by Antimycin A and Oligomycin than methanol-dependent O_2 uptake (Table 3.1). Additionally, the uncoupler CCCP failed to increase the rate of formaldehyde-dependent O_2 uptake. Butyraldehyde-dependent O_2 uptake, however, was sensitive to CCCP, which indicates the oxidation of these two aldehydes likely occurs via different mechanisms. The presence of 10 mM formaldehyde led only to a slight increase in cellular ATP compared to that of no substrate (Fig. 3.3) even though its rate of oxidation is faster than that of methanol, and only slightly slower than that of 1-

butanol. These data therefore suggest that the majority of formaldehyde consumption is likely linked to an alternative oxidase rather than an energy coupled pathway. The slight increase in cellular ATP energy may be the result of either BOH/BDH oxidation, which are known to oxidize aldehydes (Vangnai and Arp 2001; Vangnai *et al.* 2002), or oxidation by NAD^+-dependent 1-butyraldehyde dehydrogenases (Vangnai *et al.* 2002). We therefore believe that the large part of formaldehyde oxidation appears to be purely a detoxification process and little cellular energy is obtained by its oxidation. This situation contrasts significantly with C_1 utilizing organisms where formaldehyde oxidation leads to the translocation of nearly twice as many protons across the membrane as methanol oxidation, and presumably greater ATP production (Dawson and Jones 1981). Additionally, the oxidation of formaldehyde by methanotrophs is a major source of reducing power for methane oxidation (Hanson and Hanson 1996).

The oxidation of formic acid to CO_2 is the last step in C_1 oxidation by methanotrophs and methylotrophs. Like formaldehyde, formic acid can be either assimilated into biomass or oxidized to CO_2 to provide reductant for C_1 utilizing organisms (Hanson and Hanson 1996; Crowther *et al.* 2008). *T. butanivorans* does not appear to assimilate formate or further oxidize formate to CO_2 (Fig. 3.1). Indeed, even concentrated cell suspensions in pH 7 medium containing 10 mM formate did not consume formate after 48 hrs of incubation. Together, these data indicate *T. butanivorans* stoichiometrically converts the methane to formic acid, which is then exported into the medium without further processing.

In summary, the oxidation of methanol to formaldehyde provides enough ATP-producing reductant to compensate for some of the sBMO-dependent consumption of NADH used to oxidize methane. However, formaldehyde oxidation only leads to a slight increase in cellular ATP despite the energetic potential this process has to generate cellular energy. The lack of coupled formaldehyde oxidation is not surprising since the production and subsequent dissipation of a proton gradient would limit the rate of formaldehyde consumption, thereby increasing intracellular levels of

free formaldehyde. Overall, the 3-step process of oxidizing methane to formic acid likely does not yield a net loss in cellular energy. It may be possible that the conversion of formate to formic acid in the cytoplasm prior to its export would help maintain a proton gradient across the cytoplasmic membrane, which would further compensate for the initial energy loss caused by inadvertent methane oxidation. Regardless, it appears that metabolism of methane to formic acid leads to little, if any, energy loss and does not affect the ability of *T. butanivorans* to grow.

Although the presence of non-methanotrophic alkane utilizers growing in a predominantly methane environment is not novel (as evidence by the aforementioned bioprospecting efforts), it is uncertain how prevalent *T. butanivorans*' physiologic characteristics of methane processing are among C_2-C_4 alkane utilizers. Within known isolates of the *Thauera* genus, and other closely related bacteria based on 16S sequence comparison, *T. butanivorans* is unique in its ability to metabolize alkane hydrocarbons (Song *et al.* 2001). Several genetically diverse bacteria that contain a *bmoX* gene similar to that of *T. butanivorans* have also been identified (Brzostowicz *et al.* 2005; Coleman *et al.* 2006), however neither physiologic studies nor their origin of isolation would immediately suggest compatibility with growth on C_2-C_4 alkanes in high concentrations of methane. Further characterization of alkane utilizers and their ability to grow in natural gas is needed to determine how widespread or specialized the characteristics of *T. butanivorans* are.

It is intriguing to consider how *T. butanivorans* inherited a similar alkane oxidizing enzyme as methanotrophs, but rather than evolving the ability to grow on an abundant supply of methane it developed both the ability to restrict methane oxidation and detoxify C_1 metabolites in order to grow on a limited supply of longer chain alkanes. It is likely there may be more alkane utilizers similar to *T. butanivorans* that did not acquire or develop the molecular components needed to assimilate methane metabolites in favor of oxidizing longer chain alkanes that provide greater energy output for each equivalent of invested energy. Given the likelihood that most well studied methanotrophs probably originated from environments where anaerobic

carbon decomposition produces primarily CH_4, one wonders if bacteria exist in deeper substrata that have evolved the ability to oxidize and grow on C_1-C_4 alkanes.

Table 3.1. Inhibition of methanol- and formaldehyde-dependent O_2 uptake by whole cells of *T. butanivorans* grown on methane, ethane, propane and butane. [a]

Inhibitor	Methanol		Formaldehyde	
	IC_{50} (μM)	Maximal Inhibition (%)	IC_{50} (μM)	Maximal Inhibition (%)
Antimycin A[b]	25 ± 4	80 ± 3	24 ± 3	42 ± 2
Oligomycin[b]	15 ± 6	67 ± 6	35 ± 5	31 ± 2
n-Propyl gallate [c]	160 ± 12	61 ± 4	80 ± 7	59 ± 4
SHAM [c]	140 ± 20	53 ± 4	35 ± 4	49 ± 3

[a] Respiration was measured in the presence of 10 mM substrate using a Clark style oxygen electrode with a 1.6 mL reaction volume. Prior to analysis, sBMO was inactivated with acetylene in all cells as previously described (Hamamura *et al.* 1999). Final cell concentration was 0.06 mg protein ml^{-1}.
[b] Inhibitors of respiration coupled to ATP production, prepared as previously described (Vangnai *et al.* 2002).
[c] Inhibitors of respiration through an alternative oxidase, prepared as previously described (Vangnai *et al.* 2002). SHAM: Salicylhydroxamic acid

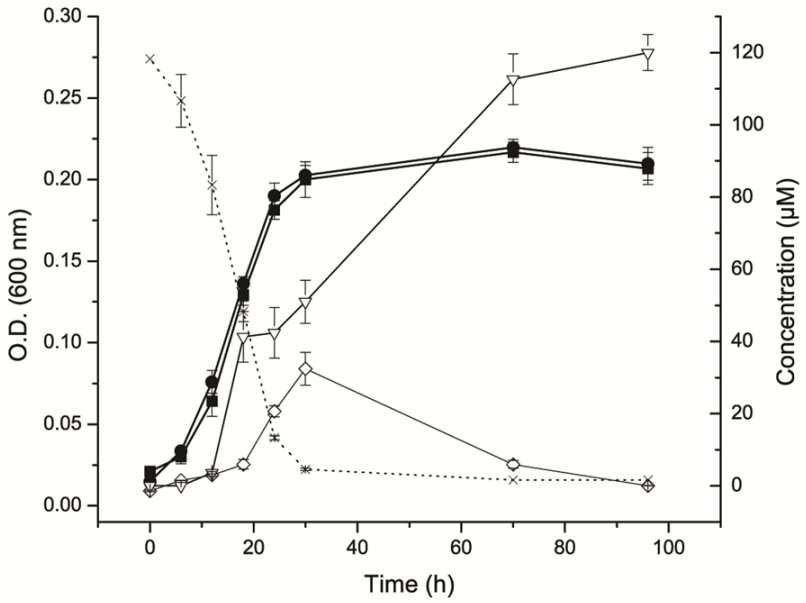


Figure 3.1. Growth of *T. butanivorans* under natural gas conditions. Cultures containing 100 ml of media, prepared as previously described (Cooley *et al.* 2009), were incubated in 750 ml sealed bottles that contained a mixture of gaseous alkanes added as an overpressure. For cultures containing methane (■, solid line), the alkane mixture consisted of 96.5% methane (2.4 mmoles), 2% ethane (0.05 mmoles), 1% propane (0.025 mmoles) and 0.5% butane (0.012 mmoles). Cultures were also grown with the same quantity of ethane, propane and butane, but no methane was added (●, solid line). Cell growth was measured by optical density at 600 nm. Methanol production (◊, solid line) was measured by gas chromatograph (GC) as previously described (Halsey *et al.* 2006). Formic acid (▽, solid line) was determined by incubating media samples with 5 mM NAD^+ and 0.35 U ml^{-1} formate dehydrogenase (Sigma) at 37 ^oC for 2 h and measuring NADH formation at 340 nm. Total C_2-C_4 alkane concentration (×, dashed line) is plotted as a fraction of the starting concentration, beginning at 1.0 and dropping to 0.0 by 30 h (y-axis not shown). Alkane concentrations were measured by GC equipped with a thermal conductivity detector. Error bars represent the standard deviation of 3 replicates.

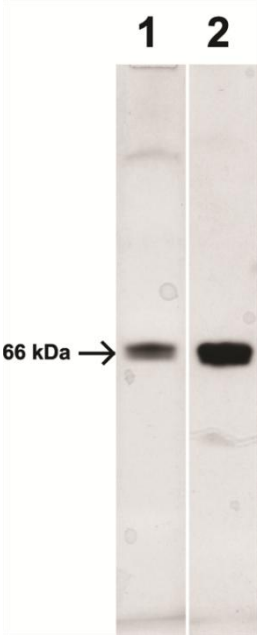


Figure 3.2. Alcohol dehydrogenase activity for methanol and 1-butanol in cell free extracts of *T. butanivorans*. The soluble fraction of cells grown on a mixture of methane, ethane, propane and butane were electrophoresed on native PAGE (20 µg protein) and stained for alcohol dehydrogenase activity with nitro-blue tetrazolium as previously described (Vangnai and Arp 2001). Lane 1 was incubated with 10 mM methanol for 20 min, while lane 2 was incubated with 2 mM 1-butanol for 5 min. The bands shown correspond to the 66 kDa BOH/BHD enzymes previously identified (Vangnai and Arp 2001).

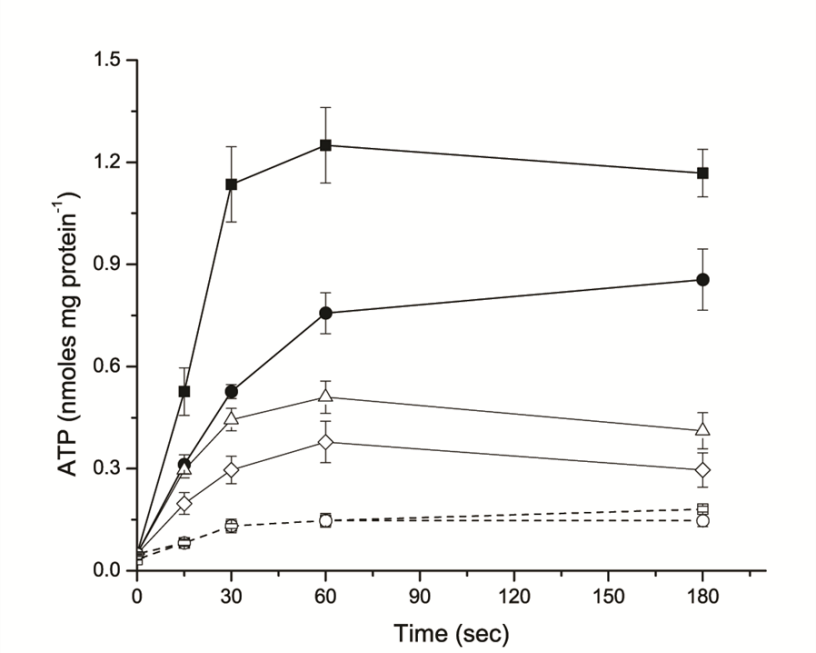


Figure 3.3. ATP formation by *T. butanivorans* in the presence of various electron donors. Cells grown with methane, ethane, propane and butane were washed 3 times and resuspended in 50 mM phosphate buffer pH 7.2 prior to purging with N_2 for 2 h. The anaerobic cell suspension was allowed to stir slowly for 24 hrs. Cells were then diluted (final concentration 1.7 mg protein mL^{-1}) in an equal volume of O_2-saturated buffer with either 2 mM 1-butanol (■, solid line), 10 mM methanol (●, solid line), 10 mM formaldehyde (Δ, solid line), no substrate (◊, solid line), 2 mM 1-butanol + 50 µM CCCP (○, dashed line) or 10 mM methanol + 50 µM CCCP (□, dashed line). Samples (0.7 mg protein) were quenched by adding an aliquot of cells to a concentrated solution of perchloric acid (final concentration 0.4 M) and EDTA (final concentration 20 mM). ATP concentrations were determined by HPLC as previously described (Manfredi *et al.* 2002). Error bars represent the standard deviation of 3 replicates.

Chapter 4

Evolutionary origin of a secondary structure: π-helices as cryptic but widespread insertional variations of α-helices enhancing protein functionality*

Richard B. Cooley, Daniel J. Arp and P. Andrew Karplus

Published in *Journal of Molecular Biology* (2010), 404(2), 232-246.

Abstract

Formally annotated π-helices are rare in protein structure but have been correlated with functional sites. Here, we analyze protein structures to show that π-helices are the same as structures known as α-bulges, α-aneurisms, π-bulges, and looping-outs and are evolutionarily derived by the insertion of a single residue into an α-helix. This newly discovered evolutionary origin explains both why π-helices are cryptic, being rarely annotated despite occurring in 15% of known proteins, and why they tend to be associated with function. An analysis of the π-helices in the diverse ferritin-like superfamily illustrates their tendency to be conserved in protein families, and identifies a putative π-helix-containing primordial precursor, a "missing link" intermediary form of the ribonucleotide reductase family, vestigial π-helices, and a novel function for π-helices we term peristaltic-like shifts. This new understanding of π-helices paves the way for this generally overlooked motif to become a noteworthy feature that will aid tracing the evolution of many protein families, guide investigations of protein and π-helix functionality, and contribute additional tools to the protein engineering toolkit.

Introduction

The diversification of protein function is an essential process that enables organisms from all kingdoms of life to thrive in an extraordinary range of environments. The goal of what has been called the functional synthesis of evolution is to gain a detailed understanding of how mutational changes in proteins give rise to novel functionality during this process (Dean and Thornton 2007; Bloom and Arnold 2009; Worth *et al.* 2009). However, how the first protein folds arose and concrete mechanisms by which protein secondary structure and protein folds evolve have been difficult to define.

While α-helices, a dominant secondary structure in proteins, are defined by main-chain hydrogen bonds (H-bonds) between residues four apart in sequence, π-helices are a protein secondary structure with main-chain H-bonds between residues five apart in sequence (referred to as π-type H-bonds). They were predicted in the 1950s (Low and Baybutt 1952), and although extended π-helices of regular conformation do not occur in proteins (Hollingsworth *et al.* 2009), short π-helical segments do. The empirical definition of a π-helix is the occurrence of at least two sequential π-type H-bonds (Fodje and Al-Karadaghi 2002). The literature for these short motifs is rather muddy, evolving from early work indicating that π-helices do not occur (Creighton 1993), to a report documenting π-helices as rare but with special functional roles (Weaver 2000), and continuing with the work of Fodje & Al-Karadaghi (Fodje and Al-Karadaghi 2002) who showed that π-helices are more common than suspected, occurring in "one of every 10 proteins." Still, π-helices tend to be overlooked in protein structures. π-helical conformations have been noted to occur in protein simulations (Lee *et al.* 2000; Mahadevan *et al.* 2001; Armen *et al.* 2003), although it is controversial whether these occurrences give informative insight into the true occurrence of such helices or reflect a shortcoming of the molecular mechanics force-fields (Feig *et al.* 2003).

The large majority of naturally occurring π-helices consist of seven-residue segments with the minimal two π-type H-bonds (Fodje and Al-Karadaghi 2002).

What has not been previously noted is that these π-helices have the same conformation and H-bonding pattern seen in protein engineering studies when a single amino acid is inserted into an existing α-helix to create what was called an α-aneurism(Keefe *et al.* 1993) or looping out (Heinz *et al.* 1993) (Fig. 4.1). Also matching this conformation are α-helical distortions seen in some natural proteins that have been called π-bulges (Cartailler and Luecke 2004) or α-bulges (Hardy *et al.* 2000). This striking similarity led us to hypothesize that natural π-helices arise during evolution via single residue insertions into α-helices. Here, through an analysis of known protein structures, we provide compelling evidence that naturally occurring π-helices are indeed evolutionarily related to α-helices. We further show that this often overlooked secondary structure is a highly informative marker with important evolutionary and functional implications, giving new insights into the origin, evolution and diversification of protein functionality, even for well characterized protein families.

Results and Discussion

The evolutionary relationship between α- and π-helices.

π-helices are associated with α-helices. To explore our hypothesis that naturally occurring π-helices arose from the insertion of a single amino acid into an α-helix, we first investigated the π-helices identified by Fodje & Al-Karadaghi (Fodje and Al-Karadaghi 2002) more closely. Reclassification of their list of π-helices (see Fig. 4.2 and *Material and Methods*) yielded 106 π-helical segments from 79 protein chains, with the longest ones having five consecutive π-type H-bonds (Table 4.1). Of these 106 π-helices, 102 were present in the midst of an α-helix, consistent with the insertion hypothesis. Interestingly, this observation completely explains why π-helices continue to be overlooked. When a π-helix is sandwiched in an α-helix, the first few and last few residues of the π-helix are also part of the bordering α-helices, and because automated secondary structure assignment algorithms such as DSSP (Kabsch and Sander 1983) and STRIDE (Frishman and Argos 1995) give precedence

to the α-helical assignment, these residues are designated as α-helical. Then, the remaining (if any) π-helix residues are too few to meet the minimal length of a π-helix*, and so are designated as "turns", leaving the π-helix completely unrecognized (e.g. Fig. 4.2). In the case of the 106 π-helices in this dataset, 105 are cryptic in this sense and only one is formally annotated by DSSP as a π-helix.

π-helices can be derived from α-helices by a one-residue insertion. For more direct evidence of an evolutionary relationship between α- and π-helices, we used the FSSP database (Holm *et al.* 2008) to survey all structurally known homologs of each π-helix-containing protein. Based on the FSSP survey, 88 of the 106 π-helices occurred in proteins with a structurally known homolog for which the equivalent segment is a pure α-helix with one less residue (Tables A1.1 and A1.2). All of the identified α-/π-helical homolog pairs were members of the same superfamily, had structural alignments with Z-scores ranging from 6.4 to 56.0 and had sequence identities ranging from 5 to 70%. For the pairs with high levels of sequence similarity, the sequences themselves provide unambiguous evidence for the insertion of a single residue.

Considering that the Protein Data Bank (PDB) is far from a comprehensive set of protein structures and that all 18 π-helices without known α-helical homologs are embedded within α-helices (Fig. A1.1), this remarkable 83% identification rate implies that the overwhelming majority of π-helices are indeed evolutionarily related to α-helices. It is noteworthy that the insertion of a single residue is able to account not for just π-helices with two H-bonds, but also those with three, four, and five H-bonds (Table 4.1, Fig. 4.3). We suggest the difference between these outcomes is how the helix adjusts its conformation to accommodate the extra amino acid (Fig. A1.2); this would, of course, depend on the local environment in which the insertion takes

* For 3_{10}-, α- and π-helices, DSSP and STRIDE generally do not assign the N- and C-capping residues as part of the helix. Thus, what we describe here as a seven residue π-helix is defined by DSSP to encompass only the central five residues.

place. Also noteworthy is that when multiple, overlapping π-helices are present within a long α-helix, a set of homologous proteins exist that document a potential stepwise pathway of accumulation of insertions (Table A1.2; Fig. 4.4).

Taken all together, these observations show that the insertion of a single residue into an α-helix is a sufficient explanation for generating the whole spectrum of π-helices observed in this dataset, and unifies π-helices, π-bulges, α-aneurisms and α-bulges as a single phenomenon. These structures encompass a breadth of conformations and H-bonding lengths and it is a matter of definition whether one considers them a perturbation of an α-helix or a π-helix. We propose that π-helix is the best designation for all these structures because they all result from a single type of event and "π-helix" is the only name that encompasses both the locally bulged structures and the more extended helical structures. In terms of how the α-helix is perturbed, the π-helices in this set demonstrate they sometimes minimally perturb the helix they enter just causing a local bulge or a bulge with a slight bend (e.g. Figs. 4.1a and 4.5c) or a more major kink and distortion (e.g. Fig. 4.5d). π-helices, as so defined, encompass a rather broad range of (ϕ,ψ)-angles (Fig. A1.2 and Fig. 4.2a from Fodje & Al-Karadaghi (Fodje and Al-Karadaghi 2002)) and this makes them more like 3_{10}-helices, which are much more conformationally diverse than α-helices (Barlow and Thornton 1988).

π-helices occur in ~15% of all proteins. The analysis of the reclassified list of π-helices from Fodje & Al-Karadaghi (Fodje and Al-Karadaghi 2002) conclusively supports the hypothesis for the relatedness of α- and π-helices and therefore it becomes of great interest to know the abundance of these motifs in nature. Using just our standard criteria and some broad (ϕ,ψ) restrictions (see *Materials and Methods*), a survey of the current PDB (Table 4.1) shows that of the 5620 diverse protein chains, 803 chains (~15%) contain a total of 1010 π-helices, with the longest being seven H-bonds in length; proportionately similar results were found with the use of more or less stringent hydrogen bonding criteria or the surveying of a larger subset of the PBD

(Table A1.3). Removing the (ϕ,ψ) restrictions in our search resulted in relatively few additional hits (~5% more), showing that aside from π-helices, very few turn structures satisfy the H-bonding criteria. The results show π-helices are widespread, occurring in one in six protein chains, yet infrequent– occurring only once per chain in 84% of cases (Table A1.4). The greatest number of π-helices within a single chain was eight, occurring in the bacterial homologue of a Na^+/Cl^--dependent neurotransmitter transporter (Yamashita *et al.* 2005) (PDB 2a65). All eight of these π-helices go unrecognized by DSSP annotation, yet two of them, which were identified in the original report as "helical disruptions" (Yamashita *et al.* 2005), are at the substrate binding site and intimately connected with function.

Although we have not inspected each individual π-helix resulting from this broader search, they too are strongly associated with α-helices. More than 95% of the π-helices have at least 3 residues within them that are designated as α-helical by DSSP and, consistent with this, they are nearly all "cryptic": compared with the 1010 π-helices we identify in this dataset, DSSP only formally annotates 44 π-helices (Table A1.3). Also, checking the six and seven H-bonded π-helices seen in the PDB survey shows that they too have homologs with a pure α-helix that differs only by a single residue (Fig. 4.3), showing that they too can arise via the same insertional phenomenon as π-helices with two, three, four and five H-bonds seen in the smaller dataset. An interesting detail consistent with previous work on π-helices (Fodje and Al-Karadaghi 2002) and π-bulges (Cartailler and Luecke 2004) is that a proline residue is commonly present where the π-helix transitions back to an α-helix as 36% of seven-residue π-helices have a proline immediately following them.

<u>The association of π-helices with protein function</u>.

The unification of π-helices and π-bulges as a single phenomenon strengthens their association with the evolution of protein function because both π-helices (Weaver 2000) and π-bulges (Cartailler and Luecke 2004) have also been shown to be

enriched at protein functional sites. Illustrating this point are three examples of enzymes from branches of well-studied, large superfamilies with an α-helix-derived π-helix critically involved in the enzyme's active site (Fig. 4.5). Because these π-helices (and most of those listed in Table A1.1) are present in discrete subgroups of larger protein superfamilies, we infer that they were evolutionarily derived from α-helix-containing ancestors (Fig. 4.5a). In each of these cases, the π-helix is conserved within the protein family described. The first example involves the phosphorylase branch of the ancient (Gibson *et al.* 2002) UDP-glucose-dependent glycosyl transferases (Fig. 4.5b). In this family, a previously unrecognized α- to π-helix conversion places a Trp residue at the pyridoxal phosphate binding site and its occurance in the superfamily correlates perfectly with the transition from UDP-glucose dependence to pyridoxal phosphate dependence. In mercuric ion reductases (Fig. 4.5c), an α- to π-helix conversion that is clearly due to the insertion of a single residue compared to more ancient members of the pyridine nucleotide linked disulfide reductases places a key catalytic Tyr residue into the mercury binding site (Rennex *et al.* 1993). In acetylcholine esterase (Fig. 4.5d), a previously unrecognized α- to π-helix conversion unique to this subgroup of the α/β hydrolase superfamily is associated with a bend in a helix that changes the shape of the active site pocket and allows a new chain segment to supply the glutamate of the Glu-His-Ser catalytic triad. In addition to these three examples, the previously described functionally important α-aneurisms and α-bulges (which can now all be called π-helices) from heat shock transcription factor (Hardy *et al.* 2000), halo- and bacteriorhodopsin (Luecke *et al.* 1999; Kolbe *et al.* 2000), and a subgroup of peroxidredoxins (Sarma *et al.* 2005) are also present and conserved within all members of a branch of broader protein families or superfamilies.

A rational for π-helix association with function. This novel explanation for the origin of π-helices provides a rationale for their association with functional sites in proteins. Because π-helix formation via the insertion of a residue has been observed to destabilize a protein by ~3-6 kcal/mol (Heinz *et al.* 1993; Keefe *et al.* 1993),

corresponding to a 100- to 10,000-fold decrease in the population of the folded state, π-helices would tend to be selected against unless they were associated with a functional advantage. This association could be direct, in that the π-helix forming mutation is directly adaptive, or it could be indirect in that the π-helix forming mutation is a non-adaptive change resulting from genetic drift in a sufficiently stable protein (Bloom and Arnold 2009) but which sets the stage for a subsequent adaptive mutation (Dean and Thornton 2007). We are aware of no other such motif or class of mutation that as a group is so consistently associated with function, making π-helices powerful novel markers for mapping the evolution of and identifying unique functionalities associated with particular branches of protein families. As such, π-helices (e.g. Tables A1.1 and A1.2, and the Supplemental PDB Survey for π-helices found on the online version of this manuscript at the *J. Mol. Biol.* website) provide fertile ground for evolutionary and functional analyses of many protein families.

Peristaltic-like shifts as a novel functionality of the π-helix. Our refined definition of π-helices also brings to light a unique structural versatility that can contribute to protein function. The hydroxylase component of soluble methane monooxygenase (MMOH), a member of the bacterial multicomponent monooxygenases (BMM), has been reported to have an active site π-helix that extends upon the binding of a product analog (Sazinsky and Lippard 2005). Closer inspection of this active site segment shows that actually no π-helical extension occurs. Instead what was thought to be one long π-helix consists of two overlapping π-helices (designated πB and πD), and during ligand binding one of them (πD) is shifted six residues in the C-terminal direction in a manner reminiscent of esophageal peristalsis (Fig. 4.6a). Such a peristaltic shift of π-helices has not been noted before. Interestingly, our inspection of the related toluene-4-monooxygenase hydroxylase structure (Bailey *et al.* 2008) shows that the binding of the regulatory subunit (common to all BMMs) causes an equivalent active site π-helical shift (Fig. 4.6b) that enlarges the buried active site cavity, presumably preparing it for substrate binding.

This π-helical shift provides a plausible mechanism for how BMMs are activated by their regulatory subunits (Cooley *et al.* 2009). Of further note, the active site π-helical shift in toluene-4-monooxygenase is coordinated with a shift in a neighboring π-helix (designated πE and conserved in all BMM enzymes) that appears to transmit the effect of regulatory subunit binding to the active site (Fig. 4.6b). This controlled pushing and pulling of π-helices in and around the buried active site of BMM enzymes provides an example of how a π-helix not directly in an active site can still specifically contribute to function.

The power of π-helices as evolutionary markers: the ferritin-like superfamily.

To illustrate the kinds of insights that can come from this awareness of π-helices, we explore here the structure-function relations and evolutionary history of the BMM family for which the best studied representative, MMOH, has a remarkable twelve π-helices within its homologous α- and β-subunits. The BMM enzymes belong to the ferritin-like superfamily, which is characterized by a conserved four-helix bundle core (referred to as helices A, B, C and D) with a carboxylate-bridged di-nuclear center usually coordinated by six protein side chains (Murzin *et al.* 1995). This superfamily includes the more ancient rubrerythrins, Δ^9-desaturases, ferritins and bacterioferritins, as well as the younger class I ribonucleotide reductase R2 subunits (RNR) and BMMs (Andrews 1998; Gomes *et al.* 2001; Leahy *et al.* 2003; Wiedenheft *et al.* 2005). Using a structure-based sequence alignment of proteins from each of the major families within this broad superfamily, we generated a maximum likelihood phylogenetic tree onto which we mapped a minimalist set of gains and losses of π-helices (Figs. 4.7a and 4.7b).

Origin of the ferritin-like superfamily. A first major insight derived by considering π-helices is a concrete proposal for the peptide-based origin of this superfamily. The root of the tree is placed near the rubrerythrins because internal symmetry in the sequence of a rubrerythrin-like protein, erythrin, from the primitive eukaryote *Cyanophora paradoxa* (31% sequence identity between homologous

helices) implies that the core four-helix bundle of the superfamily developed via the duplication and fusion of an ancient gene encoding a two-helix chain that formed a dimer (Andrews 1998) (Fig. 4.8). Here, we extend this inference by noting that ruberythrin's unique seventh iron ligating residue (a Glu in core helix C) is located on a π-helix (called πA_2), and furthermore, that the sequence of erythrin not only conserves this seventh ligating residue but also has an additional internal symmetry-related Glu residue inserted in core helix A (Fig. 4.8b). This previously unrecognized feature of erythrin strongly supports the conclusion that this original ancestor of the ferritin-like superfamily had a higher internal symmetry involving two π-helices and an unprecedented eight metal ligating residues. These two π-helices would have been generated from a single insertion into the primordial 2-helix protein (πA Fig. 4.8a) which upon gene duplication and fusion became helices πA_1 in helix A and πA_2 in helix C. An "insertion" at the level of such a peptide would not be distinct from a conformational adjustment that creates the π-helix during dimerization and iron binding. Given that the dominant pattern for characterized superfamily members is to have six ligating residues, with rubrerythrin being considered unusual due to its seventh ligating residue, this proposal for an ancestral di-iron center containing eight ligands is unanticipated and demonstrates the power that considering π-helices can have in evaluating evolutionary phenomena. The existence of erythrin as a modern day protein is remarkable and suggests it has experienced a slow rate of change making it the protein-equivalent of a living fossil, which is further supported by the observation that it is found only in organisms that themselves are a kind of living fossil (*C. paradoxa* and *Gloeobacter violaceus*). This concrete hypothesis for a novel metallocenter structure in erythrin not only defines a peptide-based origin of this protein fold, but also provides clues to the novel chemistry associated with the development of oxygenic photosynthesis on earth.

Correlation between π-helices and metallocenter geometry. A second major insight derived from considering π-helices has to do with the development of the distinct chemistry of the different families within this superfamily. With the root of

the tree placed near the erythrins, one can follow from there the formation/deletion of π-helices through the superfamily (Fig. 4.7a). This history shows mostly additions of π-helices, but in three cases – πA_1, πA_2 and πC – π-helical insertions are lost. Remarkably, every one of the π-helices in the core of the tree (πA - πD) are at the metallocenter where their addition or loss would directly influence metallocenter geometry and thus active site chemistry (Fig. 4.7b; Fig. 4.9). πA_2 contributes the seventh ligating residue characteristic to rubrerythrins and presumably erythrin but is not present in any other branches. πB from core-helix C contributes a ligating Glu residue (E193 in Fig. 4.9) which displays oxidation-state-dependent coordination in RNR Ia/b enzymes: in the oxidized (diferric) state this iron-Glu interaction is monoodentate, but in the reduced (diferrous) state the interaction is bidentate. However, in the more ancient enzymes that do not have πB, this Glu residue coordinates through bidentate interactions regardless of the metallocenter oxidation state. Insertion πC correlates with a residue change in the nearby ligand from Glu in more ancient enzymes to Asp in RNR Ia/b enzymes (red text in Fig. 4.9). Then, upon the subsequent insertion of πD and a shift in πC, this Asp ligand converts back to Glu in RNR Ic (E89) and BMM enzymes. πD also correlates with a further transition in the coordinating properties of the Glu ligand in core-helix C (E193 in RNR Ic) from oxidation-dependent in RNR Ia/b enzymes to universally monodentate regardless of the oxidation state in the BMM enzymes (blue text in Fig. 4.9). This monoodentate interaction is critical for BMM enzymes (Schwartz *et al.* 2008). Interestingly, however, the function of πB and πD have extended beyond just ligand coordination chemistry once the BMM family diverged from the RNR Ic family as both these π-helices in BMMs play roles in substrate binding as well, as shown in Fig. 4.6.

Evolutionary steps from RNR to BMMs. A particularly interesting case involves insertion πD whose appearance is correlated with a change in both metallocenter geometry and chemistry in the RNR family. Interestingly, surveying RNR Ic-like sequences while considering the πD helix led us to discover a group of sequences (designated Ia* in Fig. 4.7a) which has the πD insertion yet still possesses

the Tyr residue that is the defining characteristics of class Ia/b enzymes. This Ia* group represents an evolutionary link between RNR Ia and Ic enzymes, proving that the π-helical insertion preceded and thus could have played a role in the changes in active site chemistry of RNR Ic compared to the RNR Ia/b and BMM enzymes.

Due to the metallocenter similarity of BMMs and RNR class Ic, it has been hypothesized that the BMMs diverged from RNR class Ic (Hogbom *et al.* 2004; Andersson and Hogbom 2009). Here, we note that the presence of πD in both the BMMs and RNR class Ic reinforces this relationship and supports this more parsimonious path for the evolution of these enzymes (see dotted line in Fig. 4.7a), even though the branching pattern based purely on global sequence similarities suggests the independent generation of both πD and the accompanying active site geometry features common to BMMs and RNR Ic. The low bootstrap value of 53 for the BMM/RNR branching point indicates the relationships are sufficiently distant that the branching is not well defined, so alternate proposals may have validity. If the more parsimonious path is correct, it would imply that after the insertion of πD, the RNR Ic sequences experienced a slow rate of evolutionary change due to constraints to conserve RNR function, whereas the BMM branch experienced much more extensive sequence change associated with optimizing its new monooxygenase function. We suggest that such a conservation of an existing function also accounts for why the erythrins have diverged very little from the symmetric precursor of the superfamily while the other families — which developed novel functions — diverged much more over this same time period.

Vestigial π-helices in BMM β-subunits. Considering the roles of π-helices in BMM evolution, it is also interesting that this evolutionary history shows that the BMM β-subunits, even though they no longer bind metals, have retained π-helices πB and πD as vestigial features. This suggests that although the insertion of a residue to create a π-helix is energetically unfavorable, once a π-helix has been accepted and the protein has evolved to accommodate it, then its removal becomes unfavorable. This was experimentally documented for the π-helix (α-bulge) in heat shock factor, for

which deleting a residue did create an α-helix but significantly destabilized the protein (Hardy *et al.* 2000). Finally, we note that this tree provides concrete examples of two non-active site π-helices that play functional roles: πE in the BMM enzymes, whose function in regulating intersubunit interactions is described above, and the previously unrecognized πM in eukaryotic ferritins, which occurs at a three-fold subunit interface and plays a role in creating an iron deposition channel unique to those ferritins (Andrews 1998).

Outlook. The examples highlighted in this analysis provide compelling evidence that the overwhelming majority of π-helices are evolutionarily related to α-helices; the examples also illustrate the power that an awareness of this relationship brings to enhancing our understanding of protein structure, function and evolution, even for protein families such as glycogen phosphorylase and ribonucleotide reductase that have been extremely well studied. In the case study of the ferritin-like superfamily, the explicit consideration of π-helices provides many novel insights from specific predictions of a π-helix-containing dimeric primordial precursor of the family to the involvement of π-helices in key steps of functional diversification and the characterization of peristaltic-like shifts in the activation and substrate binding of the BMM family. This case study also underscores that α- and π-helices interconvert both directions during evolution, with the predominant direction being the creation of π-helices, and that, as can be rationalized from first principles, their gain and loss tends to be correlated both temporally (in the evolutionary process) and spatially (in the protein structure) with changes in functionality.

This analysis demonstrates the insertional origin of π-helices as a general phenomenon of protein evolution. It also resolves long standing debates about the existence of π-helices in proteins and paves the way for this generally overlooked motif to become a noteworthy feature that will aid tracing the evolution of many protein families, guide investigations of protein and π-helix functionality, and contribute additional tools to the protein engineering toolkit. Given the functional and evolutionary importance of π-helices, we recommend that contrary to current practice,

π-helices should be given precedence over α-helices in secondary structure assignment. This will allow them to be identified rather than overlooked and for those applications in which π-helices are not of interest, all π-helical residues could be counted as α-helical. Until this change happens, our π-HUNT script can be used to identify π-helices based on DSSP output. Toward the goal of identifying these cryptic and functionally relevant structures, we provide an initial framework in the form of a comprehensive list of 2967 π-helices present in over 2400 chains (see Supplemental Data Set found at the *J. Mol. Biol.* website) that was derived from our analysis of all chains in the PDB with <90% sequence identity.

Materials and Methods

The initial π-helix dataset. We have adopted the well accepted definition that helices (α, 3_{10} and π) are fundamentally defined by the presence of two sequential main chain hydrogen bonds of the correct geometry (IUPAC 1970). This is the criteria used by the common automated secondary structure assignment programs DSSP (Kabsch and Sander 1983) and STRIDE (Frishman and Argos 1995), and is also the criterion used by Fodje & Al-Karadaghi (Fodje and Al-Karadaghi 2002). For the identification of π-helices, we use here the standard criteria that these consecutive main chain hydrogen bonds be between residues five-apart in sequence (called π-type H-bonds) and for residues internal to the π-helix, the (ϕ,ψ)-values lie within the broadly defined α-helical region of the Ramachandran plot.

The 104 naturally occurring π-helices previously identified by Fodje & Al-Karadaghi (Fodje and Al-Karadaghi 2002) were reclassified using the precise definitions based on backbone H-bond information as annotated by DSSP (Kabsch and Sander 1983). As backbone NH groups can be involved in H-bonded interactions with more than one accepter, we only considered as "qualifying" those π-type H-bonds for which the π-type H-bond was the strongest of the H-bonds from that donor. Qualifying π-type H-bonds were further described as strong (≤ -2.0 kcal/mol), medium

(-1.9 to -1.0 kcal/mol) and weak (-0.9 to -0.5 kcal/mol). A π-helix was initiated by two sequential qualifying π-type H-bonds where one of them had to be of medium strength or stronger. Any break in the π-type H-bonding ended a π-helix. Figure 4.2 provides an example of this classification process.

PDB searches using π-HUNT. A non-redundant list of PDB protein chains was generated using the CullPDB server (Wang and Dunbrack 2003) such that (i) no two protein chains had greater than 25% sequence similarity, (ii) every structure was determined at or better than 2.5 Å using x-ray crystallography, and (iii) the associated R-factor for each structure was <0.25. Another non-redundant list of PDB chains was generated in the same manner except that no two chains had greater than 90% sequence identity. A Perl script was written to analyze the DSSP file of each structure in these lists (files were obtained from ftp://ftp.cmbi.kun.nl/pub/molbio/data/dssp/) and return π-helices based on π-type hydrogen bonding patterns and (ϕ,ψ)-angles. In terms of π-type H-bonding patterns, three different criteria were used: WW (at least two sequential π-type hydrogen bonds with strengths ≤ -0.5 kcal/mol), MW (at least two sequential π-type hydrogen bonds where the strength of at least one is ≤ -1.0 kcal/mol and the other is at least ≤ -0.5 kcal/mol) and MM (at least two sequential π-type hydrogen bonds with strengths ≤ -1.0 kcal/mol). In terms of the (ϕ,ψ)-criteria, torsion angles of residues internal to the π-helix were restricted to the broad helical region as defined by STRIDE (Frishman and Argos 1995) ($-180^o < \phi < 0^o$, $-120^o < \psi < 45^o$). This script, called π-HUNT, is available from the authors upon request. Searches were also performed that returned only those residues formally annotated by DSSP as π-helical. Comparison of the results from these searches was used to assess the robustness of each criterion.

Identification of α-/π-helical homologous pairs. To explore the evolutionary relationships between naturally occurring π-helices and α-helices, homologs of π-helix-containing proteins with one less residue and an α-helix in the region homologous to the π-helix were initially identified by visual inspection of the

structure-based sequence alignments provided by the Families of Structurally Similar Proteins (FSSP) database (Holm *et al.* 2008). Homology between identified α-/π-helix pairs was then confirmed by SCOP classification (Murzin *et al.* 1995), visual inspection, and consideration of similarity in protein function.

Phylogenetic analysis of the ferritin-like superfamily. Representative protein structures for each of the major families within the ferritin-like superfamily (Gomes *et al.* 2001) were selected from the Protein Data Bank (Berman *et al.* 2000). Structure-based sequence alignments for each of these proteins were obtained from the FSSP database (Holm *et al.* 2008) and manually adjusted if needed to provide consistency in the multiple sequence alignment. For the proteins in Fig. 4.7a without a known structure, sequences were added to the structure-based sequence alignment profile using the CLUSTALW algorithm (Thompson *et al.* 1994). The phylogenetic tree was then made as previously reported (Leahy *et al.* 2003) using maximum likelihood methods with the JTT model for amino acid substitutions (Jones *et al.* 1992).

Acknowledgements

We thank Joe Thornton, Jacque Fetrow and Brian Matthews for critique of an earlier manuscript. This work was supported in part by the General Medical Institute, NIH grant GM R01-083136 and the Oregon Agricultural Experiment Station.

Abbreviations

H-bonds: hydrogen bonds, MMOH: soluble methane monooxygenase hydroxylase, PH: phenol hydroxylase, ToMO: toluene monooxygenase, BMM: bacterial multi-component monooxygenases, RNR: ribonucleotide reductase, DSSP: Definition of Secondary Structure in Proteins, FSSP: Families of Structurally Similar Proteins.

Table 4.1. Classification of π-helices in the representative dataset and the current Protein Data Bank

H-bonds in π-helix [a]	Number of π-helices [b]	Number with α-helical homologs [c]	Protein Data Bank [d]
2	82	70	822
3	13	9	138
4	7	5	26
5	4	4	20
6	0	0	3
7	0	0	1
Total	106	88	1010

[a] The number of (i+5) → i hydrogen bonds in the π-helix. To qualify as a π-helix, at least two sequential (i+5) → i hydrogen bonds must exist where one is ≤-0.5 kcal/mol and the other is ≤ -1.0 kcal/mol

[b] From the reclassified Fodje & Al-Karadaghi (Fodje and Al-Karadaghi 2002) dataset

[c] Putative homologs of the π-helix-containing proteins from the Fodje & Al-Karadaghi (Fodje and Al-Karadaghi 2002) dataset with a pure α-helix at the equivalent position.

[d] Number of π-helices in 5620 diverse (<25% sequence identity with one another) protein chains determined at ≤ 2.5 Å resolution.

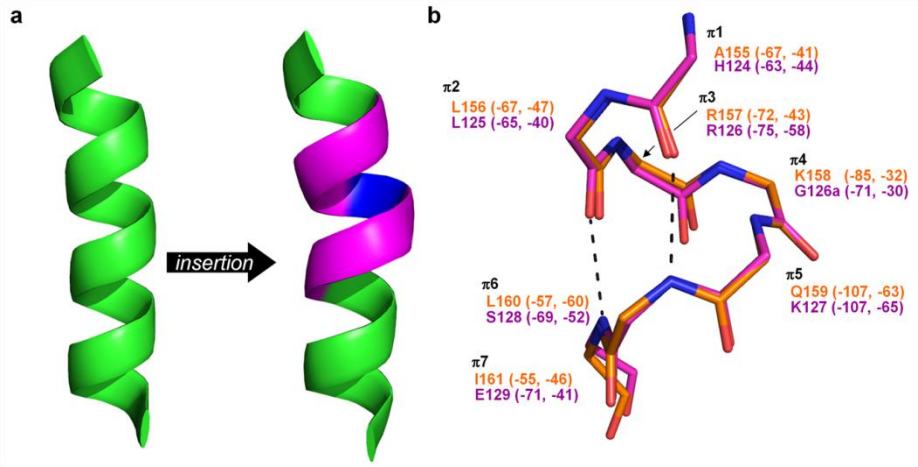


Figure 4.1. The common two-H-bond π-helix is the same motif as an engineered α-aneurism. (a) An α-helix in *Staphylococcus aureus* nuclease (left, green cartoon; PDB 1eyd) develops an α-aneurism (right, purple cartoon; PDB 1sty) when a single amino acid (highlighted blue) is inserted into it (Keefe *et al.* 1993). (b) An overlay of the π-helix from fumarase C of *Escherichia coli* (residues 155-161, orange carbon atoms; PDB 1fur) and the same engineered α-aneurism from panel a (purple carbon atoms) demonstrates their equivalence. These α-aneurism-type helical distortions have been independently characterized as looping outs (Heinz *et al.* 1993), π-bulges (Cartailler and Luecke 2004) and α-bulges (Hardy *et al.* 2000), but not as π-helices. Black-dashed lines represent the π-type H-bonds. (ϕ,ψ) torsion angles are written beside each residue. In panel b, nitrogen (blue) and oxygen (red) atoms are indicated.

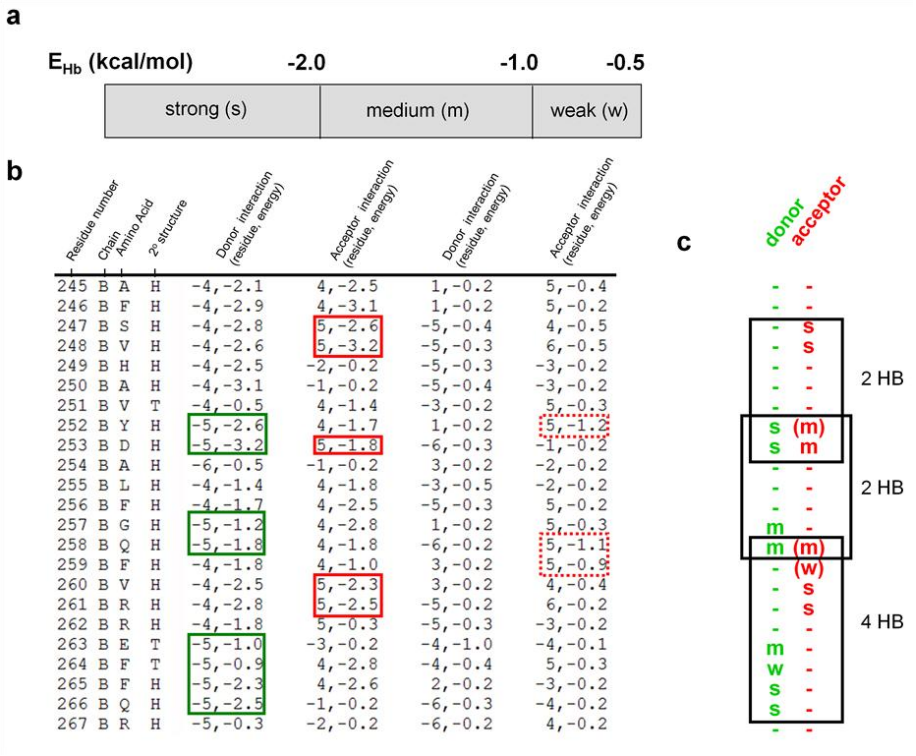


Figure 4.2. Defining π-helices based on (i+5,i) π-type H-bonds. (a) Energy cutoffs (in kcal/mol, calculated by DSSP (Kabsch and Sander 1983) for strong (s), medium (m), and weak (w) H-bonds. (b) Excerpted DSSP output for residues 245-267 of methane monooxygenase hydroxylase (MMOH) (PDB 1mty, β-chain), highlighting the π-type H-bonding patterns. Green boxes highlight π-type donor interactions and red boxes highlight the acceptor interactions. Solid outlines indicate the π-type interaction is the strongest H-bond for that residue, and dotted outlines indicate those that are the second strongest. The column contents are labeled. For each of the four H-bond interaction columns, the two numbers given are relative positions of the partner residue and the energy of the interaction (in kcal/mol). The first two of these columns are the strongest interactions and the second two columns are the second strongest interactions. (c) Short hand summary of the π-type H-bonding and the π-helix assignments. For each residue acting as a π-type donor or acceptor, the strength of the interaction is given in the appropriate column. The letter is in parentheses if the π-type interaction is not the strongest H-bond for that residue. Based on the qualifying criteria described in the *Materials and Methods*, this segment in MMOH that was previously classified as a single 20 residue π-helix with 5-(i+5,i) H-bonds (Fodje and Al-Karadaghi 2002) is seen to be three distinct, but overlapping π-helices with 2-, 2- and 4-H-bonds (see rectangles in panel c). Note that these three π-helices are cryptic in that none are assigned as π-helical (I) in the DSSP summary (fourth column).

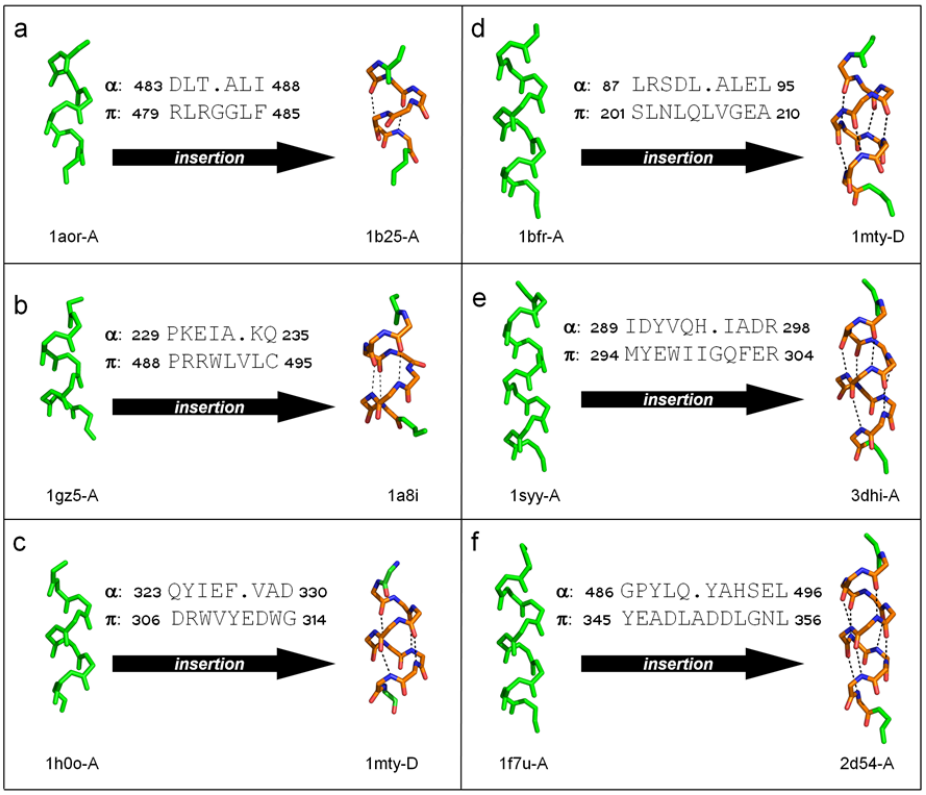


Figure 4.3. A single insertion can generate all occurring lengths of π-helices. Panels a, b, c, d, e and f show one example each of a 2-, 3-, 4-, 5-, 6- and 7-H-bond π-helix, respectively. In each panel, the equivalent α-helix (left) and π-helix (right) from a pair of homologous proteins are shown with their PDB codes. Above the black arrow that represents the insertion process is the sequence alignment from the FSSP database illustrating the single insertion. The π-helices from panels a (formaldehyde ferredoxin oxidoreductase), b (glycogen phosphorylase), and d (MMOH α-subunit; insertion πD from Fig. 4.7) are all located within the active site of their respective protein, while that shown in panel c (MMOH α-subunit) and panel e (α-subunit of Toluene-4-Monooxygenase; insertion πE from Fig. 4.7) are involved in intersubunit interactions, suggesting a functional role for these. All atoms in the α-helices are colored green, while carbon, nitrogen and oxygen atoms in the π-helices are colored orange, blue and red, respectively. Black dashed lines in the π-helices indicate the (i+5,i) π-type hydrogen bonds.

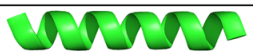
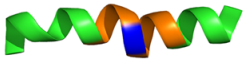
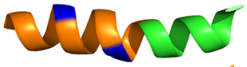
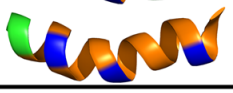
Protein	PDB	Residues	Alignment	Structure
Bacterioferritin	1bfr	86-105	MLRSD.LALEL.DGAKNL.REAI	
Class Ib RNR	1uzr	155-175	RKVAS.TLLESFLFYSGF.YLPM	
Class Ic RNR	1syy	184-205	NLVGYYIIMEGIFFYSGF.VMIL	
MMOH – β-subunit	1mty	244-266	SAFSVHAVYDALFGQFVRREFFQ	

Figure 4.4. Evidence that overlapping π-helices result from stepwise insertions of single amino acids. In the ferritin-like superfamily, putative ancestral forms in the evolution of the three, overlapping π-helices in the β-subunit of MMOH (shown in Fig. 4.2) (bottom) are represented by the structurally equivalent helices from bacterioferritin (top), which is a pure α-helix, the R2 subunit of class Ib ribonucleotide reductase (second from top), which includes one π-helix, and the R2 subunit of class Ic ribonucleotide reductase (third from top), which includes overlapping π-helices. Insertion events represented are designated πB, πD and πJ in Figure 4.7. Amino acid alignments indicate the effective points of insertion. Lines below each sequence indicate the π-helical regions. π-helices are colored orange, α-helices are colored green, and the inserted amino acid, as defined by structural alignment, is colored in blue.

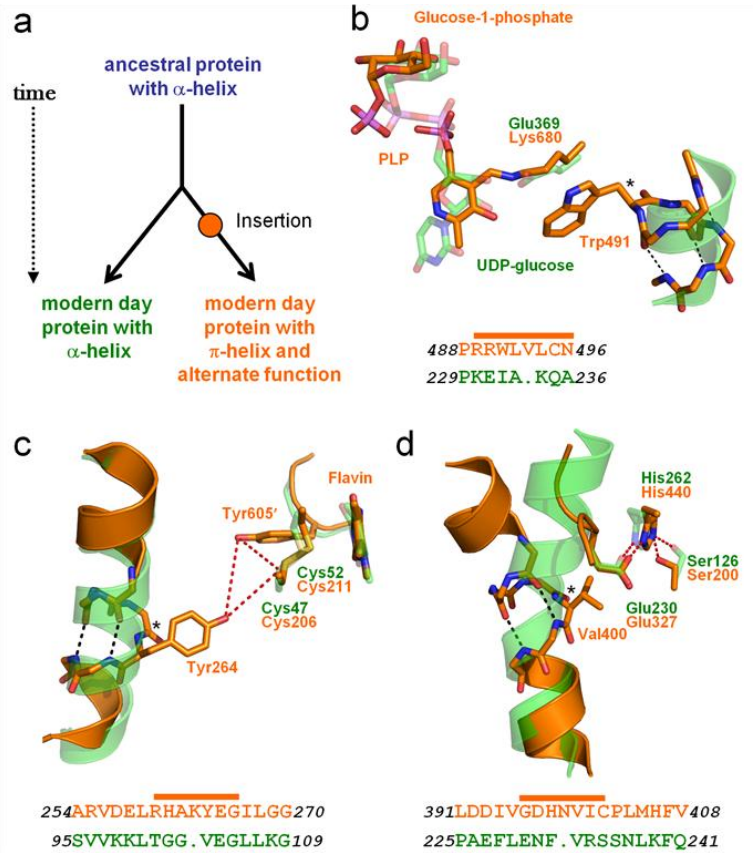


Figure 4.5. Three examples of α-helix derived active site π-helices. (a) Evolutionary path by which π-helices are derived from α-helices via a single residue insertion (orange circle). (b) Shown is the π-helix-containing PLP-bound active site of PLP-dependent glycogen phosphorylase (orange carbon atoms; PDB 3gpb) overlaid onto the α-helix-containing active site of trehalose-6-phosphate synthase (semitransparent green carbon atoms; PDB 1uqu). The structurally derived sequence alignments are shown below the structure (orange and green letters correspond to the π- and α-helix-containing sequences, respectively), which demonstrate the presence of the inserted residue that forms the π-helical segment (orange bar). (c) The active site π-helix of mercuric ion reductase (orange carbon atoms; PDB 1zk7) is overlaid on the homologous α-helix of dihydrolipoamide dehydrogenase (semitransparent green carbon atoms; PDB 1ebd). (d) A π-helix of acetylcholine esterase (orange carbon atoms; PDB 1ea5) is overlaid onto the homologous α-helix of mycolyl transferase (green transparent carbon atoms; PDB 1f0p). In all panels, dashed-black lines show the π-type H-bonds and red dashed lines show interactions within the active sites. Nitrogen (blue), oxygen (red), sulfur (yellow) and phosphorus (purple) atoms are indicated.

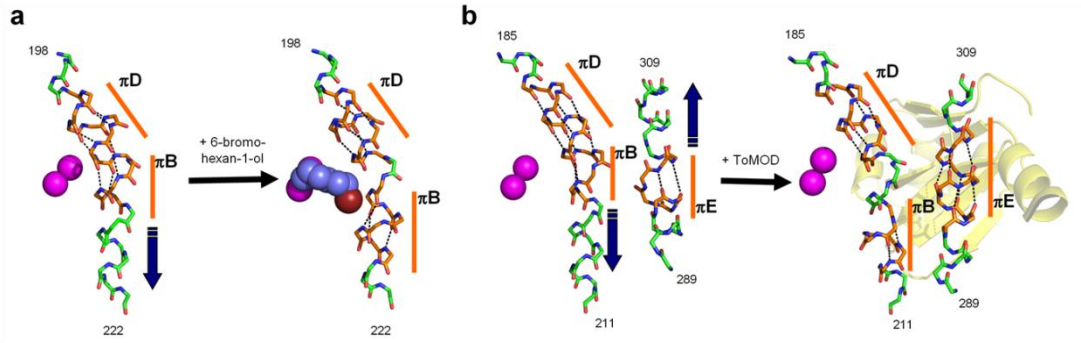


Figure 4.6. π-helical peristalsis in the active site of BMM enzymes. (a) Two π-helices (labeled πB and πD) in the active site of MMOH are overlapping in the resting state (left). Upon binding the product analogue 6-bromohexan-1-ol (blue spheres), these two π-helices no longer overlap due to a shift in πB (right). (b) The resting state structure of the toluene-4-monooxygenase hydroxylase also shows two overlapping π-helices at the active site (left). Upon binding of the regulatory component (ToMOD, semitransparent yellow), πB shifts identically to that of product-bound MMOH. As πB moves downward, an adjacent π-helix (labeled πE), which is sandwiched between the active site π-helix and the regulatory subunit, simultaneously elongates and shifts upward. For details of the π-type H-bonds, see Figure A1.3. Orange bars mark individual π-helices and blue arrows the direction of π-helix movement. π-type H-bonds are shown by black dashed lines. π-helix (orange) and α-helix carbon atoms are indicated, as are nitrogen (blue), oxygen (red), iron (purple) and bromine (brown). PDB codes are 1mty, 1xvb, 3dhg and 3dhi.

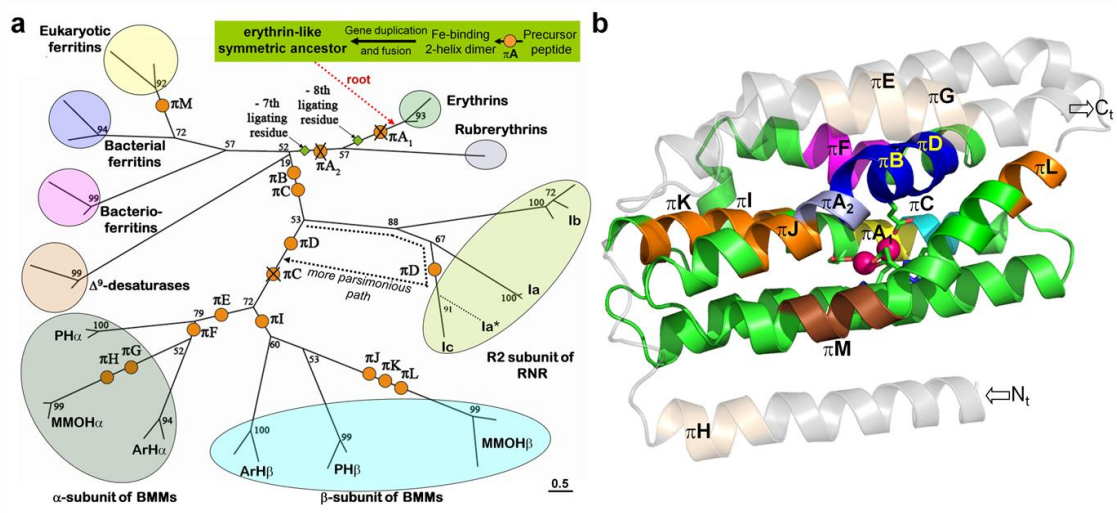


Figure 4.7. π-helices and the evolution of the ferritin-like superfamily. (a) Evolutionary tree showing that, from a symmetrical erythrin-like ancestor containing two π-helices (Fig. 4.8), a simple set of insertion (orange circles) and deletion (crossed-out orange circles) events can account for all π-helices seen in modern-day family members (Table A1.5). The branching shown is based on maximum likelihood analysis as previously described (Leahy *et al.* 2003) from structurally-derived sequence alignments (bootstrap values shown). Organism names, omitted for clarity, are shown in Fig. A1.4. The length of the bar (bottom right) corresponds to 0.5 substitutions/site. New abbreviations are PH: phenol hydroxylase family and ArH: aromatic monooxygenase hydroxylase family. (b) Ribbon diagram of the α-subunit of MMOH (PDB 1mty, chain D) showing the locations of its six π-helices (πB, πD, πE-πH) and also showing the locations of the remaining eight π-helices seen in other members of the ferritin-like superfamily (πA_1, πA_2, πC, πI-πM). Indicated are the core four-helix bundle (green), the surrounding helices (transparent gray), and the π-helices πA_1 (yellow), πA_2 (light blue), πC (cyan), πB and πD (blue), πE and πG-πL (orange), πF (purple) and πM (brown). Iron atoms (pink spheres) and the N- and C-termini are shown.

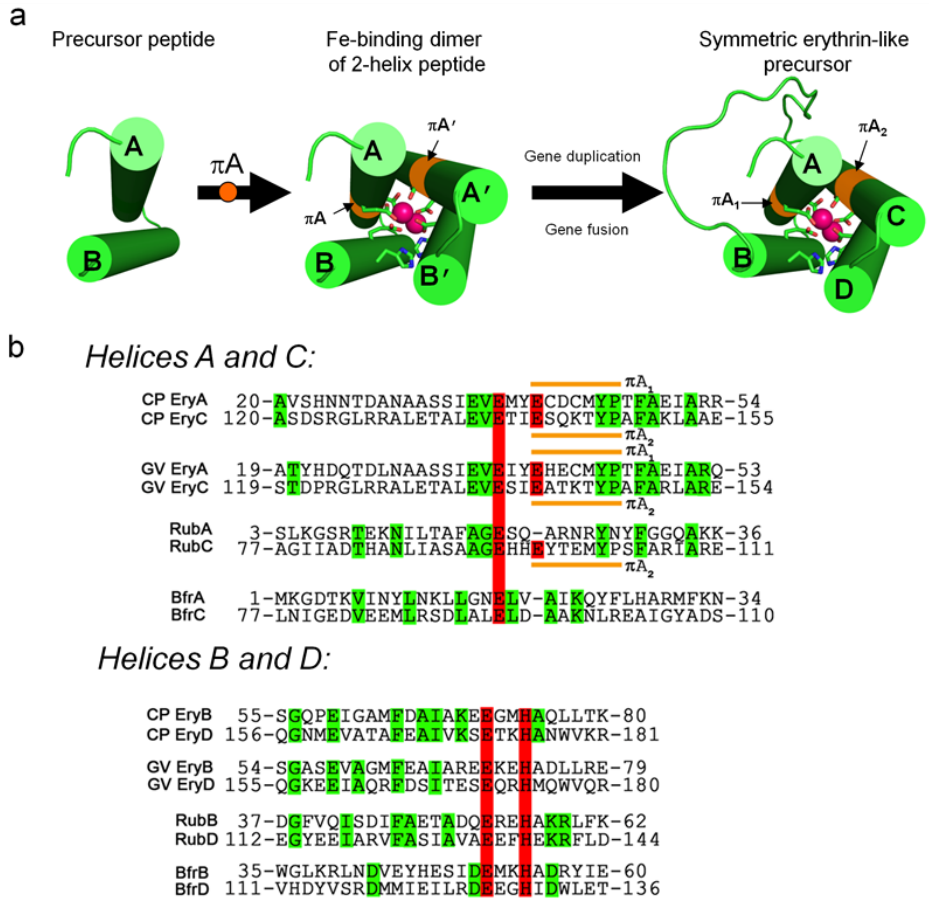


Figure 4.8. Model for the origins of the ferritin-like superfamily. (a) Scheme showing the formation of the four-helix bundle characteristic of the ferritin-like superfamily via gene duplication and fusion of a primordial two-helix precursor that, through a single insertion πA (orange circle), gave rise to two symmetry related π-helices (πA_1 and πA_2) and eight ligating residue in the erythrin-like precursor of this superfamily. (b) Sequence alignments showing the residual internal sequence similarity in erythrins and, to a lesser degree, rubrerythrin (Kurtz and Prickril 1991). Shown are the alignments of helix A with helix C and helix B with helix D in erythrin from *Cyanophora paradoxa* (CP Ery), erythrin from *Gloeobacter violaceus* (GV Ery), rubrerythrin from *Desulfovibrio vulgaris* (Rub) and bacterioferritin from *Escherichia coli* (Bfr) included as an example of a canonical superfamily member having only six ligating residues and less recognized internal symmetry. Red highlighting indicates residues involved (or proposed to be involved in the case of erythrin) in iron-ligation. Green highlighting indicates additional residues that are identical between homologous helices within the same protein. Orange bars indicate π-helical residues in rubrerythrin, and potential π-helical residues in erythrin.

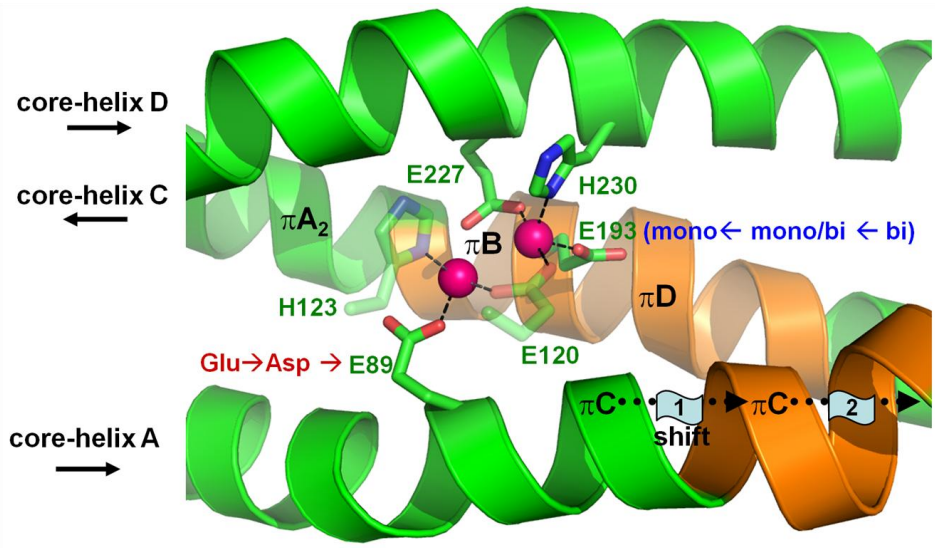


Figure 4.9. π-helices πA_2- πD correlate with changes in metallocenter geometry. Shown is an annotated view of the di-nuclear site in RNR Ic (PDB 1syy) with the backbone ribbon of core-helix B of the four-helix bundle, contributing H123 and E120, omitted for clarity. Core helices A, C and D and their direction from N- to C-terminus are labeled on the left, and π- (orange) and α-helices (green) are indicated. As described in the text, the proximity of πA_2-πD to their respective di-nuclear centers are shown and suggest they play a role in influencing metallocenter chemistry.

Chapter 5

A diiron protein autogenerates a Valine-Phenylalanine crosslink

Richard B. Cooley, Timothy W. Rhoads, Daniel J. Arp and P. Andrew Karplus

Published in *Science* (2011), **332**(6032), 929.

Abstract

All known internal covalent cross-links in proteins involve functionalized groups having oxygen, nitrogen, or sulfur atoms present to facilitate their formation. Here, we report a carbon-carbon cross-link between two unfunctionalized side chains. This valine-phenyalanine cross-link, produced in an oxygen-dependent reaction, is generated by its own carboxylate-bridged diiron center and serves to stabilize the metallocenter. This finding opens the door to new types of posttranslational modifications, and it demonstrates new catalytic potential of diiron centers.

Main Text

Many proteins generate covalent crosslinks between two or more of their own amino acids to create a cofactor near their active site (Xie and van der Donk 2001). Although the mechanisms of crosslink formation and the function of the crosslinked group are not always understood, all characterized crosslinks to date involve side-chains with nitrogen, oxygen or sulfur functional groups to facilitate the making and breaking of the covalent bond(s).

Following up our earlier work on π-helices in proteins (Cooley *et al.* 2010), we have solved the structure of a rubrerythrin-like protein that appears to conserve certain ancestral features of the ferritin-like superfamily. We call this protein symerythrin to reflect its high level of internal symmetry. The 1.20 Å resolution structure of symerythrin from the photosynthetic eukaryote *Cyanophora paradoxa* (R/R_{free} = 10.1/12.5%, Figs. 5.1A and A2.1, Table A2.1) reveals an unusual covalent crosslink connecting the aliphatic $C_{\gamma 1}$-atom of Val127 to the aromatic C_{δ}-atom of Phe17 (Fig. 5.1B) adjacent to its carboxylate-bridged diiron metallocenter (details to be described elsewhere). Liquid chromatography-tandem mass spectrometry (LC-MS/MS) analysis of in-solution proteolytic digests confirmed the location of the crosslink and its presence in the protein before crystallization (Fig. 5.1C). This crosslink is notable because it occurs between unfunctionalized side chains.

From our symerythrin expression system, crosslinked (~10%) and non-crosslinked (~90%) forms could be purified separately, allowing us to prove that the crosslink was generated by symerythrin's own metallocenter. Indeed, aerobic incubation of non-crosslinked, iron-reconstituted protein with an electron donor led to complete conversion to the crosslinked form (Fig. 5.1D). The dependency of crosslink formation on reductant, O_2, and an intact diiron center implies that symerythrin's metallocenter forms high-valent diiron-oxygen intermediates resembling those in soluble methane monooxygenase (Shu *et al.* 1997) and also proposed to be involved in formation of a recently discovered Tyr-Val ether crosslink in a ribonucleotide

reductase R2-like protein (Andersson and Hogbom 2009). By analogy with these systems (Kopp and Lippard 2002), we hypothesize that as a Q-type (i.e. Fe(IV)-Fe(IV)) intermediate forms, Val127 rotates so its $C_{\gamma 1}$-atom contacts the activated metal-bound oxygen, which abstracts a hydrogen atom to generate a primary alkyl radical. Val127 then rotates back and adds to the Phe17 phenyl ring to create a cyclohexadienyl radical. Transfer of a single electron to the diiron center, coupled with deprotonation, then generates the crosslink.

While the physiological function of symerythrin remains unknown, evidence shows that the diiron center is stabilized by the crosslink (Fig. 5.1E), which both anchors the N-terminal tail to the core of protein and creates a putative substrate binding pocket (Fig. 5.1B). Given the nature of the crosslink and the significant effort that has been placed in characterizing such hydrocarbon chemistry for its relevance in biological, bioenergetic and environmental applications (Bergman 2007), further study will be of great interest.

Acknowledgements

This work was supported in part by NIH grant GM R01-083136 (P.A.K.), the Oregon Agricultural Experiment Station (D.J.A.), and NIEHS Environmental Health Sciences Center Grant ES00210. We thank J.M. Bollinger and C. Krebs for discussions.

Protein Data Bank Depositions

Coordinates and structure factors of diferric crosslinked symerythrin are deposited as Protein Data Bank entry 3qhb.

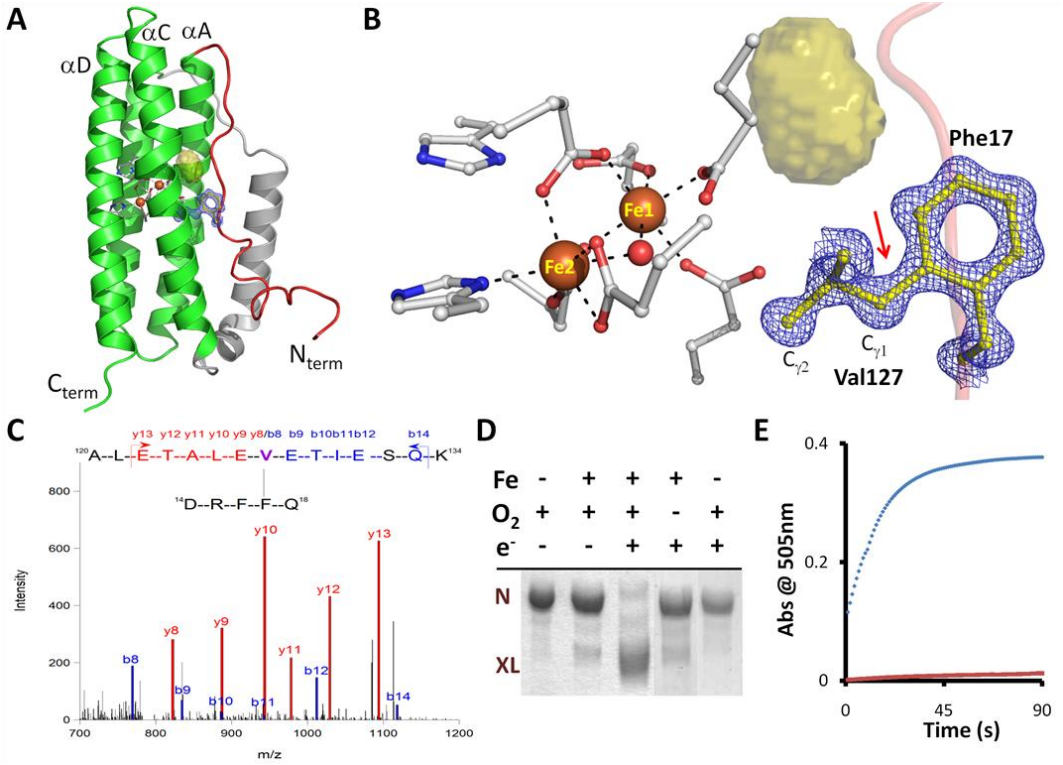


Figure 5.1. The Val-Phe crosslink. (A) The core four helices of symerythrin (green ribbon with helices αA, αC and αD labeled), the N-terminal tail (red coil), the putative substrate binding cavity (olive surface), the diiron center (brown spheres) and the ligating residues (C, N and O atoms are gray, blue and red) are shown. Electron density (blue mesh contoured at $2.2\rho_{rms}$) shows the covalent bond linking Val127 and Phe17 (yellow atoms). (B) Close-up of the cross-link (indicated by the red arrow) and diiron center (having one additional Glu ligand compared with rubrerythrins) with the same coloring as panel A. Val127 $C_{\gamma 2}$ is 4.5 Å from the bridging oxygen. (C) The MS/MS fragmentation pattern from the triply-charged parent ion at m/z 791.07 identified in proteolytic digests of crosslinked symerythrin. The parent ion mass (2370.181 Da) matches that of residues 120-134 linked to residues 14-18 (calculated MW 2370.182). (D) Purified non-crosslinked symerythrin (N, lane 1) was reconstituted with iron and exposed to O_2 (lane 2) prior to a 40 h incubation with an electron source (e^-) during which all symerythrin is converted to the crosslinked form (XL, lane 3). Crosslink formation required O_2 (lane 4) and iron (lane 5). (E) Incubation of a chromogenic iron chelator to diferrous non-crosslinked symerythrin (blue line) shows rapid iron removal compared to that of diferrous crosslinked symerythrin (red line).

Chapter 6

Symerythrin structures at atomic resolution and the origins of rubrerythrins and the ferritin-like superfamily

Richard B. Cooley, Daniel J. Arp and P. Andrew Karplus

Submitted to *Journal of Molecular Biology*

Abstract

Rubrerythrins are diiron containing peroxidases belonging to the ferritin-like superfamily (FLSF). Recently, we reported that a novel rubrerythrin variant called symerythrin from the oxygenic phototroph *Cyanophora paradoxa* was capable of autogenerating an unusual valine-phenylalanine crosslink. Here, we describe detailed characterizations of the overall structure of symerythrin with its metallocenter in three different states: diferric, azide-bound diferric and chemically reduced. The results reveal a unique and remarkably dynamic metallocenter containing an additional iron-ligating glutamate and π-helical segment not found in rubrerythrins. Reduction of the diiron center results in a 2.0 Å shift in the position of Fe1 as seen with rubrerythrins, suggesting similar functionality between the two. Of particular interest is the high internal symmetry of symerythrin, which supports the previously proposed ancestral nature of symerythrin and the notion that its core four-helix bundle was formed by the gene duplication and fusion of a two-helix peptide. Comparisons with another FLSF family distantly related to the rubrerythrins but with notable internal symmetry provides compelling evidence that, contrary to previous assumptions, there have been multiple gene-fusion events that have generated the single chain FLSF fold.

Introduction

The ferritin-like superfamily (FLSF) is a diverse group of metalloproteins named for the ferritins that are responsible for iron storage in prokaryotes and eukaryotes (Andrews 2010). Additional well-studied proteins within the FLSF include the bacterioferritins, rubrerythrins, Mn-catalases, DNA-binding proteins from starved cells (DPS), Δ^9-desaturases, the R2 subunit of class I ribonucleotide reductases, and bacterial multicomponent monooxygenases (Andrews 2010). Each of these proteins have a conserved core four-helix bundle (with helices referred to as A, B, C and D) that binds a carboxylate-bridged dinuclear metallocenter through six positionally conserved ligating residues. The canonical FLSF metallocenter ligands are four carboxylate and two histidine residues (i.e. a six-ligand motif) symmetrically arranged such that helix pairs A-B and C-D of the four-helix bundle each contribute two carboxylates and one histidine (Andrews 2010). This symmetry of the ligands has led to the proposal that the core four-helix bundle of the FLSF originated via the duplication and fusion of a gene encoding a simple two-helix protein that formed a symmetric diiron-binding dimer (Andrews 1998; Andrews 2010).

The FLSF dinuclear centers are commonly diiron centers (with sites referred to as Fe1 and Fe2) which facilitate oxygen mediated redox reactions, although some FLSF enzymes have a dimanganese (Cotruvo and Stubbe 2011) or a heteronuclear Mn/Fe cofactor (Jiang *et al.* 2007). In all cases, the reduced form of the dinuclear center is needed to react with substrate, and the redox changes are commonly associated with so-called 'carboxylate shifts' that alter the coordination geometry of the metallocenter. This conformational plasticity is a hallmark feature of FLSF metallocenters and has been the subject of numerous structure-function analyses (see review by Sazinsky and Lippard (Sazinsky and Lippard 2006).

The exception to the six-ligand motif pattern is the rubrerythrin family (including rubrerythrins, nigerythrins, and sulerythrins). These proteins are unusual among FLSF proteins in that they contain a seventh metal ligating residue. This

additional ligand is a glutamate which resides within a π-helical bulge of core helix C and reflects an insertion compared to other FLSF proteins (Cooley *et al.* 2010); it gives rubrerythrins the unusual ability among FLSF proteins to preferentially react with hydrogen peroxide rather than dioxygen (Coulter *et al.* 2000). As a result, rubrerythrins are believed to act as peroxidases *in vivo* in order to mitigate toxicity from reactive oxygen species (Zhao *et al.* 2007), particularly in anaerobes where they are widely distributed (Gomes *et al.* 2001). Interestingly, rubrerythrins have recognizable internal sequence similarity that goes beyond just the metallocenter ligands and this is thought to reflect the origins of the fold from a gene duplication and fusion event (Kurtz and Prickril 1991; Andrews 1998).

Rubrerythrins have the most plastic metallocenter among FLSF proteins in that although seven residues are involved in metal ligation, only six coordinate the metallocenter at any one time. In the diferric state the His from helix B is not a ligand, but on reduction to the diferrous state, Fe1 shifts ~2.0 Å so that it breaks its bond with the additional glutamate from helix C and makes a new ligating bond with the His from helix B (Jin *et al.* 2002; Iyer *et al.* 2005). This redox-dependent toggling of Fe1 positions is thought to be part of rubrerythrin's catalytic cycle (Dillard *et al.* 2011).

During a recent analysis of π-helices in the FLSF (Cooley *et al.* 2010), we noted that genes found only in the oxygenic phototrophs *Gloeobacter violaceus* and *Cyanophora paradoxa* (NCBI gene IDs 801647 and 2602414, respectively) encode a distinct variant of the rubrerythrin family having even higher internal sequence similarity than the characterized rubrerythrins. The encoded protein sequences include an inserted glutamate in helix A that, according to the internal symmetry of the FLSF four-helix bundle, is placed symmetrically to the ‘additional’ glutamate (inserted in helix C) of the rubrerythrins (Cooley *et al.* 2010). Thus, we proposed this glutamate in helix A reflects an ancestral feature of the rubrerythrin family and serves as an unprecedented eighth metal ligating residue residing in a π-helical segment of helix A symmetric with the π-helical segment of helix C of rubrerythrins.

As a first step to characterize these novel rubrerythrin-like proteins, we have cloned and expressed the *C. paradoxa* protein and solved its crystal structure at 1.20 Å resolution. As we have briefly reported elsewhere, this protein, now named symerythrin[†] to reflect its high level of internal symmetry, autogenerates an unprecedented carbon-carbon crosslink between unfunctionalized Val and Phe side chains (Cooley *et al.* 2011). Here, we provide a detailed description of the overall structure of symerythrin and its novel diferric metallocenter in oxidized, azide-bound and chemically reduced forms. The plasticity of this eight-residue metallocenter and the implications of this structure for understanding the origins of rubrerythrin-like proteins are also discussed.

Results

Expression and purification of recombinant symerythrin.

Both *G. violaceus* and *C. paradoxa* symerythrin were expressed without an affinity tag in order to minimize artifactual metal binding during overexpression and handling. All studies reported here are based on the *C. paradoxa* protein, for which the expression was more efficient and reliable. Recombinant *C. paradoxa* symerythrin was obtained in major and minor forms (~10:1) migrating on denaturing gels at apparent molecular weights of 20 and 17 kDa, respectively. Mass spectrometric analyses showed that both forms were full-length symerythrin and both had the N-terminal methionine cleaved during overexpression in *E. coli*. The only difference between the two forms was that the lower band contained the Val127-Phe17 crosslink described earlier (Cooley *et al.* 2011). For characterization, these isoforms could be separated from each other through common chromatographic procedures (see *Materials and Methods*).

[†]The protein now named symerythrin (Cooley *et al.* 2011) was originally referred to by us (Cooley *et al.* 2010) and Andrews (Andrews 1998) as 'erythrin' because it appeared to have a rubrerythrin-like metallocenter and no rubredoxin domain. Later, Andrews (Andrews 2010) used 'erythrin group' for a very diverse set of FLSF proteins without a rubredoxin domain and 'erythrin subfamily' for a specific subset having notable internal sequence similarity. Here, we use 'erythrins' to refer to this latter group.

Solution characterization of crosslinked and non-crosslinked symerythrin

In contrast with known rubrerythrins and rubrerythrin-like proteins, which are dimeric, analytical gel-filtration and dynamic light scattering analyses demonstrate that both isoforms of symerythrin are monodisperse monomers at plausibly physiologic conditions (pH 7.4, 150 mM NaCl) (data not shown). In terms of metal content, crosslinked symerythrin purified with ~2 Fe per monomer (1.7-1.8 by the ferrozine assay (Percival 1991), 2.2-2.3 by inductively coupled plasma optical emission spectroscopy (ICP-OES)) and no detectable quantities of manganese or zinc. That the iron remains bound even after elution from a phenyl sepharose column shows that the metallocenter of crosslinked symerythrin is stable in up to 40% ethanol. Non-crosslinked symerythrin, on the other hand, purified with just 0.4-0.5 iron atoms per monomer. Reconstitution of the non-crosslinked form by incubation with excess ferrous iron resulted in 1.6-1.8 iron atoms per monomer. As we reported earlier, the non-crosslinked iron center is labile to the iron chelator 1,10-phenanthroline, but the crosslinked metallocenter is not (Cooley *et al.* 2011).

The UV/visible spectra of resting iron-loaded symerythrin displayed a distinctive absorption band from 300-350 nm (Fig. 6.1a) typical of the oxo-to-Fe(III) charge transfer of an Fe(III)-Fe(III) (e.g. diferric) center (Fox *et al.* 1993). Anaerobic incubation of this diferric metallocenter with two reducing equivalents of sodium dithionite in the presence of methylviologen resulted in the loss of this band (Fig. 6.1a and inset). Additionally, the titration of diferric symerythrin with sodium azide yielded a new chromophore at 450 nm (Fig. 6.1b) as seen with other diferric-azide adducts (Fox *et al.* 1993; Gupta *et al.* 1995). The azide binding showed saturation with a K_d of ~50 mM (data not shown). In all of these experiments, non-crosslinked and crosslinked symerythrin behaved similarly (Fig. 6.1).

The structure of oxidized crosslinked symerythrin

Overall fold. All structures of symerythrin reported here are of the crosslinked form because the non-crosslinked form did not crystallize in any of the tested conditions. The structure of oxidized crosslinked symerythrin as seen in crystals grown at pH 5.5 was solved by molecular replacement and refined at 1.20 Å resolution to an R/R_{free} of 0.101/0.125 with reasonable geometry (Table 6.1) (Cooley *et al.* 2011). The complete chain of crosslinked symerythrin, from residues 2-180, was modeled in both molecules in the asymmetric unit. No substantive differences existed between the two chains in the asymmetric unit (0.04 Å RMSD for 179 Cα-atoms), and all descriptions of symerythrin provided here are based on molecule A.

Symerythrin adopts the expected up-down-down-up four-helix bundle fold characteristic of the FLSF with two metal atoms bound at the core of the protein (Figs. 6.2a and 6.2b). These metal ions were assigned as irons at 100% occupancy based on quantitative trace metal analyses. Helices B and C are connected by a 31 residue extended 'linker' which has a 5-residue stretch of P_{II}–spiral (i.e. polyproline-II or polypeptide-II (Hollingsworth and Karplus 2010)) and two α-helices. The N-terminal segment preceding helix A adopts an extended conformation, including 5- and 6-residue segments of P_{II}-spirals, and lies in the groove between core helices A and C; it is covalently anchored to helix C via the Val127-Phe17 crosslink. The peptide chain terminates shortly after helix D with a short, 3-residue C-terminal tail.

Like rubrerythrins, symerythrin contains a π-helical segment in helix C (πC) adjacent to the metallocenter. In addition, as predicted (Cooley *et al.* 2010), a π-helical segment (πA) is present in helix A of symerythrin (Fig. 6.2b). Each of these two π-helices has a Pro at its C-terminal end (Fig. 6.2a) and is associated with a ~20^{o} kink in its core helix. These two kinked helices pack such that a pocket is created roughly halfway up the four-helix bundle adjacent to the metallocenter. This pocket is isolated from the outside solvent by the N-terminal tail, which is secured to the core four-helix bundle via the Val127-Phe17 crosslink (Fig. 6.2b). A DALI (Holm *et al.* 2008) search revealed nigerythrin (PDB 1yux) as the structurally characterized protein

most similar to symerythrin (Z-score = 17.9, 27% sequence identity, Cα RMSD of 1.8 Å over 137 residues).

A novel variant of the FLSF metallocenter. Given eight potential metal ligands (six Glu and two His), describing the metallocenter coordination and comparing it with other FLSF proteins (having different residue numbers) can be rather confusing. To facilitate this, we introduce here a nomenclature in which each ligand residue is uniquely identified by the helix from which it comes, with the eight ligands thus being identified as $Glu_{\alpha A}$, $Glu_{\pi A}$, $Glu_{\alpha B}$, $His_{\alpha B}$, $Glu_{\alpha C}$, $Glu_{\pi C}$, $Glu_{\alpha D}$, and $His_{\alpha D}$. In symerythrin, these correspond to Glu37, Glu40, Glu71, His74, Glu128, Glu131, Glu162, His165, respectively (Fig. 6.2a). This nomenclature emphasizes the internal symmetry of the fold and allows residues involved in the metallocenter to be uniquely identified without the use of specific residue numbers. In addition to the direct metal ligands, FLSF metallocenters often have a conserved pair of Tyr residues from helices A and C that provide second-shell support for the metallocenter. In symerythrin, these are Tyr45 and Tyr136. Designating these as $Tyr_{\alpha A}$ and $Tyr_{\alpha C}$, respectively, $Tyr_{\alpha A}$ hydrogen-bonds to $Glu_{\alpha C}$ and $Tyr_{\alpha C}$ hydrogen-bonds to $Glu_{\alpha A}$.

As seen at 1.20 Å resolution, the coordination geometry of the carboxylate-bridged diiron center is very well defined (Fig. 6.3a). The iron atoms are 3.6 Å apart, compared with just 3.3 Å in rubrerythrin. Both metals have octahedral coordination involving a μ-oxo/hydroxo bridge. The coordinating residues match those known in the seven-ligand metallocenters of rubrerythrins plus one additional residue, $Glu_{\pi A}$, acting as a monodentate ligand to Fe1. The electron density shows that $Glu_{\pi A}$ adopts two equally populated conformations (Fig. 6.3a). The two Fe1 coordination sites occupied by the monodentate interactions of symerythrin's $Glu_{\alpha A}$ and $Glu_{\pi A}$ are filled by a bidentate interaction of $Glu_{\alpha A}$ in rubrerythrins (Fig. 6.3b). This change makes the octahedral coordination of Fe1 in symerythrin less distorted than that of rubrerythrin. In contrast, the coordination geometries of Fe2 in symerythrin and rubrerythrin are nearly identical. Also as seen for other rubrerythrins, the $His_{\alpha B}$ side

chain is not involved in metal coordination in the diferric state, but in contrast to those structures it enters a hydrogen bond with the non-coordinating carboxylate oxygen of $Glu_{\alpha A}$. A 2.0 Å dataset collected at room temperature revealed that no significant changes in the metallocenter occur upon freezing of the crystals (data not shown).

The azide-bound diferric complex.

A 250 mM sodium azide soak slightly impaired the diffraction of the crystals, and, as little to no density was present for residues 2-4, these were left out of the model. The electron density maps at 1.40 Å resolution revealed a mixture of resting state and azide-bound metallocenters. To guide the model building, an F_o(azide)-F_o(resting) difference map was generated to provide the most direct image of the changes induced by azide binding (Fig. 6.4a). This map unambiguously revealed the orientation of the azide as well as the altered coordinating residues. The data were well fit with the metallocenter modeled as 50% resting state and 50% azide bound (R/R_{free} of 0.114/0.156, Table 6.1). Unfortunately, we were unable to obtain a structure of azide bound to the diferrous metallocenter because crystals were unstable in solutions with both azide and sodium dithionite.

In the azide-bound diferric metallocenter, the azide moiety displaces the bridging oxo/hydroxo-ligand and binds in a μ-1,1 orientation (Fig. 6.4b) similar to toluene monooxygenase (Sazinsky *et al.* 2004) and a Y208F variant of the R2 subunit of ribonucleotide reductase (Andersson *et al.* 1999). There are also substantial changes in the coordination geometry of three glutamates. Most notably, $Glu_{\pi C}$ moves away from the Fe1 and is replaced by a terminally coordinating water molecule (HOH^{az}); this displacement of $Glu_{\pi C}$ can be understood in that its former position is sterically incompatible with the second nitrogen of the azide. $Glu_{\alpha A}$ undergoes an unusual carboxylate shift in which its Fe1-ligating oxygen changes from Oε1 to Oε2, and $Glu_{\pi A}$ becomes fixed in a single orientation. These alterations allow both $Glu_{\alpha A}$ and $Glu_{\pi A}$ to hydrogen bond HOH^{az}. Fe1 and Fe2 shift ~0.3 Å and 0.1 Å, respectively,

in the direction indicated by the difference map peaks (Fig. 6.4a). The geometry of Fe2 ligating residues remain unchanged.

The dithionite-reduced metallocenter adopts two conformers.

To determine the structure of the diferrous metallocenter, crystals were soaked in an oxygen-free environment containing sodium dithionite and methyl viologen. Structural changes happened only at the metallocenter, and the 1.25 Å resolution data were well accounted for using two distinct conformations of equal population (R/R_{free} = 0.115/0.141, Fig. 6.5a, Table 6.1). We will refer to these as conformations ***A*** and ***B***. In conformation ***A*** (Fig. 6.5b), the major change is an Fe1 shift of ~2 Å to position Fe1***A*** with the iron-iron distance increasing to 4.1 Å. $His_{\alpha B}$ becomes a ligand and the ligating interactions with $Glu_{\pi A}$, $Glu_{\pi C}$ and the bridging μ-oxo/hydroxo atom are lost. The resulting two open coordination sites are satisfied by a new water molecule very close the old position of Fe1 (HOH-***A***) and a change in $Glu_{\alpha A}$ coordination from monodentate to bidentate. This shift of Fe1 closely matches the redox dependent shift observed in rubrerythrins (Fig. 6.5c), however, in the those structures 100% of Fe1 undergoes the shift (Jin *et al.* 2002; Iyer *et al.* 2005; Dillard *et al.* 2011).

In conformation ***B*** (Fig. 6.5c), Fe1 remains in the same position as it was in the oxidized structure, but its ligation changes. Both residues $Glu_{\pi A}$ and $Glu_{\pi C}$ remain as ligands but alter their conformation and ligating geometry from the oxidized structure. Because there is no electron density for the positions that $Glu_{\pi A}$ and $Glu_{\pi C}$ occupied in the reference oxidized structure we can conclude that this ***B*** conformation is not simply due to a portion of the crystal that has remained in the diferric state. Also, trace metal analyses showed symerythrin contains no metals other than iron so that the 50% of Fe1 that did not move also cannot be due to a different kind of redox insensitive metal.

Internal symmetry of symerythrin.

A sequence alignment of the N- and C-terminal halves of the four-helix bundle of symerythrin demonstrates a 32% internal sequence identity (Cooley *et al.* 2010). The structure of symerythrin shows that, at the level of the folded protein, the internal symmetry is even more striking. First, the backbones of the two halves of the core four-helix bundle overlay to within 0.77 Å (Fig. 6.6a), a much better internal symmetry than any other structurally characterized FLSF member (Table 6.2). Second, the π-helical bulges in helices A and C are perfectly symmetric in their placement. Third, the metallocenter is exquisitely symmetric as 102 atoms (51 atom pairs) comprising the irons and their ligands overlay to within 0.71 Å in oxidized symerythrin and to within 0.61 Å in dithionite-reduced symerythrin (Fig. 6.6b).

Exploratory studies of enzymatic activity.

The auto-generation of the Val-Phe crosslink near the active site of symerythrin proves its metallocenter has notable catalytic potential,(Cooley *et al.* 2011) yet the physiologic function of symerythrin remains unknown. The strong structural similarities of the symerythrin metallocenter with that of rubrerythrins makes it plausible that its activity is related. As rubrerythrin is known to have strong peroxidase and much weaker oxidase activity (Riebe *et al.* 2009; Dillard *et al.* 2011), we tested symerythrin for such activities. In the absence of a known physiological reductant for symerythrin, for initial tests we exposed stoichiometrically reduced symerythrin to O_2 and monitored its return to the diferric state via recovery of its characteristic absorption band at 350 nm (Fig. 6.7a). This process of chemically reducing symerythrin followed by oxidation with O_2 could be repeated multiple times. Similarly, recovery of the 350 nm absorption band occurred upon exposure to hydrogen peroxide (Fig. 6.7b). Crosslinked and non-crosslinked symerythrin behaved equivalently in these experiments demonstrating that both forms have both catalytic oxidase- and peroxidase-like activities.

Discussion

Since its amino acid sequence was first described sixteen years ago (Stirewalt *et al.* 1995), symerythrin was expected to have a rubrerythrin-like diiron center involving seven ligating residues. The structures we report here confirm our prediction (Cooley *et al.* 2010) that the symerythrin metallocenter is distinct from rubrerythrins in that it includes as an eighth metal ligating residue, a glutamate that comes from a π-helical segment in helix A of the four-helix bundle.

The metallocenter structure. With crystallographic data to 1.20 Å resolution, the structure of diferric symerythrin provides the highest resolution description of a carboxylate bridged dinuclear center in the FLSF. The high resolution was crucial in allowing us to sort out details of partially occupied conformations associated with each of the structures and especially the azide-bound and dithionite-reduced metallocenters. We suspect that much of this heterogeneity is due to the crystals being at pH 5.5, which compared to a more physiological pH near 7, would alter the protonation state of a substantial portion of any ligands with pK_a values in the range of 5 to 7.5. Although our analyses are not at a sufficient resolution to visualize H-atoms, we can make some inferences of protonation states based on the interactions present, which help in the interpretation of the structures.

First, in the diferric structure we infer that the bridging oxygen is a μ-hydroxo-bridge. This accounts for bond lengths to the irons being 2.1-2.2 Å, compared with values of ~1.8 Å expected for a μ-oxo-bridge (Kurtz 1990), and for the close (2.6 Å) H-bonding interaction with $Glu_{\pi C}$ for which we assume the Glu is the acceptor (Fig. 6.3a). Interestingly, one of the two conformations seen for $Glu_{\pi A}$ H-bonds to the bridging hydroxo, so for 50% of the molecules, $Glu_{\pi A}$ itself is protonated. Assuming these protonation states, the ligands sum to a net charge of -6 or -7 (for the two populations differing in the protonation state of $Glu_{\pi A}$), which matches reasonably well with the +6 charge of the two ferric ions. Finally, an H-bond interaction between $His_{\alpha B}$ and $Glu_{\alpha A}$ is consistent with the His being in the positively charged imidazolium state.

For the azide-bound diferric complex (Fig. 6.4b), the interactions of HOH^{az} with the metallocenter allow us to suggest that it is a water donating H-bonds to $Glu_{\alpha A}$ and $Glu_{\pi A}$, which are both deprotonated. Assuming these protonation states, with the swapping out of $Glu_{\pi C}$ ligation for HOH^{az} and with $Glu_{\pi A}$ always being deprotonated, the ligands in this complex sum to a net charge of -6, matching the +6 charge of the two ferric ions. That the azide-bound diferric structure is a mix of unliganded and liganded forms is not surprising given that the soak concentration (250 mM) is not much higher than the azide K_d of 50 mM determined at neutral pH.

In both conformations of the dithionite-reduced structure, $Glu_{\pi C}$ has moved so it points at $Glu_{\pi A}$ placing their carboxylate Oε2-atoms just 2.6 Å apart (Fig. 6.5). A proton must be between these atoms, and we assign $Glu_{\pi C}$ as protonated because its Cδ-Oε2 bond length refined to a longer lengths (1.29 vs 1.25 Å) and has a less polar environment. In conformation ***A*** (with Fe1 shifted to position Fe1***A***; Fig. 6.5b), the two terminal oxygens can be identified as water molecules based both on the longer coordination distances of ~2.3 Å (2.26 – 2.36 Å) and on their network donating two H-bonds each to the Oε1-atoms of $Glu_{\pi A}$ and $Glu_{\pi C}$. With these assignments, the conformation ***A*** ligands sum to a charge of -4 exactly matching the charge of the diferrous center. For conformation ***B*** (without the Fe1 shifted) little has changed from the diferric state, except $Glu_{\pi C}$ has become protonated. If the μ-oxo-bridge also became protonated to an aqua-bridge that would help the charge balance, but there is nothing in the structure to support that assignment. As no other rubrerythrin family member has adopted such a diferrous structure, we suggest that this conformation is not of catalytic interest but is an artifact caused by the low pH of the crystal, which leads to $His_{\alpha B}$ being protonated and unable to serve as a metal ligand in half of the molecules in the crystal. Whereas a crystal structure at pH 7 would be desirable, the current crystal form disintegrated rapidly when placed in such buffers.

Taken together, and discounting the heterogeneities that appear to be associated with the low pH of the crystals, the structures of these three forms of symerythrin reveal that from a structural perspective the metallocenter is remarkably

similar to that of rubrerythrin family proteins (Figs. 6.3b and 6.5c). In addition to providing additional, similar views of the diferric and diferrous structures, the azide bound diferric complex provides the first such complex for the rubrerythrin family. It directly shows that the μ-hydroxo ligand is labile and so could be displaced by substrates, products or reaction-intermediates.

Considerations of symerythrin function. In thinking about symerythrin's function, we have little to go on aside from the structural information and the *in vitro* measurements showing that both oxygen and hydrogen peroxide can be reduced by symerythrin. From a structural viewpoint, most important to consider is the diferrous structure which is both the form from which a catalytic cycle begins and the form that would have predominated in the anaerobic world at the time of the origin of the rubrerythrin family. Compared to the rubrerythrins (Fig. 6.5c), the symerythrin diferrous structure is a remarkable match, having equivalent coordination geometry and a very similar active site pocket, simply having one more glutamate, $Glu_{\pi A}$, present making the active site pocket more polar and more negatively charged. The extra $Glu_{\pi A}$ would also serve to make the interactions with hydrogen peroxide more symmetric. In this light, the symerythrin active site is perfectly compatible with the peroxidase mechanism proposed for rubrerythrins where the peroxide binds the diferrous enzyme in place of the two terminal water ligands (Jin *et al.* 2002; Dillard *et al.* 2011).

Although inspection of the genomic context of symerythrin in the genomes of *C. paradoxa* and *G. violaceus* did not reveal additional clues about its physiologic function or interacting partner molecules, a reasonable working hypothesis is that symerythrin will also have peroxidase or closely related activity. As is observed for rubrerythrin, it seems non-polar substrates like oxygen and negatively charged substrates like superoxide would be discriminated against by this active site environment. The extra negative charge density could also provide stabilization for the previously proposed (Cooley *et al.* 2011) high-valent Fe(IV)-oxygen intermediates involved in the formation of the Val127-Phe17 crosslink.

Finally, we note that in these structural studies, we have only been able to visualize symerythrin forms containing the Val-Phe crosslink, but we do not yet know whether symerythrin in its natural setting contains the crosslink. Indeed, if the protein is in a microaerobic environment within the cells, the crosslink may not exist since its formation is dependent on oxygen (Cooley *et al.* 2011). Both the preferential crystallization of the crosslinked form and the observed stabilization it provides the metallocenter (Cooley *et al.* 2011) show that the crosslinked form is more conformationally rigid. However, the crosslinked and non-crosslinked proteins behave similarly in all of the spectroscopic and activity studies, which we take as evidence that the structures reported here are also relevant for non-crosslinked symerythrin. Supporting this conclusion is the observation that the symerythrin metallocenter structures agree well with other rubrerythrin family members, none of which have crosslinks. We are now working to culture these phototrophs in order to identify symerythrin's presence, location, and crosslink status in its natural setting, as well as other properties such as its metal content, its interacting protein partners, and how its expression responds to different environmental conditions.

Insights into the evolution of the rubrerythrins and the FLSF. As first described by Andrews (Andrews 1998), rubrerythrin family members (and especially symerythrin[†]) have a notable level of internal two-fold sequence similarity at the subdomain level which when seen for proteins is considered an ancestral feature reflecting that the protein fold originated through the duplication and fusion of a gene encoding a half-protein that assembled into a symmetric dimer (e.g. (Dauter *et al.* 1997; Lang *et al.* 2000)). For the rubrerythrin family, this putative ancestor would have consisted of a two-helix peptide having metal binding residues in the appropriate positions, and given the ubiquitous distribution of rubrerythrins in anaerobic organisms (Gomes *et al.* 2001), the fold would seem to have originated before the advent of oxygen in the atmosphere. One plausible scenario for the origin of the family is that the ancestral form of the rubrerythrin-symerythrin families had a rubrerythrin-like metallocenter and the symerythrins were formed via the insertion of

an eighth metal-ligating residue in helix A. Given the very narrow phylogenetic distribution of the 8-residue metallocenter variant compared with the ubiquitous presence of the 7-residue metallocenter (Gomes *et al.* 2001), this scenario would seem attractive.

However, an alternative scenario that we proposed previously (Cooley *et al.* 2010), is that the seven residue metallocenter version was formed by the deletion of the $Glu_{\pi A}$ metal-ligating residue from a symerythrin-like ancestor. Now, along with the greater internal sequence symmetry of symerythrin compared to rubrerythrins, additional evidence for the ancestral form being symerythrin-like comes from the much higher internal backbone symmetry of symerythrin (Table 6.2, Fig. 6.6a), the additional symmetry associated with the two π-helices, and the near perfect internal symmetry of the metallocenter itself (Fig. 6.6b). By Dollo parsimony (Farris 1977), the principle that complex features are easier to lose than gain during evolution, the remarkable internal similarities of symerythrin make a compelling case that its eight-ligand motif is an ancestral trait of the family. The narrow phylogenetic distribution of symerythrin would be understandable if the seven-residue version of the metallocenter had improved functionality allowing it to outcompete the ancestral form over time. Interesting in this light is that the two organisms in which symerythrin has been preserved are not anaerobes, but are both aerobic organisms that are among the early organisms that carried out oxygenic photosynthesis. They have both been referred to as "living fossils," a term coined by Darwin to refer to organisms that have changed little over time because their niche remains constant and have experienced little competition in that niche. Based on this concept, the relatively slow evolution of the rubrerythrin family proteins such that residual internal symmetry from their origin has been maintained suggests that their function has changed little over time.

So if the seven-residue metallocenter was derived from an eight-residue metallocenter, how did the more predominant six-residue metallocenter of the FLSF arise? One possible scenario is the six-residue metallocenter splintered off from a rubrerythrin-like ancestor via the deletion of the seventh metal-ligating residue.

Alternatively, the six-residue metallocenter could have formed independently (one or more times) via a distinct gene-duplication and fusion event of a variant of the dimerizing two-helix ancestral peptide that had three rather than four metal-ligating residues. Sequence comparisons lend support to such a model. In particular, there exists one FLSF family with a six-ligand metallocenter for which family members have residual internal sequence identity at a level similar to that of the rubrerythrins (~30% identity), and for which a crystal structure shows high internal structural similarity (RMSD of 1.1 Å, Table 6.2). Because of their similarly high internal sequence similarity, Andrews (Andrews 2010) had considered these proteins to be closely grouped with the rubrerythrins, naming them 'erythrins' because they do not have a rubredoxin domain.

However, a remarkable finding not noted by Andrews (Andrews 2010) is that these erythrins are actually not very sequence similar to the rubrerythrin family (~16% identity). In fact, aside from the four metallocenter-associated residues in the half-protein (2 Glu, 1 His, and 1 Tyr), there are 14 positions showing strong internal sequence symmetry in the erythrins and 15 positions showing strong internal symmetry in the rubrerythrins but only one match between the two (Fig. 6.8a). This very low congruence approximates what would be expected for two independent sequences and is not compatible with a model in which the internal sequence similarities are conserved from a common ancestor. Thus, this provides very strong evidence for independent origins for these two families resulting from the fusion of distinct two-helix peptides (Fig. 6.8c). Further support for this conclusion is another feature symmetric within each family, but differing between them: the loops connecting the A to B (or C to D) helices in erythrins are three residues shorter than those in the rubrerythrin family (Fig. 6.8a,b).

These observations support the hypothesis that the single-chain FLSF fold originated at least twice via independent gene duplication/fusion events of what may have been homologous two-helix peptides (Fig. 6.8c). Evidence that this process of FLSF fold generation may continue to this day comes from a modern-day family of

proteins that are two-helix peptides that assemble into the topology of the FLSF four-helix bundle containing the key residues of the canonical FLSF six-residue metallocenter. These proteins have not been functionally characterized, but are known from the crystal structure of a protein from *Nitrosomonas europaea* (PDB 3k6c). This two-helix peptide has a loop length matching the erythrins rather than the rubrerythrins (Fig. 6.8a).

As for the more divergent FLSF proteins, these proteins have no notable internal sequence symmetry beyond the metal-ligating residues themselves (Table 6.2). We would suggest that for these other FLSF superfamily members there is insufficient evidence at this time to conclude whether they derived from the rubrerythrin line, the erythrin line or from one or more additional independent gene-fusion events of homo- or hetero-dimers of two-helix peptides having the appropriate ligands. In any case, the strong evidence for independent gene-fusion events giving rise to the rubrerythrin and erythrin families of the FLSF implies that our earlier proposal that the whole FLSF derived from a symerythrin-like ancestor (Cooley *et al.* 2010) was an oversimplification, and that previous assumptions of a single origin for all FLSF families (Andrews 1998; Andrews 2010) must be revisited. Nevertheless, the insights generated through the characterization of symerythrin presented here underscore the value of using π-helices as guides to discover structure-function and evolutionary relationships in proteins.

Materials and Methods

Expression and purification of symerythrin from *Cyanophora paradoxa*

*Development of a recombinant expression system***.** ORF180 from the cyanelle genome of *C. paradoxa* (NCBI Gene ID: 801647) was purchased from GenScript (Piscataway, NJ) in the pUC57 cloning vector with NdeI and BamHI restriction sites added to the 5' and 3' ends of the gene, respectively. This plasmid was transformed into chemically competent *E. coli* DH5α cells. After cell growth in the presence of

100 µg/ml ampicillin, plasmid DNA was isolated and digested with NdeI and BamHI. The resulting 0.5 kb insert encoding the symerythrin gene was gel purified using the QIAquick Gel Extraction Kit (Qiagen) and ligated to the NdeI/BamHI-digested pT7-7 expression vector. The resulting plasmid, pT7CPERY, was transformed by electroporation into *E. coli* BL21(DE3) (Novagen) cells for subsequent expression and purification.

*Expression and purification of symerythrin***.** *E. coli* BL21(DE3) cells containing the pT7CPERY plasmid were grown in 4 L of LB media containing 100 µg/ml Ampicilin at 37 ^{o}C to OD_{600} 0.6 – 0.8. Expression of symerythrin was then induced with 1 mM isopropyl β-D-1-thiogalactopyranoside, at which time 150 µM $FeSO_4$ was also added. After 5 h, cells were harvested by centrifugation at 5000 **g** for 15 minutes and frozen at -70 ^{o}C. Yields of 3-4 g cell paste per liter culture were typical.

For protein purification, 12-14 g cell paste was suspended in 25 mM MOPS, 5% glycerol and 50 mM NaCl pH 7.2 to a total volume of 40 ml. Cells were lysed by two or three passes through a French press at 52,000 Pa, and then centrifuged at 10,000 **g** for 10 min. The supernatant was carefully decanted, diluted to 60 ml and centrifuged at 150,000 **g** for 1 h. The resulting supernatant was decanted, adjusted to pH 7.2. This cell free extract was loaded at 2.0 ml/min onto a DEAE Sepharose FF column (2.5 cm x 17 cm) pre-equilibrated with 25 mM MOPS, 5% glycerol and 50 mM NaCl pH 7.2. The column was washed with five column volumes (400 ml) of the same buffer at a flow rate of 1.5 ml/min, after which a linear 400 ml gradient from 0 to 0.5 M NaCl was applied. Each 10 ml fraction was analyzed by SDS-PAGE: non-crosslinked symerythrin migrated with an apparent molecular weight of 20 kDa, while crosslinked symerythrin migrated with an apparent molecular weight of 17 kDa. Both crosslinked and non-crosslinked symerythrin eluted between 0.1 and 0.15 M NaCl (pool I), while enriched non-crosslinked symerythrin eluted between 0.15 and 0.2 M NaCl (pool II).

*Purification of non-crosslinked symerythrin***.** Fractions from pool II of the anion exchange fractionation were concentrated to 2.5 ml by ultrafiltration and applied to a Superdex-75 gel filtration column (2.5 cm x 55 cm) equilibrated with 25 mM MOPS, 5% glycerol and 150 mM NaCl pH 7.2 at a flow rate of 0.25 ml/min. The purity of the resulting fractions was analyzed by SDS-PAGE. Fractions containing pure non-crosslinked symerythrin were pooled, repeatedly exchanged into 10 mM Tris pH 7.4 by ultrafiltration, and concentrated to ~30 mg/ml before being flash frozen in liquid N_2. From 4 L of culture, 15-20 mg of non-crosslinked symerythrin was typically obtained.

*Purification of crosslinked symerythrin***.** Fractions from pool I of the anion exchange chromatography were concentrated to 2.5 ml by ultrafiltration and applied to the same Seperdex-75 column described above. Both non-crosslinked and crosslinked symerythrin eluted simultaneously but separate from other contaminants. The pooled fractions was adjusted to 1 M NH_4Cl from a 4 M stock to ensure proper binding to a Phenyl Sepharose CL4B column (2 cm x 90 cm). After application to the column, and washing with two column volumes of buffer (25 mM MOPS and 175 mM NH_4Cl, pH 7.2), the protein was eluted with a 750 ml (five column volume) linear gradient from 0-50% ethanol in 25 mM MOPS, pH 7.2. Crosslinked symerythrin eluted at ~35-40% ethanol while the non-crosslinked form eluted at ~40-45% ethanol. Fractions containing >95% crosslinked symerythrin were pooled together, exchanged into 10 mM Tris pH 7.4 and concentrated to ~15 mg/ml prior to flash freezing in liquid N_2. From 4 L of culture, ~2-3 mg of crosslinked symerythrin was obtained.

Biochemical analyses of purified symerythrin.

Protein concentrations were determined by BCA assays (Pierce) using bovine serum albumin as a standard. Analytical gel filtration and dynamic light scattering were used to determine the quaternary structure of crosslinked and non-crosslinked symerythrin at concentrations ~5 mg/ml. In-solution proteolytic digests of crosslinked

and non-crosslinked forms were conducted and analyzed by mass spectrometry as previously described.(Cooley *et al.* 2011)

Trace metal analysis. Iron content was typically determined by the ferrozine colorimetric assay (Percival 1991). For determining the content of iron, manganese and zinc by Inductively coupled plasma optical emission spectroscopy (ICP-OES), 10 μL of 70% trace metal grade HNO_3 was added to an equal volume of 15 mg/ml symerythrin. The mixture was incubated at 60 ^{o}C for 3 h and then diluted to 5 ml in metal-free water for subsequent analysis.

*Metal reconstitution***.** Purified non-crosslinked symerythrin was incubated under argon with 1 mM 1,10-phenathronline and 5 mM EDTA at 2 mg/ml in 25 mM MOPS pH 7.2, 150 mM NaCl, 5% glycerol and 1 mM dithionite. After 3 h at room temperature, the resulting metal-free symerythrin was repeatedly exchanged into 10 mM Tris pH 7.4, concentrated to ~20 mg/ml and flash frozen in liquid nitrogen.

To prepare Fe-reconstituted non-crosslinked symerythrin, metal-free protein was diluted to 2 mg/ml (100 μM) in a sealed vial containing an anaerobic solution of 25 mM MOPS pH 7.2, 150 mM NaCl and 5% glycerol. Anaerobically prepared 20 mM stock solutions of sodium dithionite and $Fe(NH_4)_2(SO_4)_2$ were added to this solution to final concentrations of 2 mM and 1 mM, respectively. All manipulations of anaerobic solutions were carried out using gas-tight Hamilton syringes. After 3 h at room temperature under argon, excess dithionite and iron were removed by dialysis. Typically, 1.7 to 1.8 Fe per active site were obtained from this procedure.

UV/Visible spectroscopic characterization. Identical procedures were used for non-crosslinked and crosslinked symerythrin. Iron reconstituted, symerythrin was diluted to 1 mg/ml into an anaerobically prepared solution of 25 mM MOPS pH 7.2, 150 mM NaCl and 50 μM methyl viologen. This solution was transferred to a 1 cm path length quartz cuvette purged with argon. Spectra were recorded at room temperature using a Beckman DU-640 spectrophotometer. Sodium dithionite was then added from an anaerobically prepared 1 mM stock solution and incubated for 3 minutes, at which time the spectrum was recorded. Subsequent dithionite additions

were repeated until absorption peaks characteristic of reduced methyl viologen were observed, indicating all symerythrin metallocenters were reduced. At this time, the cuvette was exposed to either O_2 by removing the septa on the cuvette or H_2O_2 by stepwise additions of an anaerobically prepared 1 mM hydrogen peroxide stock solution. Azide-bound spectra were recorded after aerobically incubating a 1 mg/ml symerythrin solution with various quantities of sodium azide from a 5 M stock solution in 25 mM MOPS, 150 mM NaCl, pH 7.2.

Consensus sequences for rubrerythrin- and erythrin-like sequences

Non-redundant sequences of rubrerythrin family members were obtained from the NCBI database by performing a PSI-BLAST search using *D. vulgaris* rubrerythrin (PDB 1lkm). Using a 50% sequence identity threshold, 250 rubrerythrin sequences were obtained and aligned via the CLUSTALW algorithm. The consensus sequence of these aligned sequences was derived using a SeqLogo plot (Schneider and Stephens 1990) generated by the WebLogo server (Crooks *et al.* 2004). Residues in each position were considered highly, moderately, or poorly conserved if the identity of the top amino acid had a bit value of at least 2/3, between 2/3 and 1/3, and less than 1/3 the maximum bit value, respectively. A similar PSI-BLAST search was performed using the erythrin sequence from *Pyrococcus furiosus* (PDB 2fzf) for which 11 sequences matched the same 50% sequence identity threshold used for rubrerythrins. For the oligomerizing $\alpha\alpha$-hairpins, the sequence of PDB 3k6c was used for the PSI-BLAST search. This resulted in 80 sequences with at least 50% sequence identity. Alignments and consensus sequences for these protein families were obtained as for rubrerythrins. For the consensus sequence of symerythrins, residues were considered highly conserved if the residues were identical in the two symerythrin sequences. For the internal similarity assessment, a residue was considered moderately conserved if it matched an internally symmetric highly conserved residue.

Crystallization and Structure Determination

Crystallization and crystal handling. Crystals were grown at 6 ^{o}C in hanging drops on siliconized cover slips from 2.5 µl of purified crosslinked symerythrin (8 mg/ml in 10 mM Tris pH 7.4) and 1 µl of reservoir solution consisting of 22-28% PEG 3350, 0.1 M Bis-Tris pH 5.5 and 0.35-0.4 M NH_4Cl. Diffraction quality crystals (400 x 75 x 75 µm) were observed within 4-10 days. No crystals grew when non-crosslinked protein was used, and irregular crystal growth was observed when heterogeneous mixtures of crosslinked and non-crosslinked symerythrin were used. Purification of crosslinked symerythrin was therefore essential for obtaining diffraction quality crystals under these conditions. For data collection of all forms, single crystals were suspended on a rayon loop, briefly dipped into paratone oil and immediately frozen in liquid N_2.

To form the azide-bound diferric complex, crystals were incubated for 20 min in an artificial mother liquor (AML) of 28% PEG 3350, 0.37 M NH_4Cl and 0.1 M BisTris pH 5.5 supplemented with 250 mM sodium azide. Attempts to obtain greater occupancies of azide binding were not successful as the use of higher concentrations of sodium azide or longer incubation periods resulted in the disintegration of the crystals.

For preparing the dithionite-reduced form, crystals before freezing were incubated for 45 min under argon in anaerobically prepared solution of 5 mM dithionite and 0.5 mM methylviologen in AML. The blue color of reduced methyl viologen was present during the course of the incubation, confirming the solution remained anaerobic. To ensure the results were not due to partial or incomplete reduction, we varied the dithionite concentration from 1 mM to 20 mM and the length of incubation from 20 min to 3 h; no differences were found between these treatments. To address the possibility that the paratone oil in which the crystals were frozen contained small quantities of oxygen that could affect the structure of the reduced metallocenter, crystals were soaked as above except with 15% PEG400 added as a cryoprotectant so that after the 45 min incubation, crystals could be directly frozen in liquid N_2. The results were unchanged.

Data Collection and structure determination. Data were collected using beamline 5.0.3 and 5.0.1 at the Advanced Light Source (Lawrence Berkeley National Laboratory). Data were processed and scaled using iMosflm v1.0.040 (Leslie 1992) and SCALA (Evans 1997), and 5% of the data were flagged for use in R_{free}. Statistics are in Table 6.1.

The structure of differic symerythrin was solved by molecular replacement using the Phenix software package (Adams *et al.* 2010). Rubrerythrin (PDB 1lkm) without the rubredoxin domain, iron atoms and water molecules was used as the initial search model and a solution was obtained with spacegroup $P3_2$ and 2 molecules in the asymmetric unit (initial R/R_{free} = 0.294/0.353). In each monomer, two $2F_o$-F_c peaks of >20 ρ_{rms} were observed in a similar position as the dinuclear center of rubrerythrin. Modeling two iron atoms in these peaks dropped the R/R_{free} to 0.288/0.323, after which the sequence of symerythrin from residues 2-180 was built manually using Coot (Emsley *et al.* 2010) and refined with REFMAC (Murshudov *et al.* 1997). For this and all subsequent structures, anisotropic B-factors and riding hydrogens were used during refinements. Standard criteria were used for modeling water molecules (>1 ρ_{rms} intensity in the $2F_o - F_c$ map, >2.4 Å distance from nearest contact). Molprobity (Davis *et al.* 2007) was used to monitor the model geometry. For the final rounds of refinement, B-factor and geometry restraint weights were optimized. The final R/R_{free} of the diferric structure was 0.101/0.125 (Table 6.1).

For refinement of the dithionite-treated structure, the diferric structure (with water molecules) was used as the starting model except that both iron atoms and all eight side chains of the inner sphere iron-ligating residues from both chains were removed to minimize model bias (initial R/R_{free} = 0.194/0.205). Iron atoms and side chains of the metallocenter were then manually built into both chains. Water molecules were added or removed according the criteria described for the oxidized structure. As with the oxidized structure, interpretable electron density was observed for residues 2-180 in both chains. Final refinements were done as for the diferric state, leading to a final R/R_{free} of 0.115/0.141 (Table 6.1).

The azide bound structure (initial R/R_{free} = 0.238/0.286) was solved using the same process as the dithionite structure. The positions of both iron atoms and the side chains of $Glu_{\alpha B}$, $His_{\alpha B}$, $Glu_{\alpha C}$, $Glu_{\alpha D}$ and $His_{\alpha D}$ were not substantially different from the unliganded structure and were therefore modeled first (R/R_{free} = 0.137/0.175). $2F_o$-F_o density for the remaining ligating residues revealed a mixture of states consisting of the unliganded conformation and an azide-induced conformation. To facilitate building the azide and the azide-induced conformations of $Glu_{\alpha A}$, $Glu_{\pi A}$ and $Glu_{\pi C}$, an F_o(azide)-F_o(unliganded) difference map was generated using phases from the unliganded model (Fig. 6.4a). Modeling the occupancies of the unliganded and azide-bound conformations at 0.5 each gave consistent B-factors between the iron atoms and all ligands, including the azide molecule. With the slightly lower resolution of these data compared with the oxidized and dithionite-treated data, residues 2-4 had much weaker density and were not modeled. π-helices were confirmed by analysis with DSSP (Kabsch and Sander 1983) and π-HUNT (Cooley *et al.* 2010).

Accession Numbers

Coordinates and structure factors for crosslinked oxidized, dithionite-reduced and azide-bound diferric models have been deposited in the Protein Data Bank with accession numbers 3qhb, 3qhc and 3sid, respectively.

Acknowledgements

This work was supported in part by the General Medical Institute, NIH grant GM R01-13 083136 and the Oregon Agricultural Experiment Station.

Table 6.1. Data Collection and Refinement Statistics for Symerythrin

	Diferric	**Dithionite-Treated**	**Azide-bound diferric**
Data collection [a]			
Space group	$P3_2$	$P3_2$	$P3_2$
Unit cell axes (Å)	*a*=*b*=81.47, *c*=46.26	*a*=*b*=81.50 *c*=46.34	*a*=*b*=82.01 *c*=46.44
Resolution Limits (Å)	30.57-1.20 (1.25-1.20)	30.60-1.25 (1.31-1.25)	30.74-1.40 (1.48-1.40)
Unique Observations	107,444 (15,684)	95,243 (13,875)	68,857 (10,086)
Multiplicity	17.8 (10.7)	11.9 (10.4)	12.2 (12.0)
Average I/σ	23.4 (6.3)	15.0 (5.4)	18.9 (5.6)
R_{meas} [b] (%)	8.1 (35.2)	10.6 (42.6)	7.7 (46.1)
Refinement			
R_{cryst} / R_{free} (%)	10.1/12.5	11.5/14.1	11.5/15.6
No. protein molecules	2	2	2
No. protein residues	358	358	352
No. water molecules	555	555	562
Total number atoms	3534	3549	3543
rmsd bond angles (^{o})	3.6	2.8	2.1
rmsd bond lengths (Å)	0.019	0.022	0.026
<B> protein (Å^2)	12.9	14.7	16.2
<B> water (Å^2)	26.2	27.3	30.24
<B> Fe (Å^2)	6.73	10.1	10.0
Ramachandran Plot (%)[c]			
Favored	97.2	97.2	98.0
Outliers	0.0	0.0	0.0
PDB code	3qhb	3qhc	3sid

[a] Numbers in parentheses correspond to values in the highest resolution bin
[b] R_{meas} is the multiplicity-weighted merging *R*-factor (Diederichs and Karplus 1997)
[c] Ramachandran plot generated using Molprobity (Davis *et al.* 2007)

Table 6.2. Internal symmetry statistics for structure-based comparisons of the A-B with the C-D core-helix pairs from representatives of FLSF proteins. Alignments and the reported statistics were generated by the DALI server (Holm *et al.* 2008).

Protein	**PDB**	**No. aligned residues**	**% Sequence Identity**	**RMSD (Å)**	**Z-score**
Symerythrin	3qhb	57	32	0.77	10.6
Rubrerythrin	1lkm	58	29	1.5	9.8
Erythrin	2fzf	58	31	1.1	8.5
DPS	2d5k	59	7	2.0	8.9
Bacterioferritin	1bfr	62	16	1.3	9.3
Ferritin	1vlg	62	11	1.5	8.4
Δ^9-desaturase	2uw1	55	4	2.5	4.3
RNR Ib	1uzr	59	15	2.8	6.4
MMOH α-subunit	1mty:D	60	8	2.8	5.8
MMOH β-subunit	1mty:B	57	9	2.7	4.6

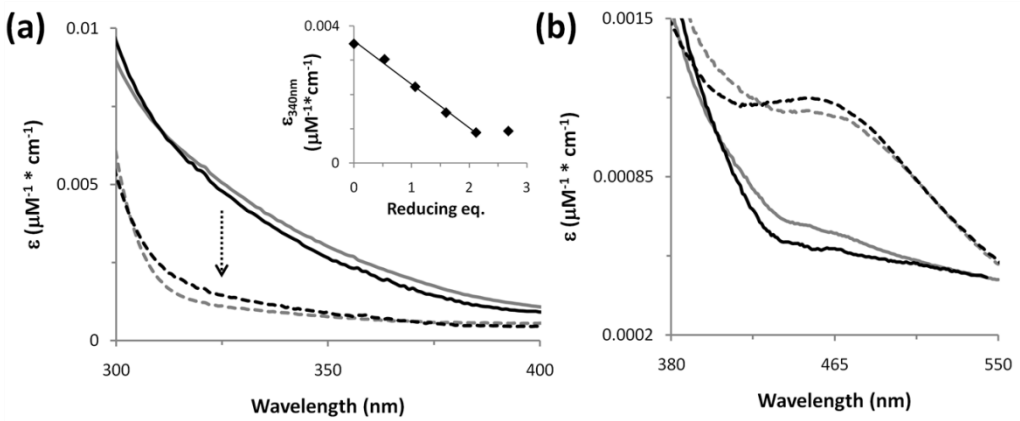


Figure 6.1. Spectra of oxidized, reduced and azide bound symerythrin. (a) Reduction (dotted arrow) converts spectra of diferric non-crosslinked (solid gray line) and crosslinked (solid black line) symerythrin to spectra of diferrous symerythrin (dashed gray and black lines, respectively). Inset: The change in absorption at 340 nm as a function of reducing equivalents during a dithionite titration of diferric, non-crosslinked symerythrin. Equivalent results were obtained for the crosslinked isoform (data not shown). (b) Spectra of diferric non-crosslinked and crosslinked symerythrin before (gray and black lines, respectively) and after (dashed gray and black lines, respectively) addition of 0.5 M sodium azide .

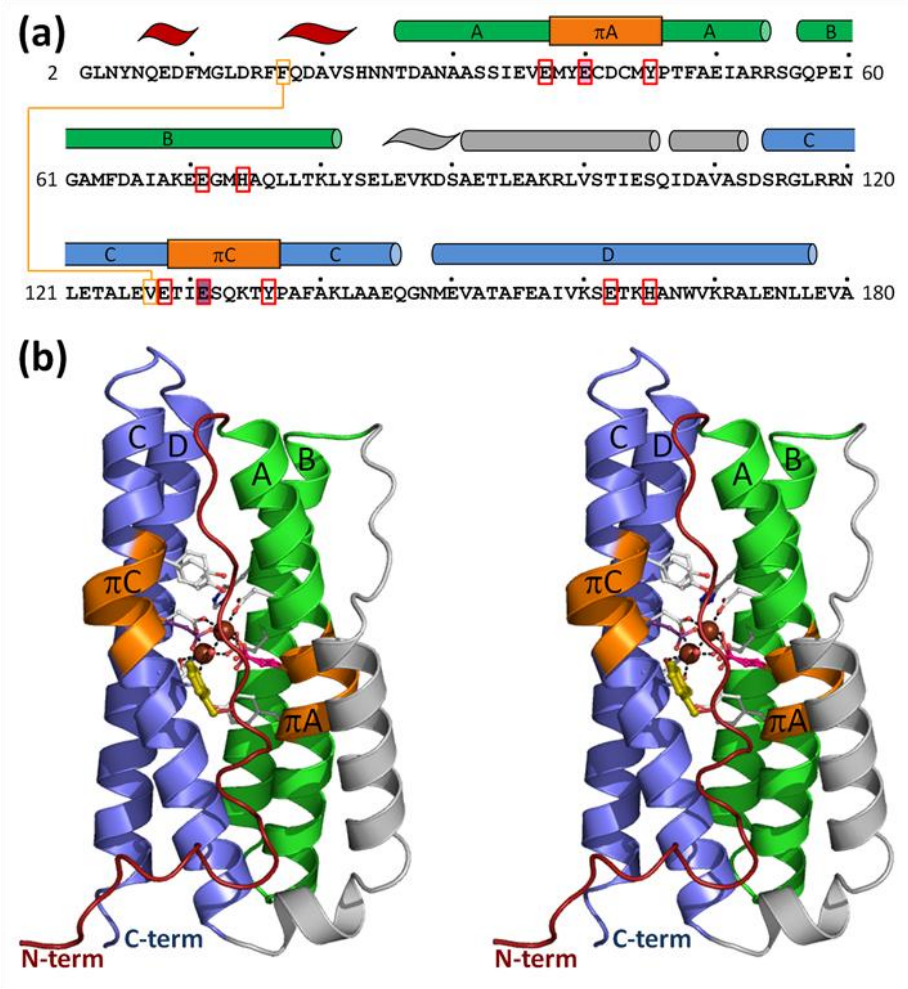


Figure 6.2. Structure of crosslinked diferric symerythrin. (a) The sequence of symerythrin with secondary structure segments annotated: P_{II} spirals (Hollingsworth and Karplus 2010) (tilda shapes), α-helices (cylinders – with A and B green, C and D blue, and linker helices grey), and π-helices (orange rectangles). Metal ligands (including the second-shell Tyr residues) are outlined in red with the seventh and eighth metal-ligating residues distinguished by purple and pink backgrounds, respectively. The Val127-Phe17 crosslink is indicated using yellow boxes and lines. Black dots are located above every tenth residue. (b) Stereoview ribbon diagram of symerythrin emphasizing the αα-units of helices A-B (green) and C-D (blue), the N-terminal tail (dark red), the linker between helices B and C (gray), and the π-helical segments πA and πC (orange). Iron atoms (brown), atoms in the Val127-Phe17 crosslink (yellow) and metallocenter residues (colored as in Fig. 6.3) are shown as balls-and-sticks. The N- and C-termini are labeled. Molecular graphics were prepared using PYMOL (Schrodinger 2010).

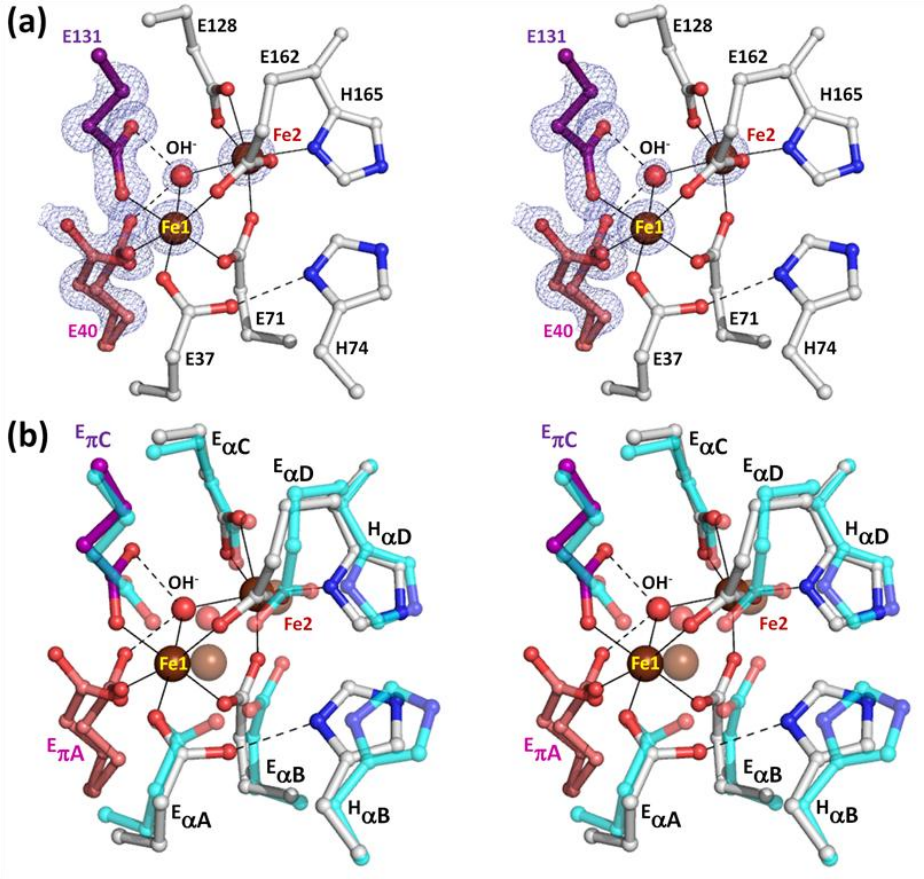


Figure 6.3. Stereoviews of the diferric metallocenter of symerythrin and its comparison with rubrerythrin. (a) Ball-and-stick model of the symerythrin metallocenter showing the six canonical metal-ligating residues of the FLSF (gray carbons), the seventh ligating residue characteristic of rubrerythrins (Glu131; purple carbons), and the eighth ligating residue unique to symerythrin (Glu40; pink carbons). Nitrogen and oxygen atoms are blue and red, respectively. $2F_o$-F_c density for the side chains of Glu131 and Glu40 is contoured at 2.0 ρ_{rms}, and that for Fe1 and Fe2 (brown spheres) and the μ-hydroxo bridge is contoured at 5.0 ρ_{rms}. Solid and dashed lines show metal-ligand and hydrogen bond interactions, respectively. (b) Overlay of the diferric metallocenter of symerythrin (as in panel A) with the diferric metallocenter of rubrerythrin from *D. vulgaris* (PDB 1lkm) shown semi-transparently with carbon colored cyan and other atoms colored as for symerythrin. In panel B and further figures, metallocenter residues are labeled according to the helix on which they reside, as described in the main text.

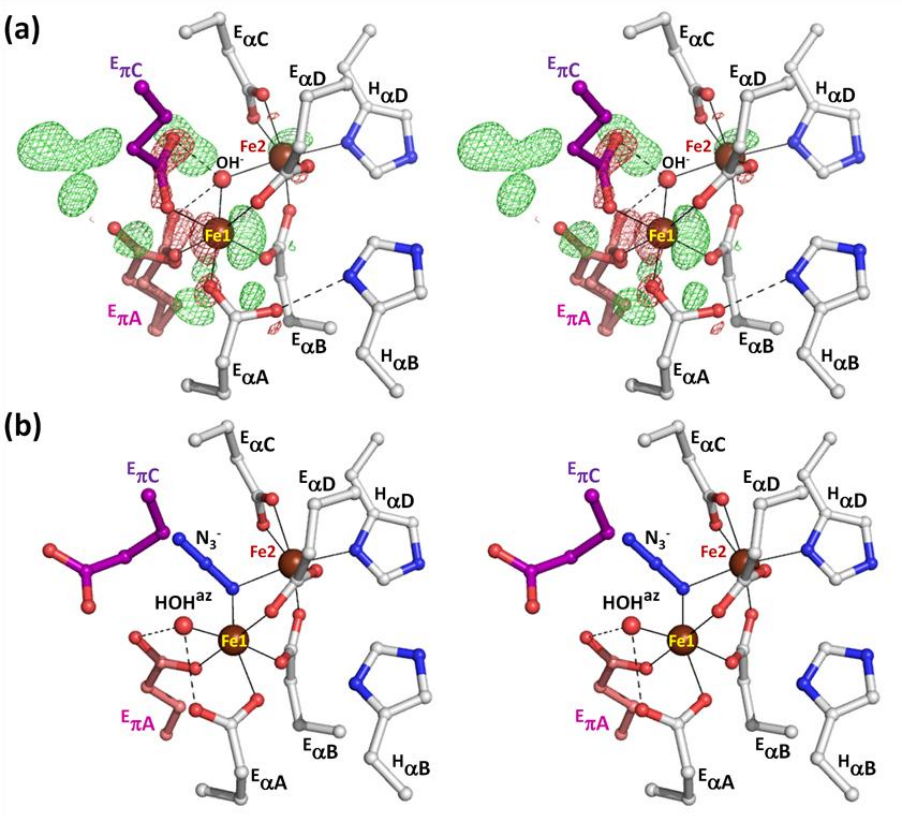
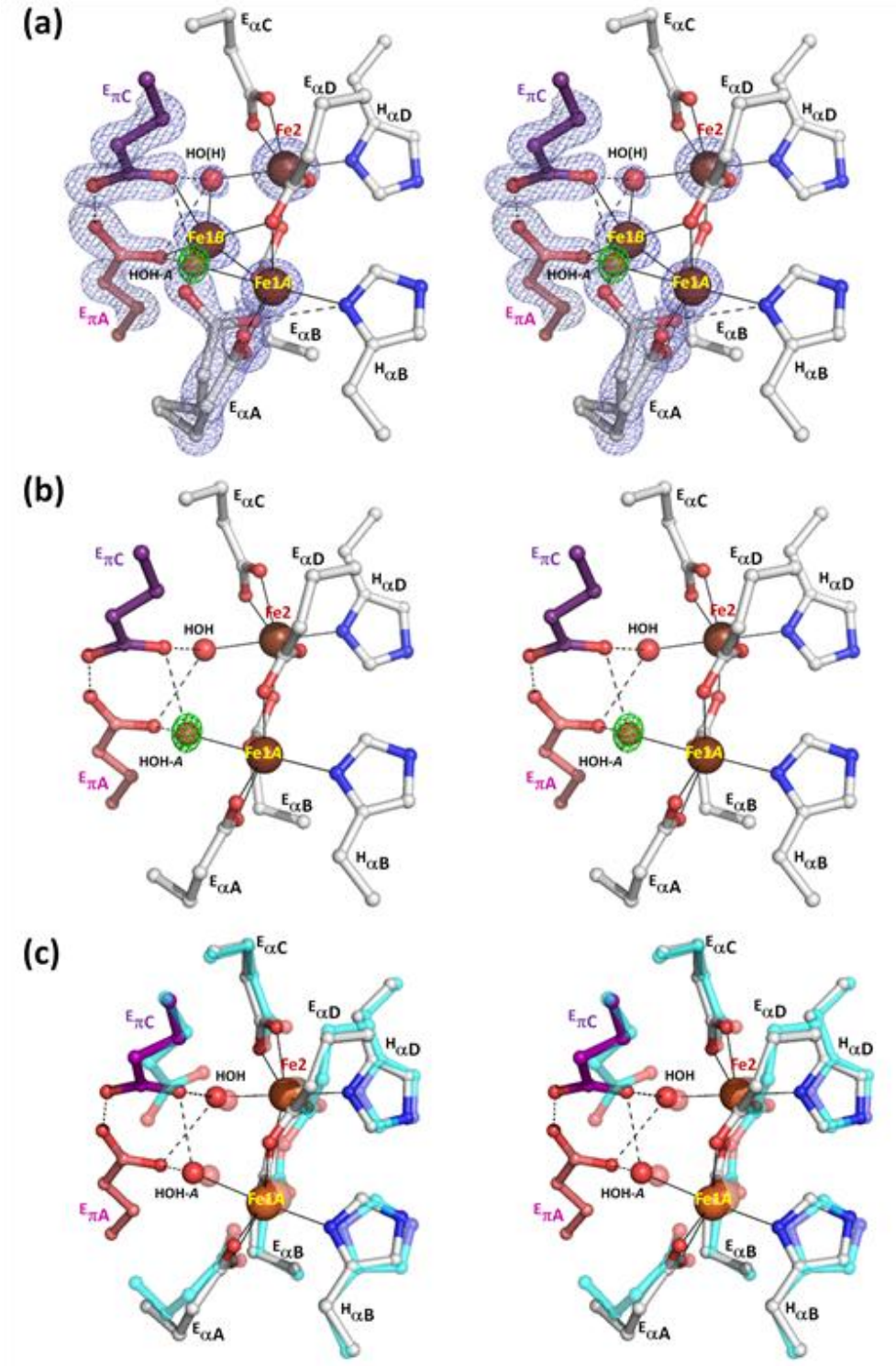


Figure 6.4. Stereoviews of the azide-bound diferric metallocenter of symerythrin. (a) The unliganded diferric metallocenter model overlayed on the F_o(azide) - F_o(unliganded) difference map contoured at +4.5 ρ_{rms} (green mesh) and -4.5 ρ_{rms} (red mesh). Atom coloring and residue labeling are as described in Fig. 6.3. (b) Model for azide-bound diferric metallocenter (seen at 0.5 occupancy) including the newly bound HOH^{az}. Lines and coloring are as in Fig. 6.3a.

Figure 6.5. Stereoview of the diferrous symerythrin metallocenter and its comparison with rubrerythrin. (a) The reduced symerythrin metallocenter modeled as two equally populated conformations, ***A*** and ***B***, each at 50% occupancy. $2F_o$-F_c density contoured at 2.0 ρ_{rms} (blue mesh) reveals single conformations for residues $E_{\pi A}$ and $E_{\pi C}$ and two conformations for $E_{\alpha A}$, and $2F_o$-F_c density contoured to 10.0 ρ_{rms} shows two positions for Fe1 (Fe1***A*** and Fe1***B***). An F_o-F_c omit map (omitting HOH-***A***) contoured to +4.0 ρ_{rms} (green mesh) shows evidence for modeling an Fe1***A***-ligating water molecule (labeled HOH-***A*** in panel B). Lines and atom coloring are as in Fig. 6.3a. (b) The reduced conformation ***A*** including the omit map density for HOH-***A*** described in panel A. (c) Overlay of conformation ***A*** of diferrous symerythrin with the diferrous metallocenter of rubrerythrin (transparent atoms with same coloring as Fig. 6.3b; PDB 1lko).

Figure 6.5. (continued)



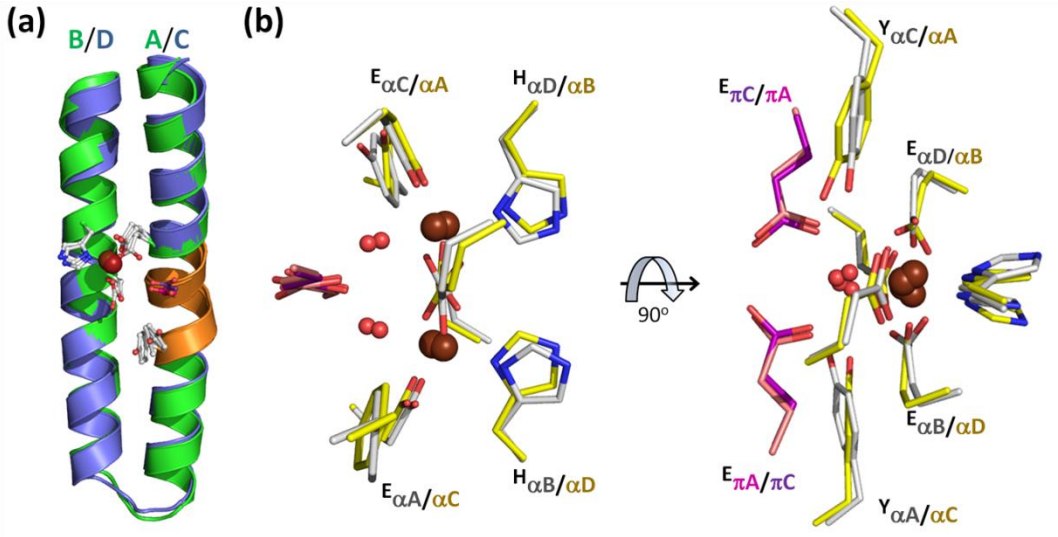


Figure 6.6. Internal symmetry in diferrous symerythrin. (a) Overlay of helix pair A-B (green) onto C-D (blue) highlighting iron-ligating residues (sticks), the two π-helices (orange) and the two iron atoms (brown and light brown spheres). (b) Orthogonal views (direction of rotation indicated) of an overlay of conformation ***A*** of the diferrous symerythrin metallocenter (coloring is the same as Fig. 6.3) with itself (yellow carbon atoms), based on the superimposing of helix pairs A-B and C-D onto helix pairs C-D and A-B, respectively. Direct metal ligands and the second-shell Tyr residues are shown and labeled in the panel in which they are best seen.

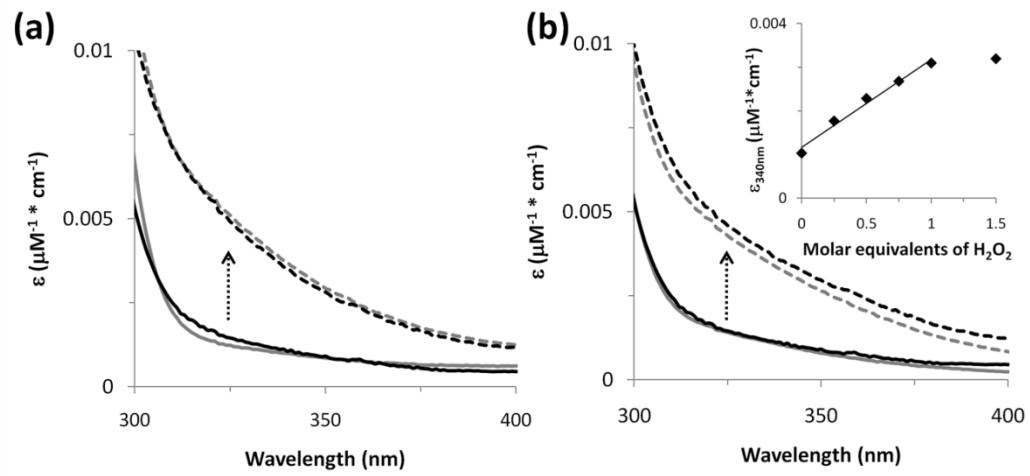
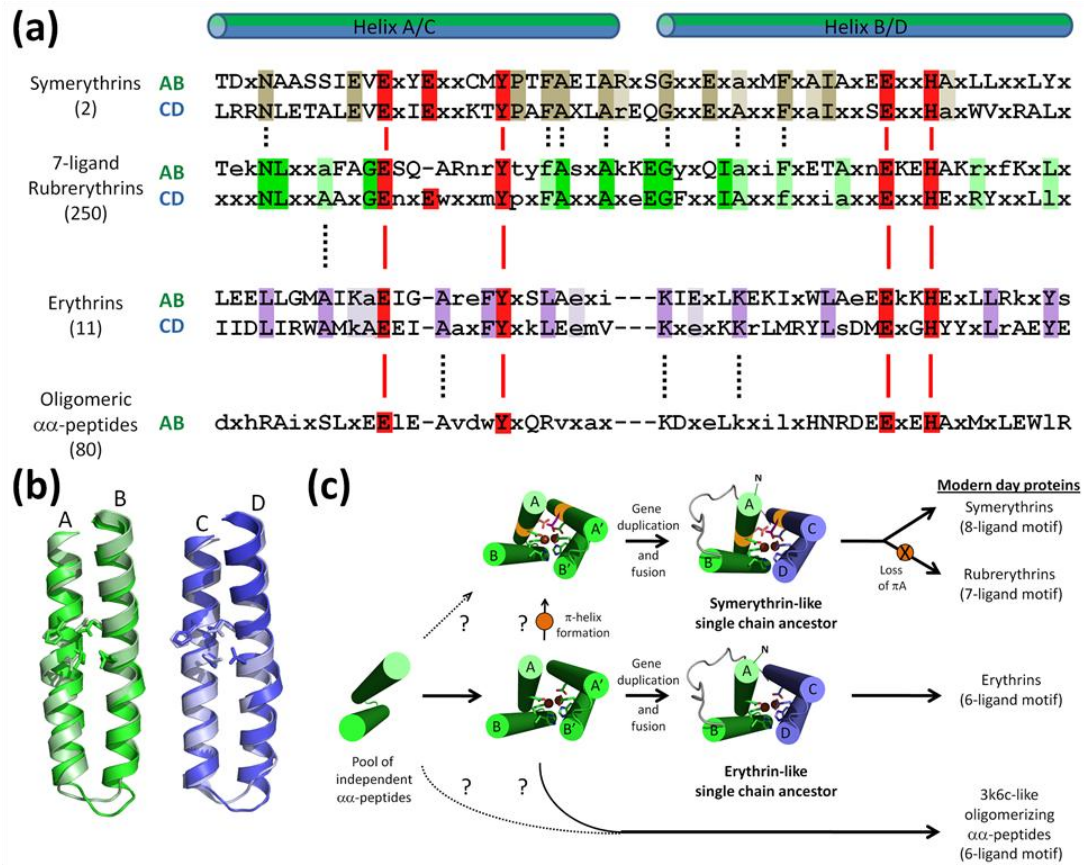


Figure 6.7. Oxidase and peroxidase cycling of *C. paradoxa* symerythrin. (a) Absorption spectra of diferrous non-crosslinked (gray lines) and crosslinked (black lines) symerythrin before (solid lines) and after (dashed lines) exposure to O_2. Dotted arrow emphasizes the direction of the spectral changes. (b) As for A but showing spectra before and after exposure to hydrogen peroxide. Inset: stoichiometry of the change in absorption coefficient at 340 nm as a function of hydrogen peroxide added to 50 μM diferrous, non-crosslinked symerythrin. Equivalent results were obtained for the crosslinked isoform.

Figure 6.8. Internal sequence similarities of three FLSF families and a model for their origins. (a) Internal alignment of consensus sequences from helix pairs A-B and C-D of symerythrins, of the rubrerythrin family members having seven metal ligating residues, of erythrins, and of the oligomerizing αα-peptide family. Upper- and lower-case letters give the identities of highly and moderately conserved residues, respectively, while "x" designates positions for which there is lesser sequence conservation (see *Material and Methods*). Conserved metallocenter residues (including the secondary sphere Tyr residues) are highlighted in red and are connected across families with solid red lines. Residues that are internally symmetric in symerythrins, 7-ligand rubrerythrins, and erythrins are highlighted in gray, green and purple, respectively. A dark highlight coloring is used for positions that match highly conserved residues and a light highlight color is used for positions matching a moderately conserved residue with either a moderately conserved or highly conserved residue. The complete SeqLogos plot(Schneider and Stephens 1990) of these consensus sequences are shown in Fig. A3.1. Dashed black lines connect residue pairs that match between the groups. In parenthesis below the family name is the number of sequences used to generate the consensus sequences for that group. (b) Overlay of helix pairs A-B and C-D of symerythrin (dark green and dark blue, respectively) on the equivalent helices of erythrin (light green and light blue, respectively; PDB 2fzf) shows the structural difference associated with the three residue gap in erythrins compared to rubrerythrins. Side chains of metallocenter residues are shown as sticks. (C) Schematic for the origins of modern day symerythrins, 7-ligand rubrerythrins, erythrins and the oligomerizing αα-peptides. Solid lines represent a simple direct pathway for the formation of each modern day family assuming that the dimerizing two-helix peptides are all homologous with each other. The dotted lines (with "?") are alternatives to the thinner solid lines (with "?") recognizing that with the small number of residues involved in the metallocenter, it is also possible that the distinct dimerizing peptides were not homologous.

Figure 6.8. (continued)



Chapter 7

Conclusion

Impacts

One important quality of research is its ability to influence not only the immediate field of study but also fields further removed. Although some of the implications reported in this dissertation are applicable to only a narrow set of disciplines, many results provide important insights with more general applications to the fields of protein evolution, protein structure and even bioinformatics. I will first summarize what are in my opinion the important impacts of the research presented in this dissertation and then will outline areas of future work.

BMM structure, function, physiology and evolution. Out of Chapters 2, 3 and 4 come several important advances in our understanding of BMM function. One of the more unexpected observations was the finely tuned substrate specificity of sBMO for substrates on which the host organism can grow. Generally speaking, such a phenomenon is not all that surprising. In the case of sBMO, however, the substrates in question have no functional groups with which to make a specific network of 3-dimensional hydrogen-bonded or ionic interactions. Nobel laureate Konrad Bloch once noted, "the stereospecific removal of hydrogen in the formation of oleate…would seem to approach the limits of the discriminatory power of enzymes" (Bloch 1969). Without the ability to make such a network of substrate-enzyme interactions, the simplest explanation for the discriminatory power of sBMO is that its active site is rigid and perfectly shaped to accommodate the small linear alkanes C_2 to C_5. However, the notion that its fully buried active site is large enough to accommodate molecules at least the size of naphthalene discredits this possibility. Thus, our observations suggest there are still as of yet poorly recognized, highly complex mechanisms of substrate selection in BMMs that go beyond the traditional understandings of enzyme-substrate interactions.

In Chapter 4, we provide a new way to think about substrate selection in BMM enzymes like sBMO that could impact how future studies in this area are approached. In particular, we uncover a novel function of π-helices that we termed "peristaltic-like

shifts" in which a π-helical bulge slides up and down along an α-helix to alter the size and shape of the active site in preparation for substrate binding and product release. This shifting of π-helices, which occurs in response regulatory subunit binding in toluene monooxygenase, provides a plausible explanation the data presented in Chapter 2 in which the binding affinities of alcohols to the active site of sBMO change in response to the binding of sBMO's regulatory subunit. It is reasonable to suggest therefore that the dynamic nature of the π-helices in BMMs is an important factor in controlling substrate specificity for these enzymes. Furthermore, there are as many as ten other π-helices littered throughout the structure of BMMs, making this family of enzymes the current record holder for harboring the most π-helices. Since they are typically associated with functional sites, these π-helices provide clues for uncovering novel functional sites in BMMs. To date, no study has attempted to experimentally characterize such functions, largely because the π-helices in BMMs have gone unnoticed and also because of limitations in our ability to create site-directed variants of these proteins. With our new understanding of the importance of π-helix functionality and with recent advancements in the engineering of sMMO and sBMO variants, new insights will most certainly be made in our understanding of BMM catalysis.

The characterization of π-helices not only impacted our understanding of BMM catalysis, it also provided new clues about the evolutionary origins of these enzymes. Prior to this work, researchers hypothesized that the α- and β-subunits of BMMs originated from the gene duplication of an RNR-like ancestor based on similarities in overall structure. Our work in Chapter 4 takes this hypothesis a step further and provides compelling evidence that both of the BMM subunits originated from an RNR Ic-like ancestor largely because of the strict conservation of a particular active site π-helix across all three proteins (RNR Ic, MMOHα and MMOHβ). This observation suggests that formation of this π-helix was an important evolutionary step in the creation of BMMs. This is a novel observation because as noted in Chapter 1, comparative studies between RNRs and BMMs have only focused on changes in

residues at or near the metallocenters rather than backbone geometry. As 15% of all proteins have at least one π-helix, the tracking of their formation and loss across evolutionarily related families should form a new and informative tool for future analysis.

In addition to these functional and evolutionary implications, another important impact of determining the substrate specificities of sBMO is that it suggested *T. butanivorans*, as a non-methanotroph, could grow in an environment where the only source of carbon and energy is natural gas. The studies in Chapter 3 demonstrated that although sBMO provides an effective filter to minimize methane oxidation in the presence of trace quantities of growth substrate, detectable quantities of methanol (the oxidation product of methane) were observed. *T. butanivorans* not only efficiently detoxifies these methane metabolites but could also extract energy (in the form of ATP) from this process. These observations suggest that an ancestral heterotrophic *T. butanivorans*-like organism acquired an sMMO-like gene cluster from a methanotroph and, because it didn't have the biochemical machinery to assimilate C_1 compounds into biomass, the monooxygenase evolved to preferentially oxidize substrates the cell could metabolize downstream (i.e. ethane, propane and butane). Given how relatively little is understood about the physiology and enzymology of short chain gaseous alkane utilizing bacteria compared to methanotrophs (Shennan 2006), natural gas seepages may be a prime spot to search for more such alkane utilizing organisms in order to expand our understanding of BMM diversity and potentially discover new enzymes and organisms useful to the fields like bioremediation.

Protein structure and evolution. The largest bottleneck in understanding how modern day proteins came to be is that no fossil records of ancient genes or proteins exist, forcing researchers to rely only on information from extant organisms to recreate the past. Fortunately, recent advances in genomic sequencing techniques and protein structure determination have opened new doors in the area of protein and organismal

evolution. The results outlined in Chapter 4 provide a great example of this with our uncovering of the evolutionary origin of π-helices.

π-helices. The biggest impact of these findings is that they bring the importance of π-helices to the forefront of protein structure and evolution by demonstrating (i) they occur in 1 in 6 proteins despite previous assertions that they either don't exist (Creighton 1993) or are extremely rare (Weaver 2000), (ii) they are evolutionarily related to α-helices via the insertion of a single amino acid, and (iii) their formation/loss is often correlated with changes in protein functionality. Together, these points reinforce the importance of checking structures carefully for the presence of π-helices, and if any are found then those sites are likely of functional importance. Furthermore, we provide a new experimental approach to testing the functional relevance of any given π-helix by deleting a single amino acid within the π-helical region to convert it to an α-helix.

As most structural biologists do not immediately recognize π-helices in proteins (a perfect example is the Na^+/Cl^--dependent neurotransmitter transporter, which has eight π-helices in a single chain but none were noticed by the authors (Yamashita *et al.* 2005)), I wrote a freely available program (called π-HUNT) that can analyze thousands of structures to look for π-helices and output the all the information into a spreadsheet. This simple tool will allow more researchers without expertise in the details of secondary structure conformation to recognize π-helices. To facilitate this even more, we provided with the publication of Chapter 4 an Excel spreadsheet cataloging 15,000 non-redundant structures with how many π-helices they have and where to find them in the structure. This database is a treasure trove of information for structural and evolutionary biologists which can be used to (i) identify previously unrecognized functional sites of proteins, (ii) provide insight into protein evolution and (iii) uncover new functions of π-helices. While such follow up studies are not within the scope this dissertation, these tools undoubtedly offer great potential for exciting future research projects.

The biggest reason the evolutionary relationship between π-helices and α-helices has not been previously recognized is because commonly used automated secondary structure software almost always fail (~95% of the time) to recognize π-helices. However, there is also a more subtle reason this association was not recognized: sequence alignment software typically enforce heavy "gap penalties" during the alignment of two sequences in order to prevent the opening of gaps that would otherwise result in over fitting or over estimating the similarity of two sequences. Furthermore, when secondary structure is assigned to a particular sequence prior to alignment (either through secondary structure prediction programs, homology modeling or experimental information), gap penalties are typically increased by as much as eight times that of non-helical residues (Lesk *et al.* 1986). As a result, sequence alignment programs rarely incorporate single amino acid gaps in α-helices, thereby hiding the relationship of α- and π-helices.

If the three-dimensional structure of the protein of interest is known, one can avoid such alignment errors by using structure-based sequence alignment software. These alignments depend on the overlay of two similar structures rather than the alignment of two sequences. This method is particularly beneficial in cases where the sequences are too divergent to generate reliable sequence-based alignments. Thus structure-based alignments are always preferable provided the necessary information is available. Yet surprisingly I have used several popular structure-based alignment servers (such as PDBeFold (Velankar *et al.* 2010)) that output erroneous sequence alignments when comparing α-/π-helical homologs even though the global structure alignment is of high quality. The inability to recognize these errors in aligning two sequences in which one has a π-helix and the other an α-helix is quite evident in many alignments reported in the literature today, and is particularly common in the alignment of FLSF members in which π-helices are widespread. In fact, Andrews' most recent alignment of π-helix-containing rubrerythrin proteins with the purely α-helical erythrins (see Fig. 5 of (Andrews 2010)) did not incorporate a gap in the midst of helix C, resulting in the misalignment of the highly conserved metal-ligating

residues. Based on the reported alignment, the reader is left to assume that the metal-ligating residues are not positionally conserved across these families. This incorrect alignment is particularly surprising given that there exist representative structures for both families that when structurally aligned demonstrate the metal-ligating residues are in fact in structurally equivalent positions. From my experience, the DALI server (Holm *et al.* 2008) is the most reliable in this aspect and is the database we used to conduct our studies discussed in Chapter 4. Wider use of the π-HUNT to identify π-helices in protein structures would help promote an appropriate level of awareness to ensure gaps are placed in the proper spots in their sequence alignments. Without these tools, current sequence alignment programs cannot be relied on to produce accurate alignments of α-/π-helical homologs.

Understanding the relationship between α- and π-helices and the ability to identify the latter in sequence alignments was essential to realizing the uniqueness of the symerythrin sequence. This sequence from *Cyanophora paradoxa* was published in 1995 (Stirewalt *et al.* 1995), and the first mention of its relationship to rubrerythrin-like proteins was in 1998 (Andrews 1998). Since then, it has been mentioned in the literature on a handful of occasions (Morton 1998; Gomes *et al.* 2001; Wakagi 2003; Pinto *et al.* 2011), but only in reference to it being a divergent rubrerythrin. One of the most likely reasons that other researchers did not notice the uniqueness of this sequence is that alignments of symerythrin created by the NCBI BLAST server with other rubrerythrin proteins do not include the first forty amino acids of symerythrin, which is exactly the region where the π-helix and additional metal-ligating residue unique to symerythrin resides. Furthermore, the homology model deposited in the SWISS-MODEL repository (Kiefer *et al.* 2009) for symerythrin doesn't contain the first 40 amino acids either. In other words, the most commonly used tools to search and compare protein sequences literary hid the unique aspects of symerythrin. One must wonder, therefore, how many other sequences with cryptic π-helices and novel functionalities are waiting to be discovered.

Peptide-based origins of the FLSF. From our analysis of π-helices in the FLSF, we proposed that symerythrin possessed higher symmetry between the two halves of its four-helix bundle compared to any other characterized FLSF protein. The results of our structural characterization (Chapter 6) supported our proposal and provided original evidence for the "proteins from peptides" hypothesis of protein evolution. Of the ten superfolds (that is, folds into which a quarter of all domains can be categorized (Soding and Lupas 2003)), only three have substantial evidence that the fold originated from the concatenation of smaller peptides. These are(the TIM-barrel (Lang *et al.* 2000), β-trefoil (Lee and Blaber 2011) and ferredoxin-like (Dauter *et al.* 1997). This dissertation provides new experimental evidence that the four-helix bundle fold also had similar peptide-based origins. An unexpected result, however, was that the single chain four-helix bundle fold of the FLSF originated more than once. This observation changes the generally accepted notion that all members of a superfamily are homologous at the domain level and therefore share a common ancestor at the level of the fold. It also underscores the notion that function likely predated fold formation. Having a specific functionality preexisting the formation of a fold makes sense because if there were no functionality (other than folding), there would be no selective pressure to maintain the genes for that fold.

The valine-phenylalanine crosslink. The discovery of the autogenerated Val-Phe crosslink is important because it opens new doors to the types of chemistry proteins are capable of performing. To our knowledge, no protein has been observed to make a carbon-carbon bond between two unfunctionalized hydrocarbons, whether between two substrates or to form a crosslink. Although this finding was most unexpected, I would argue its discovery cannot be considered pure luck. We hypothesized in Chapter 4 that symerythrin would have a novel metallocenter, and given how sensitive the reactivity of diiron centers are to their local environment, it doesn't come as a surprise that symerythrin's novel metallocenter is capable of performing novel chemistry. Currently it is not clear whether this carbon-carbon bond making functionality has any relevance to its physiologic function, but the fact that

symerythrin can do it provides new groundwork for making biologically inspired catalysts. Today most of these studies are based on the diiron center of methane monooxygenase, which performs one of the most energetically challenging reactions in biology. Knowing that symerythrin can perform analogous reactions (that is the breaking of stable C-H bonds of hydrocarbons) should lead to new approaches in this field.

Rubrerythrins and the origins of oxygenic photosynthesis. The peptide-based model for the origins of the FLSF (Chapter 6) provides an important explanation for a noteworthy correlation between the functional evolution of the FLSF and the transition from an anoxic to oxic atmosphere. The correlation is that most FLSF members with six-residue metallocenters are highly reactive with O_2 and even require it as a substrate. This implies that their functionality evolved after O_2 was present in the atmosphere. In contrast, rubrerythrins do not preferentially react with O_2 but instead are found only in obligate anaerobes or microaerobes to protect against hydrogen peroxide (Riebe *et al.* 2009), while symerythrin is found in the earliest branching lineage of oxygenic photosythetic organisms (Nelissen *et al.* 1995). Given that life and the first protein folds evolved under anoxic conditions (Alva *et al.* 2010), this observation is consistent with the notion that symerythrins and rubrerythrins, as the earliest members of the FLSF, formed pre-O_2 or during the early stages of the microaerobic world. The ensuing increase in atmospheric oxygen could have resulted in a selective advantage for the 6-ligand metallocenter (formed either through an independent gene duplication/fusion event of a two-helix peptide or from the rubrerythrin/symerythrin family) which allowed for a major expansion of FLSF functions utilizing O_2.

Combining this correlation of the origins of the FLSF and the beginnings of Earth's oxygenic atmosphere with the fact that symerythrin is only found in the most primitive oxygenic phototrophs suggests that symerythrin played an important role in the origins of oxygenic photosynthesis. At this point the function of symerythrin remains a mystery and unfortunately there is limited literature regarding the

physiology of both *G. violaceus* and *C. paradoxa*. Thus the role of symerythrin in arguably the greatest biological innovation after the originsof life remains highly speculative. Nevertheless, such correlations certainly warrant future investigations.

Future Work

The studies described in this dissertation have provided the foundation for several interesting follow-up research projects. In this section, I briefly outline four such projects that extend these studies and rely on a similar multi-disciplinary approach.

The role of π-helices in butane and methane monooxygenase. The findings described in Chapters 2 and 4 indicate that π-helices play an important role in the function of sMMO and sBMO. Knowing that π-helices and α-helices can interconvert via the insertion/deletion of a single amino acid, we now have a strategy to begin uncovering the function and dynamics of π-helices in greater detail. Historically, a major complication that hindered the creation of site-directed variants of sMMO and sBMO was the inability to find a suitable recombinant expression system. The reasons for this are not well understood, but it suggests that poorly understood chaperone-assisted processes are necessary for proper folding of these multi-subunit complexes. There are still no reports of a recombinant expression system for sMMO, however strategies for developing site-directed variants in the native bacterium have recently been described (Smith and Murrell 2011). Furthermore, our laboratory has developed similar strategies with sBMO from *T. butanivorans* (Halsey *et al.* 2006). These experimental advances, combined with a novel approach for studying the function of π-helices, provides a fruitful area for understanding substrate specificity and the activation of the hydroxylase by the regulatory subunits.

Because π-helices πD, πB and πE from Fig. 4.6 are now directly implicated in the function of these monooxygenases, these would be the π-helices worth investigating initially. To convert these π-helices into α-helices, variants in which

single amino acids are deleted from these regions would be created. Even if they are non-functional, the methanotroph can still grow on methane using the particulate form of methane monooxygenase as long as there is sufficient copper in the media. Once copper becomes depleted, expression of the sMMO is induced, at which time the cells can be harvested and the protein isolated. Assuming they are well folded, an advantage to using sMMO over sBMO is that sMMO variants can be crystallized in the same manner as the wild-type form, allowing for direct visualization of these α-helical variants. Soaking substrate analogues like 6-bromohexan-1-ol, which induced π-helical peristalsis in wild-type sMMO, into the crystals of these variants would provide insight into the role of πB in substrate binding. Lastly, assuming functional α-helical variants of sMMO are identified, making analogous variants of sBMO and conducting kinetic analyses with and without its regulatory component (similar to those described in Chapter 2) will undoubtedly give valuable insight in the mechanisms of substrate specificity and hydroxylase activation by regulatory subunit. Having both systems to work with (sMMO and sBMO) is highly valuable because experiments that are difficult or impossible in one system may be feasible in the other.

Mechanism of crosslink formation in symerythrin. Given the effort that has been placed in characterizing the functionalization of hydrocarbons for generating catalysts and for their applications in bioenergetics and bioremediation (Bergman 2007), understanding the mechanism by which the valine-phenyalanine crosslink is generated is of great interest. In Chapter 5, we proposed that crosslink formation proceeds initially via formation of a high-valent iron-oxygen Q-like intermediate similar to that of sMMO (see Fig. 1.5), which is the most similar diiron protein capable of breaking the C-H bonds of unfunctionalized substrates. This intermediate then abstracts a hydrogen atom from Val127 to form a radical, which then adds to the phenylalanine ring of Phe17. After deprotonation and return of an electron to the diiron center from the cyclohexadienyl radical, the crosslink is formed.

Currently there is no direct experimental evidence to support this mechanism, and certainly alternative mechanisms can be envisioned. Given the differences

between the diiron metallocenters of symerythrin (Figs. 6.3A and 6.5B) and sMMO (Fig. 1.2C), it seems unlikely that the exact natures of the iron-oxygen intermediate responsible for breaking C-H bonds are identical. The combination of a variety of techniques such as sequential mixing stopped-flow UV/Vis spectroscopy, rapid freeze quenching, electron paramagnetic resonance and Mossbauer spectroscopy will prove useful for elucidating the interactions of molecular oxygen with the diferrous metallocenter of symerythrin. These results will provide a novel framework for understanding oxygen activation by metalloproteins, which is an essential part of many metabolic processes.

After characterizing the nature of oxygen activation by the diiron center of symerythrin, the second goal of this project would be to understand the mechanism by which the crosslinking bond is formed. Although it is commonly assumed to proceed via radical chemistry, the exact mechanism of C-H bond breakage in methane by the Q-intermediate of sMMO is still poorly understood. This is largely due to the difficulty in trapping the presumed free radicals (particularly methyl radicals) for long enough to characterize them spectroscopically. It is reasonable to suspect we would run into similar obstacles in characterizing the reaction that leads to crosslink formation in symerythrin. This is further complicated by unpublished observations that show crosslink formation does not occur each time the diferrous metallocenter of symerythrin reacts with molecular oxygen. The reason for this is not understood, but it may result from either inefficient formation of the appropriate iron-oxygen intermediate, the low likelihood of Val127 interacting with the activated oxygen intermediate due to its distance from the diiron center, or low probability of Val127 and Phe17 interacting with each other at the right time and orientation. One potential approach to improve the efficiency of crosslink is to use a V127L variant of symerythrin. As its side chain is also an unfunctionalized hydrocarbon but with one additional carbon compared to that of valine, leucine may be more likely interact with the activated iron-oxygen intermediate each time it is formed. This greater proximity could allow for more efficient crosslink formation, thus facilitating its study. In terms

of the metallocenter chemistry, we do know that crosslink formation is dependent on the metal-ligating glutamate unique to symerythrin because when this residue is changed to an Ala, no crosslinked symerythrin is observed after expression in *E. coli* (Fig. 7.1). This shows that unique metal ligating residue of symerythrin, E40, is essential for crosslink formation. Whether it is involved in stabilizing the iron-oxygen intermediates or the putative radical intermediates of Val127 or Phe17 remains to be determined.

Relevance of crosslinked symerythrin *in vivo*. All of the symerythrin studies reported in this dissertation have been conducted on recombinant symerythrin expressed in *E. coli*. Currently, we are growing the two oxygenic photosynthetic organisms *Cyanophora paradoxa* and *Gloeobacter violaceus* and attempting to detect the presence of symerythrin in them. Preliminary results obtained by immuno-blotting has shown that non-crosslinked symerythrin is present at very low levels in the cyanelles of *C. paradoxa* (Fig. 7.2B). Upon exposure to increased light intensities or hydrogen peroxide, crosslinked symerythrin was detected in similarly low quantities.

Closer inspection of these immuno-blots revealed an unexpected finding: a more intense band migrating at approximately twice the expected molecular weight of non-crosslinked symerythrin (Fig. 7.2A). Given that these blots are done in denaturing conditions with a disulfide reducing agent, this result suggests that symerythrin exists as a covalent dimer *in vivo*. This can be explained if symerythrin forms an interpeptide crosslink rather than the intrapeptide crosslink that we have characterized thus far. Structurally, such a domain-swapped dimer would be possible given the presumed flexibility of the N-terminal tail. Although symerythrin at neutral pH is a monomer, the crystal structure suggests that at least in acidic pH values, it could exist as a dimer since the interface between the two copies of symerythrin in the asymmetric unit of the crystal covers over 1000 $Å^2$ making eight hydrogen bonds and six salt bridges. Clearly, characterizing the *in vivo* oligomerization and crosslink status of symerythrin will yield some novel and interesting results.

Physiological role of symerythrin. In Chapter 6, we report that symerythrin has enzymatic oxidase and peroxidase activity. However, its physiologic function in these two oxygenic phototrophs remains to be determined. One of the most informative experiments would be to knock out the symerythrin gene via insertion of an antibiotic resistance gene. Unfortunately, there is no precedent for mutational analyses in either of these organisms, and given how slowly *G. violaceus* grows (doubling time of 17 days (Rexroth *et al.* 2011)) such studies would take too long to do in a reasonable amount of time.

Alternatively, I would argue the most important series of experiments required to determine the physiological function of symerythrin should be in regards to the identification of the physiologically relevant electron relay system, if one exists. Thus far we have been utilizing an electron relay system that appears to be slow and inefficient in terms of its ability to reduce the diiron center of symerythrin (spinach ferredoxin and ferredoxin reductase). One way to identify such an electron relay system is to form a symerythrin affinity column by covalently attaching recombinant symerythrin to Sepharose beads and then flow *C. paradoxa* cyanelle extracts through this column. After thorough washing of the column, only proteins that interact with symerythrin would presumably still be bound. Elution of these interacting proteins follow by characterization via mass spectrometry methods could provide important clues in indentifying symerythrin's physiologically relevant reductase. This would be an ideal starting place to hone in on symerythrin's function and will lead to a greater understanding of its role in the development of oxygen photosynthesis and why, given its high residual sequence similarity, it appears to have evolved so slowly since its formation.

Concluding Remarks

In this dissertation, I have presented the results from five years of research regarding the structure, function and evolution of diiron proteins from the ferritin-like

superfamily. These research projects have led to five publications and have spanned a wide range of techniques from molecular biology and cellular physiology to x-ray crystallography and computer programming. As with any research project, we did not achieve several goals that we initially set for ourselves (i.e. crystallizing the hydroxylase component of soluble butane monooxygenase), but at the same time answered many other important questions and opened the doors to new and unexpected projects. The priority of our laboratory right now is to use these results as the basis for grant submissions so that we can continue understanding the fascinating chemistry performed by these ferritin-like metalloproteins. Further work in this area will no doubt prove to be exciting.

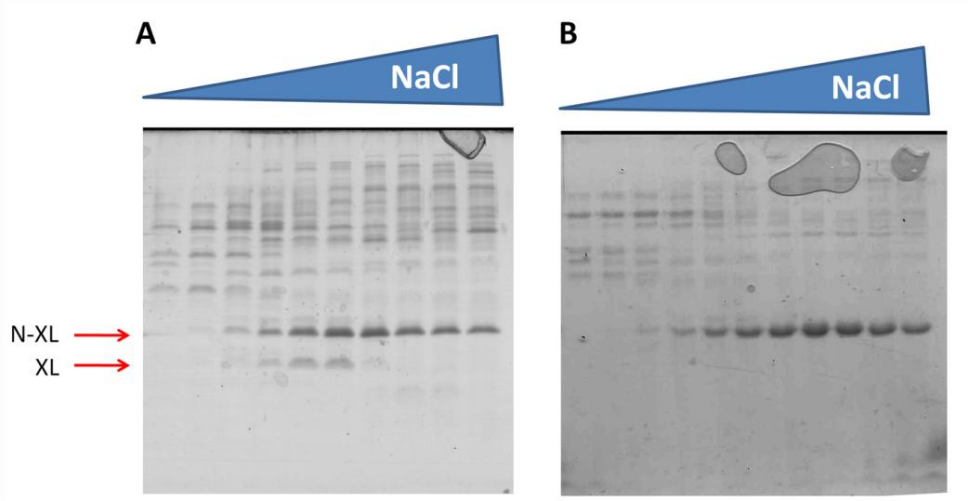


Figure 7.1. The E40A variant of symerythrin does not generate the Val-Phe crosslink after expression in *E. coli*. (A) Anion exchange fractionation of soluble extracts from *E. coli* after overexpression of wild-type symerythrin shows both non-crosslinked (N-XL) and crosslinked (XL) isoforms elute from the column. (B) The same experiment as panel A except that the E40A variant of symerythrin is used. This variant in which the metal-ligating glutamate unique to symerythrin is a altered to an alanine does not form the crosslink.

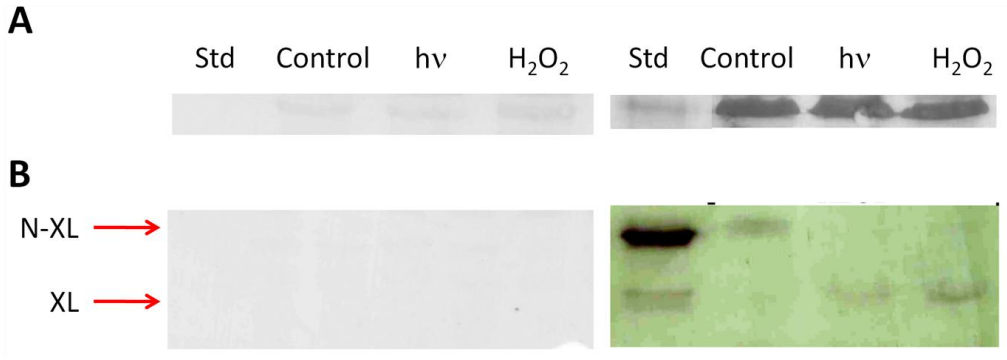


Figure 7.2. Identification of symerythrin in the cyanelles of *C. paradoxa* by immunoblot analysis. (A) The ~45 kDa molecular weight region of blots performed using pre-immune rabbit serum (left) and rabbit serum with α-symerythrin polyclonal antibodies (right). The four lanes of each blot from left to right are purified recombinant symerythrin (150 ng), solubilized cyanelles from control cells, solubilized cyanelles from cells exposed to excess light, and solubilized cyanelles from cells exposed to hydrogen peroxide (50 μg of total protein in each lane). Each exposure lasted 6 hours. We hypothesize that this 45 kDa band recognized by the α-symerythrin antibody is the crosslinked symerythrin dimer. (B) The 15-22 kDa range of the same blots described in panel A. Arrows labeled with N-XL and XL correspond to non-crosslinked and crosslinked symerythrin, respectively. Blots were developed using goat-α-rabbit antibodies conjugated to alkaline phosphatase.

Bibliography

Adams, P. D., P. V. Afonine, G. Bunkoczi, *et al.* (2010). "PHENIX: a comprehensive Python-based system for macromolecular structure solution." Acta Crystallogr D Biol Crystallogr **66**(Pt 2): 213-221.

Alva, V., M. Remmert, A. Biegert, *et al.* (2010). "A galaxy of folds." Protein Sci **19**(1): 124-130.

Andersson, C. S. and M. Hogbom (2009). "A *Mycobacterium tuberculosis* ligand-binding Mn/Fe protein reveals a new cofactor in a remodeled R2-protein scaffold." Proc Natl Acad Sci U S A **106**(14): 5633-5638.

Andersson, M. E., M. Hogbom, A. Rinaldo-Matthis, *et al.* (1999). "The crystal structure of an azide complex of the diferrous R2 subunit of ribonucleotide reductase displays a novel carboxylate shift with important mechanistic implications for diiron-catalyzed oxygen activation." Journal of the American Chemical Society **121**(11): 2346-2352.

Andrews, S. (1998). Iron Storage in Bacteria. Advances in Microbial Physiology. R. K. Poole. San Diego, Academic Press. **40:** 281-349.

Andrews, S. C. (2010). "The Ferritin-like superfamily: Evolution of the biological iron storeman from a rubrerythrin-like ancestor." Biochim Biophys Acta **1800**(8): 691-705.

Anthony, C. (1993). The Role of Quinoproteins in Bacterial Energy Transduction. Principles and Applications of Quinoproteins. V. L. Davidson. New York, CRC Press.

Armen, R., D. O. Alonso and V. Daggett (2003). "The role of α-, 3_{10}-, and π-helix in helix → coil transitions." Protein Sci **12**(6): 1145-1157.

Arp, D. J. (1999). "Butane metabolism by butane-grown *'Pseudomonas butanovora'*." Microbiology **145**: 1173-1180.

Ausubel, F. M., R. Brent, R. E. Kingston, *et al.*, Eds. (2003). Current Protocols in Molecular Biology, John Wiley & Sons Inc

Bailey, L. J., J. G. McCoy, G. N. Phillips, Jr., *et al.* (2008). "Structural consequences of effector protein complex formation in a diiron hydroxylase." Proc Natl Acad Sci U S A **105**(49): 19194-19198.

Balasubramanian, R., S. M. Smith, S. Rawat, *et al.* (2010). "Oxidation of methane by a biological dicopper centre." Nature **465**(7294): 115-119.

Baldwin, J., C. Krebs, L. Saleh, *et al.* (2003). "Structural Characterization of the Peroxodiiron (III) Intermediate Generated during Oxygen Activation by the W48A/D84E Variant of Ribonucleotide Reductase Protein R2 from *Escherichia coli*." Biochemistry **42**(45): 13269-13279.

Baldwin, J., W. C. Voegtli, N. Khidekel, *et al.* (2001). "Rational reprogramming of the R2 subunit of Escherichia coli ribonucleotide reductase into a self-hydroxylating monooxygenase." J Am Chem Soc **123**(29): 7017-7030.

Barlow, D. J. and J. M. Thornton (1988). "Helix geometry in proteins." J Mol Biol **201**(3): 601-619.

Behan, R. K. and S. J. Lippard (2010). "The aging-associated enzyme CLK-1 is a member of the carboxylate-bridged diiron family of proteins." Biochemistry **49**(45): 9679-9681.

Bergman, R. (2007). "C–H activation." Nature **446**(7134): 391-393.

Berman, H. M., J. Westbrook, Z. Feng, *et al.* (2000). "The Protein Data Bank." Nucleic Acids Res **28**(1): 235-242.

Bloch, K. (1969). "Enzymic synthesis of monounsaturated fatty acids." Accounts of Chemical Research **2**(7): 193-202.

Bloom, J. D. and F. H. Arnold (2009). "In the light of directed evolution: pathways of adaptive protein evolution." Proc Natl Acad Sci U S A **106 Suppl 1**: 9995-10000.

Bochevarov, A. D., J. Li, W. J. Song, *et al.* (2011). "Insights into the different dioxygen activation pathways of methane and toluene monooxygenase hydroxylases." J Am Chem Soc **133**(19): 7384-7397.

Borodina, E., T. Nichol, M. G. Dumont, *et al.* (2007). "Mutagenesis of the "leucine gate" to explore the basis of catalytic versatility in soluble methane monooxygenase." Appl Environ Microbiol **73**(20): 6460-6467.

Brazeau, B. J. and J. D. Lipscomb (2000). "Kinetics and activation thermodynamics of methane monooxygenase compound Q formation and reaction with substrates." Biochemistry **39**(44): 13503-13515.

Brzostowicz, P. C., D. M. Walters, R. E. Jackson, *et al.* (2005). "Proposed involvement of a soluble methane monooxygenase homologue in the cyclohexane-dependent growth of a new *Brachymonas* species." Environ Microbiol **7**(2): 179-190.

Burton, S. G. (2003). "Oxidizing enzymes as biocatalysts." Trends Biotechnol **21**(12): 543-549.

Cartailler, J. P. and H. Luecke (2004). "Structural and functional characterization of π-bulges and other short intrahelical deformations." Structure **12**(1): 133-144.

Chang, S. L., B. J. Wallar, J. D. Lipscomb, *et al.* (1999). "Solution structure of component B from methane monooxygenase derived through heteronuclear NMR and molecular modeling." Biochemistry **38**(18): 5799-5812.

Chatwood, L. L., J. Muller, J. D. Gross, *et al.* (2004). "NMR structure of the flavin domain from soluble methane monooxygenase reductase from Methylococcus capsulatus (Bath)." Biochemistry **43**(38): 11983-11991.

Choi, Y. S., H. Zhang, J. S. Brunzelle, *et al.* (2008). "In vitro reconstitution and crystal structure of *p*-aminobenzoate N-oxygenase (AurF) involved in aureothin biosynthesis." Proc Natl Acad Sci U S A **105**(19): 6858-6863.

Colby, J., D. I. Stirling and H. Dalton (1977). "The soluble methane mono-oxygenase of *Methylococcus capsulatus* (Bath). Its ability to oxygenate n-alkanes, n-alkenes, ethers, and alicyclic, aromatic and heterocyclic compounds." Biochem J **165**(2): 395-402.

Coleman, N. V., N. B. Bui and A. J. Holmes (2006). "Soluble di-iron monooxygenase gene diversity in soils, sediments and ethene enrichments." Environ Microbiol **8**(7): 1228-1239.

Cooley, R. B., D. J. Arp and P. A. Karplus (2010). "Evolutionary Origin of a Secondary Structure: π-Helices as Cryptic but Widespread Insertional Variations of α-Helices That Enhance Protein Functionality." J Mol Biol **404**(2): 232-246.

Cooley, R. B., B. L. Dubbels, L. A. Sayavedra-Soto, *et al.* (2009). "Kinetic characterization of the soluble butane monooxygenase from *Thauera butanivorans*, formerly *'Pseudomonas butanovora'*." Microbiology **155**: 2086-2096.

Cooley, R. B., T. W. Rhoads, D. J. Arp, *et al.* (2011). "A diiron protein autogenerates a valine-phenylalanine cross-link." Science **332**(6032): 929.

Cornish-Bowden, A. (1995). Fundamentals of Enzyme Kinetics. London, Portland Press Ltd.**:** 108.

Cotruvo, J. A. and J. Stubbe (2011). "*Escherichia coli* class Ib ribonucleotide reductase contains a dimanganese(III)-tyrosyl radical cofactor *in vivo*." Biochemistry **50**(10): 1672-1681.

Coulter, E. D., N. V. Shenvi, Z. M. Beharry, *et al.* (2000). "Rubrerythrin-catalyzed substrate oxidation by dioxygen and hydrogen peroxide." Inorg. Chim. Acta **297**(1-2): 231-241.

Coulter, E. D., N. V. Shenvi, Z. M. Beharry, *et al.* (2000). "Rubrerythrin-catalyzed substrate oxidation by dioxygen and hydrogen peroxide." Inorganica Chimica Acta **297**(1-2): 231-241.

Creighton, T. E. (1993). Proteins: Structures and Molecular Properties. New York, Freeman and Company**:** 182-186.

Crichton, R. R. and J. P. Declercq (2010). "X-ray structures of ferritins and related proteins." Biochim Biophys Acta **1800**(8): 706-718.

Crooks, G. E., G. Hon, J. M. Chandonia, *et al.* (2004). "WebLogo: a sequence logo generator." Genome Res **14**(6): 1188-1190.

Crowther, G. J., G. Kosaly and M. E. Lidstrom (2008). "Formate as the main branch point for methylotrophic metabolism in *Methylobacterium extorquens* AM1." J Bacteriol **190**(14): 5057-5062.

Dauter, Z., K. S. Wilson, L. C. Sieker, *et al.* (1997). "Atomic resolution (0.94 A) structure of *Clostridium acidurici* ferredoxin. Detailed geometry of [4Fe-4S] clusters in a protein." Biochemistry **36**(51): 16065-16073.

Davis, I. W., A. Leaver-Fay, V. B. Chen, *et al.* (2007). "MolProbity: all-atom contacts and structure validation for proteins and nucleic acids." Nucleic Acids Res **35**(Web Server issue): W375-383.

Dawson, M. J. and C. W. Jones (1981). "Respiration-linked proton translocation in the obligate methylotroph *Methylophilus methylotrophus*." Biochem J **194**(3): 915-924.

Dean, A. M. and J. W. Thornton (2007). "Mechanistic approaches to the study of evolution: the functional synthesis." Nat Rev Genet **8**(9): 675-688.

deMare, F., D. M. Kurtz, Jr. and P. Nordlund (1996). "The structure of Desulfovibrio vulgaris rubrerythrin reveals a unique combination of rubredoxin-like FeS4 and ferritin-like diiron domains." Nat Struct Biol **3**(6): 539-546.

deMare, F., P. Nordlund, N. Gupta, *et al.* (1997). "Re-engineering the diiron site in rubrerythrin towards that in ribonucleotide reductase." Inorg. Chim. Acta **263**(1): 255-262.

Diederichs, K. and P. A. Karplus (1997). "Improved R-factors for diffraction data analysis in macromolecular crystallography." Nat Struct Biol **4**(4): 269-275.

Dillard, B. D., J. M. Demick, M. W. Adams, *et al.* (2011). "A cryo-crystallographic time course for peroxide reduction by rubrerythrin from *Pyrococcus furiosus*." J Biol Inorg Chem: E-pub ahead of print.

Do, L. H. and S. J. Lippard (2011). "Toward Functional Carboxylate-Bridged Diiron Protein Mimics: Achieving Structural Stability and Conformational Flexibility Using a Macrocylic Ligand Framework." J Am Chem Soc.

Doughty, D. M., K. H. Halsey, C. J. Vieville, *et al.* (2007). "Propionate inactivation of butane monooxygenase activity in *'Pseudomonas butanovora'*: biochemical and physiological implications." Microbiology **153**(Pt 11): 3722-3729.

Doughty, D. M., E. G. Kurth, L. A. Sayavedra-Soto, *et al.* (2008). "Evidence for involvement of copper ions and redox state in regulation of butane monooxygenase in Pseudomonas butanovora." J Bacteriol **190**(8): 2933-2938.

Doughty, D. M., L. A. Sayavedra-Soto, D. J. Arp, *et al.* (2006). "Product repression of alkane monooxygenase expression in Pseudomonas butanovora." J Bacteriol **188**(7): 2586-2592.

Dubbels, B. L., L. A. Sayavedra-Soto and D. J. Arp (2007). "Butane monooxygenase of *'Pseudomonas butanovora'*: purification and biochemical characterization of a terminal-alkane hydroxylating diiron monooxygenase." Microbiology **153**(6): 1808-1816.

Dubbels, B. L., L. A. Sayavedra-Soto, P. J. Bottomley, *et al.* (2009). "Thauera butanivorans sp. nov., a C2-C9 alkane-oxidizing bacterium previously referred to as *'Pseudomonas butanovora'*." Int J Syst Evol Microbiol **59**(Pt 7): 1576-1578.

Dubbels, B. L., L. A. Sayavedra-Soto, P. J. Bottomley, *et al.* (2009). "*Thauera butanivorans* sp. nov., a C_2-C_9 alkane oxidizing bacterium previously referred to as *'Pseudomonas butanovora'*." Int J Syst Evol Microbiol: (in press).

Elliott, S. J., M. Zhu, L. Tso, *et al.* (1997). "Regio- and Stereoselectivity of Particulate Methane Monooxygenase from *Methylococcus capsulatus* (Bath)." J Am Chem Soc **119**(42): 9949-9955.

Emsley, P., B. Lohkamp, W. G. Scott, *et al.* (2010). "Features and development of Coot." Acta Crystallogr D Biol Crystallogr **66**(Pt 4): 486-501.

Enzien, M. V., F. Picardal, T. C. Hazen, *et al.* (1994). "Reductive Dechlorination of Trichloroethylene and Tetrachloroethylene under Aerobic Conditions in a Sediment Column." Appl Environ Microbiol **60**(6): 2200-2204.

Eriksson, M., A. Jordan and H. Eklund (1998). "Structure of *Salmonella typhimurium* nrdF ribonucleotide reductase in its oxidized and reduced forms." Biochemistry **37**(38): 13359-13369.

Evans, P. R. (1997). Recent advances in phasing. Proc. CCP4 Study Weekend. K. S. Wilson, G. Davies, A. W. Ashton and S. Bailey. Daresbury Laboratory, Warrington, UK**:** 97–102.

Farris, J. (1977). "Phylogenetic analysis under Dollo's Law." Syst Zool **26**(1): 77-88.

Feig, M., A. D. MacKerell Jr and C. L. Brooks III (2003). "Force field influence on the observation of π-helical protein structures in molecular dynamics simulations." J. Phys. Chem. B **107**(12): 2831-2836.

Fenton, H. (1894). "LXXIII.-Oxidation of tartaric acid in presence of iron." Journal of the Chemical Society, Transactions **65**: 899-910.

Fodje, M. N. and S. Al-Karadaghi (2002). "Occurrence, conformational features and amino acid propensities for the π-helix." Protein Eng **15**(5): 353-358.

Fox, B. G., J. Shanklin, C. Somerville, *et al.* (1993). "Stearoyl-acyl carrier protein Δ^9-desaturase from *Ricinus communis* is a diiron-oxo protein." Proc Natl Acad Sci U S A **90**(6): 2486-2490.

Frishman, D. and P. Argos (1995). "Knowledge-based protein secondary structure assignment." Proteins **23**(4): 566-579.

Froland, W. A., K. K. Andersson, S. K. Lee, *et al.* (1992). "Methane monooxygenase component B and reductase alter the regioselectivity of the hydroxylase component-catalyzed reactions. A novel role for protein-protein interactions in an oxygenase mechanism." J Biol Chem **267**(25): 17588-17597.

Funhoff, E. G., U. Bauer, I. Garcia-Rubio, *et al.* (2006). "CYP153A6, a soluble P450 oxygenase catalyzing terminal-alkane hydroxylation." J Bacteriol **188**(14): 5220-5227.

Gibson, R. P., J. P. Turkenburg, S. J. Charnock, *et al.* (2002). "Insights into trehalose synthesis provided by the structure of the retaining glucosyltransferase OtsA." Chem Biol **9**(12): 1337-1346.

Gomes, C. M., J. Le Gall, A. V. Xavier, *et al.* (2001). "Could a diiron-containing four-helix-bundle protein have been a primitive oxygen reductase?" Chembiochem **2**(7-8): 583-587.

Green, J. and H. Dalton (1986). "Steady-state kinetic analysis of soluble methane mono-oxygenase from *Methylococcus capsulatus* (Bath)." Biochem J **236**(1): 155-162.

Gupta, N., F. Bonomi, D. M. Kurtz, Jr., *et al.* (1995). "Recombinant *Desulfovibrio vulgaris* rubrerythrin. Isolation and characterization of the diiron domain." Biochemistry **34**(10): 3310-3318.

Guy, J. E., I. A. Abreu, M. Moche, *et al.* (2006). "A single mutation in the castor Δ^9-18:0-desaturase changes reaction partitioning from desaturation to oxidase chemistry." Proc Natl Acad Sci U S A **103**(46): 17220-17224.

Halsey, K. H., D. M. Doughty, L. A. Sayavedra-Soto, *et al.* (2007). "Evidence for modified mechanisms of chloroethene oxidation in *Pseudomonas butanovora* mutants containing single amino acid substitutions in the hydroxylase alpha-subunit of butane monooxygenase." J Bacteriol **189**(14): 5068-5074.

Halsey, K. H., L. A. Sayavedra-Soto, P. J. Bottomley, *et al.* (2006). "Site-directed amino acid substitutions in the hydroxylase alpha subunit of butane

monooxygenase from *Pseudomonas butanovora*: Implications for substrates knocking at the gate." J Bacteriol **188**(13): 4962-4969.
Hamamura, N., R. T. Storfa, L. Semprini, *et al.* (1999). "Diversity in butane monooxygenases among butane-grown bacteria." Appl Environ Microbiol **65**(10): 4586-4593.
Hanson, R. S. and T. E. Hanson (1996). "Methanotrophic bacteria." Microbiol Rev **60**(2): 439-471.
Hardy, J. A., S. T. Walsh and H. C. Nelson (2000). "Role of an α-helical bulge in the yeast heat shock transcription factor." J Mol Biol **295**(3): 393-409.
Harris, T. K. and V. L. Davidson (1993). "A new kinetic model for the steady-state reactions of the quinoprotein methanol dehydrogenase from *Paracoccus denitrificans*." Biochemistry **32**(16): 4362-4368.
Heinz, D. W., W. A. Baase, F. W. Dahlquist, *et al.* (1993). "How amino-acid insertions are allowed in an α-helix of T4 lysozyme." Nature **361**(6412): 561-564.
Heinz, D. W., W. A. Baase, X. J. Zhang, *et al.* (1994). "Accommodation of amino acid insertions in an α-helix of T4 lysozyme. Structural and thermodynamic analysis." J Mol Biol **236**(3): 869-886.
Hindupur, A., D. Liu, Y. Zhao, *et al.* (2006). "The crystal structure of the E. coli stress protein YciF." Protein Sci **15**(11): 2605-2611.
Hobson, G. D. and E. N. Tiratsoo (1981). Introduction to Petroleum Geology. Houston, Gulf Publishing Company.
Hogbom, M., P. Stenmark, N. Voevodskaya, *et al.* (2004). "The radical site in chlamydial ribonucleotide reductase defines a new R2 subclass." Science **305**(5681): 245-248.
Hollingsworth, S. A., D. S. Berkholz and P. A. Karplus (2009). "On the occurrence of linear groups in proteins." Protein Sci **18**(6): 1321-1325.
Hollingsworth, S. A. and P. A. Karplus (2010). "A fresh look at the Ramachandran plot and the occurrence of standard structures in proteins." Biomol Concepts **1**(3-4): 271-283.
Holm, L., S. Kaariainen, P. Rosenstrom, *et al.* (2008). "Searching protein structure databases with DaliLite v.3." Bioinformatics **24**(23): 2780-2781.
Hunt, J. M. (1979). Petroleum Geochemistry and Geology. San Francisco, W.H. Freeman and Company.
IUPAC (1970). "IUPAC-IUB Commission on Biochemical Nomenclature. Abbreviations and symbols for the description of the conformation of polypeptide chains. Tentative rules (1969)." Biochemistry **9**(18): 3471-3479.
Iyer, R. B., R. Silaghi-Dumitrescu, D. M. Kurtz, Jr., *et al.* (2005). "High-resolution crystal structures of *Desulfovibrio vulgaris* (Hildenborough) nigerythrin: facile, redox-dependent iron movement, domain interface variability, and peroxidase activity in the rubrerythrins." J Biol Inorg Chem **10**(4): 407-416.
Jiang, W., D. Yun, L. Saleh, *et al.* (2007). "A manganese(IV)/iron(III) cofactor in *Chlamydia trachomatis* ribonucleotide reductase." Science **316**(5828): 1188-1191.

Jin, S., D. M. Kurtz, Jr., Z. J. Liu, *et al.* (2002). "X-ray crystal structures of reduced rubrerythrin and its azide adduct: a structure-based mechanism for a non-heme diiron peroxidase." J Am Chem Soc **124**(33): 9845-9855.

Johnson, E. L. and M. R. Hyman (2006). "Propane and *n*-butane oxidation by *Pseudomonas putida* GPo1." Appl Environ Microbiol **72**(1): 950-952.

Jones, D. T., W. R. Taylor and J. M. Thornton (1992). "The rapid generation of mutation data matrices from protein sequences." Comput Appl Biosci **8**(3): 275-282.

Jordan, A. and P. Reichard (1998). "Ribonucleotide reductases." Annu Rev Biochem **67**: 71-98.

Kabsch, W. and C. Sander (1983). "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features." Biopolymers **22**(12): 2577-2637.

Keefe, L. J., J. Sondek, D. Shortle, *et al.* (1993). "The α-aneurism: a structural motif revealed in an insertion mutant of *Staphylococcal nuclease*." Proc Natl Acad Sci U S A **90**(8): 3275-3279.

Kendrew, J. C., G. Bodo, H. M. Dintzis, *et al.* (1958). "A three-dimensional model of the myoglobin molecule obtained by x-ray analysis." Nature **181**(4610): 662-666.

Kido, Y., T. Shiba, D. K. Inaoka, *et al.* (2010). "Crystallization and preliminary crystallographic analysis of cyanide-insensitive alternative oxidase from Trypanosoma brucei brucei." Acta Crystallogr Sect F Struct Biol Cryst Commun **66**(Pt 3): 275-278.

Kiefer, F., K. Arnold, M. Künzli, *et al.* (2009). "The SWISS-MODEL Repository and associated resources." Nucleic acids research **37**(suppl 1): D387.

Kolbe, M., H. Besir, L. O. Essen, *et al.* (2000). "Structure of the light-driven chloride pump halorhodopsin at 1.8 A resolution." Science **288**(5470): 1390-1396.

Kopp, D. A. and S. J. Lippard (2002). "Soluble methane monooxygenase: activation of dioxygen and methane." Curr Opin Chem Biol **6**(5): 568-576.

Korth, H.-G. and W. Sicking (1997). "Prediction of methyl C-H bond dissociation energies by density functional theory calculations." J. Chem. Soc., Perkin Trans. 2: 715-720.

Kotani, T., Y. Kawashima, H. Yurimoto, *et al.* (2006). "Gene structure and regulation of alkane monooxygenases in propane-utilizing *Mycobacterium* sp. TY-6 and *Pseudonocardia* sp. TY-7." J Biosci Bioeng **102**(3): 184-192.

Kotani, T., T. Yamamoto, H. Yurimoto, *et al.* (2003). "Propane monooxygenase and NAD+-dependent secondary alcohol dehydrogenase in propane metabolism by *Gordonia* sp. strain TY-5." J Bacteriol **185**(24): 7120-7128.

Kurtz, D. M. (1990). "Oxo-Bridged and Hydroxo-Bridged Diiron Complexes - a Chemical Perspective on a Biological Unit." Chemical Reviews **90**(4): 585-606.

Kurtz, D. M., Jr. and B. C. Prickril (1991). "Intrapeptide sequence homology in rubrerythrin from *Desulfovibrio vulgaris*: identification of potential ligands to the diiron site." Biochem Biophys Res Comm **181**(1): 337-341.

Lang, D., R. Thoma, M. Henn-Sax, *et al.* (2000). "Structural evidence for evolution of the beta/alpha barrel scaffold by gene duplication and fusion." Science **289**(5484): 1546-1550.

Leahy, J. G., P. J. Batchelor and S. M. Morcomb (2003). "Evolution of the soluble diiron monooxygenases." FEMS Microbiol Rev **27**(4): 449-479.

Lee, J. and M. Blaber (2011). "Experimental support for the evolution of symmetric protein architecture from a simple peptide motif." Proc Natl Acad Sci U S A **108**(1): 126-130.

Lee, K. H., D. R. Benson and K. Kuczera (2000). "Transitions from α to π helix observed in molecular dynamics simulations of synthetic peptides." Biochemistry **39**(45): 13737-13747.

Lee, S. K., J. C. Nesheim and J. D. Lipscomb (1993). "Transient intermediates of the methane monooxygenase catalytic cycle." J Biol Chem **268**(29): 21569-21577.

LeGall, J., B. C. Prickril, I. Moura, *et al.* (1988). "Isolation and characterization of rubrerythrin, a non-heme iron protein from *Desulfovibrio vulgaris* that contains rubredoxin centers and a hemerythrin-like binuclear iron cluster." Biochemistry **27**(5): 1636-1642.

Lehmann, Y., L. Meile and M. Teuber (1996). "Rubrerythrin from Clostridium perfringens: cloning of the gene, purification of the protein, and characterization of its superoxide dismutase function." J Bacteriol **178**(24): 7152-7158.

Lesk, A. M., M. Levitt and C. Chothia (1986). "Alignment of the amino acid sequences of distantly related proteins using variable gap penalties." Protein Eng **1**(1): 77-78.

Leskovac, V. (2003). Comprehensive Enzyme Kinetics. New York, Kluwer Academic / Plenum Publishers.

Leslie, A. (1992). "Recent changes to the MOSFLM package for processing film and image plate data." Joint CCP4+ ESF-EAMCB newsletter on protein crystallography **26**(1).

Levinson, G. and G. A. Gutman (1987). "Slipped-strand mispairing: a major mechanism for DNA sequence evolution." Mol Biol Evol **4**(3): 203-221.

Li, Q., P. Sritharathikhun and S. Motomizu (2007). "Development of novel reagent for Hantzsch reaction for the determination of formaldehyde by spectrophotometry and fluorometry." Anal Sci **23**(4): 413-417.

Lide, D. R. and H. P. R. Frederikse, Eds. (1995). CRC Handbook of Chemistry and Physics. Boca Raton, FL, CRC Press, Inc.

Lieberman, R. L. and A. C. Rosenzweig (2005). "Crystal structure of a membrane-bound metalloenzyme that catalyses the biological oxidation of methane." Nature **434**(7030): 177-182.

Lopes Ferreira, N., H. Mathis, D. Labbe, *et al.* (2007). "*n*-Alkane assimilation and tert-butyl alcohol (TBA) oxidation capacity in *Mycobacterium austroafricanum* strains." Appl Microbiol Biotechnol **75**(4): 909-919.

Low, B. W. and R. B. Baybutt (1952). "The π-helix- A Hydrogen Bonded Configuration of the Polypeptide Chain." J Am Chem Soc **74**(22): 5806-5807.

Luecke, H., B. Schobert, H. T. Richter, *et al.* (1999). "Structure of bacteriorhodopsin at 1.55 A resolution." J Mol Biol **291**(4): 899-911.

Mahadevan, J., K. H. Lee and K. Kuczera (2001). "Conformational Free Energy Surfaces of Ala_{10} and Aib_{10} Peptide Helices in Solution." J. Phys. Chem. B **105**(9): 1863-1876.

Manfredi, G., L. Yang, C. D. Gajewski, *et al.* (2002). "Measurements of ATP in mammalian cells." Methods **26**(4): 317-326.

Mathevon, C., F. Pierrel, J. L. Oddou, *et al.* (2007). "tRNA-modifying MiaE protein from Salmonella typhimurium is a nonheme diiron monooxygenase." Proc Natl Acad Sci U S A **104**(33): 13295-13300.

Mitic, N., J. K. Schwartz, B. J. Brazeau, *et al.* (2008). "CD and MCD studies of the effects of component B variant binding on the biferrous active site of methane monooxygenase." Biochemistry **47**(32): 8386-8397.

Moenne-Loccoz, P., J. Baldwin, B. A. Ley, *et al.* (1998). "O_2 activation by non-heme diiron proteins: identification of a symmetric μ-1,2-peroxide in a mutant of ribonucleotide reductase." Biochemistry **37**(42): 14659-14663.

Moore, A. L. and M. S. Albury (2008). "Further insights into the structure of the alternative oxidase: from plants to parasites." Biochem Soc Trans **36**(Pt 5): 1022-1026.

Morton, B. R. (1998). Selection on the codon bias of chloroplast and cyanelle genes in different plant and algal lineages. J Mol Evol. **46:** 449-459.

Muller, J., A. A. Lugovskoy, G. Wagner, *et al.* (2002). "NMR structure of the [2Fe-2S] ferredoxin domain from soluble methane monooxygenase reductase and interaction with its hydroxylase." Biochemistry **41**(1): 42-51.

Murray, L. J. and S. J. Lippard (2007). "Substrate trafficking and dioxygen activation in bacterial multicomponent monooxygenases." Acc Chem Res **40**(7): 466-474.

Murray, L. J., S. G. Naik, D. O. Ortillo, *et al.* (2007). "Characterization of the arene-oxidizing intermediate in ToMOH as a diiron(III) species." J Am Chem Soc **129**(46): 14500-14510.

Murshudov, G. N., A. A. Vagin and E. J. Dodson (1997). "Refinement of macromolecular structures by the maximum-likelihood method." Acta Crystallogr D Biol Crystallogr **53**(Pt 3): 240-255.

Murzin, A. G., S. E. Brenner, T. Hubbard, *et al.* (1995). "SCOP: a structural classification of proteins database for the investigation of sequences and structures." J Mol Biol **247**(4): 536-540.

Nelissen, B., Y. Van de Peer, A. Wilmotte, *et al.* (1995). "An early origin of plastids within the cyanobacterial divergence is suggested by evolutionary trees based on complete 16S rRNA sequences." Mol Biol Evol **12**(6): 1166-1173.

Nesheim, J. C. and J. D. Lipscomb (1996). "Large kinetic isotope effects in methane oxidation catalyzed by methane monooxygenase: evidence for C-H bond cleavage in a reaction cycle intermediate." Biochemistry **35**(31): 10240-10247.

Nordlund, P. and P. Reichard (2006). "Ribonucleotide reductases." Annu Rev Biochem **75**: 681-706.

O'Connor, C. (2008). "Isolating Hereditary Material: Frederick Griffith, Oswald Avery, Alfred Hershey, and Martha Chase." Nature Education **1**(1).

Pace, C. N., B. A. Shirley, M. McNutt, *et al.* (1996). "Forces contributing to the conformational stability of proteins." Faseb J **10**(1): 75-83.

Parales, R. E., N. C. Bruce, A. Schmid, *et al.* (2002). "Biodegradation, biotransformation, and biocatalysis (b3)." Appl Environ Microbiol **68**(10): 4699-4709.

Percival, M. D. (1991). "Human 5-lipoxygenase contains an essential iron." J Biol Chem **266**(16): 10058-10061.

Pinto, A. F., S. Todorovic, P. Hildebrandt, *et al.* (2011). "Desulforubrerythrin from Campylobacter jejuni, a novel multidomain protein." J Biol Inorg Chem **16**(3): 501-510.

Rennex, D., R. T. Cummings, M. Pickett, *et al.* (1993). "Role of tyrosine residues in Hg(II) detoxification by mercuric reductase from Bacillus sp. strain RC607." Biochemistry **32**(29): 7475-7478.

Rexroth, S., C. W. Mullineaux, D. Ellinger, *et al.* (2011). "The Plasma Membrane of the Cyanobacterium Gloeobacter violaceus Contains Segregated Bioenergetic Domains." Plant Cell.

Riebe, O., R. J. Fischer, D. A. Wampler, *et al.* (2009). "Pathway for H_2O_2 and O_2 detoxification in Clostridium acetobutylicum." Microbiology **155**(Pt 1): 16-24.

Rosenzweig, A. C., H. Brandstetter, D. A. Whittington, *et al.* (1997). "Crystal structures of the methane monooxygenase hydroxylase from *Methylococcus capsulatus* (Bath): implications for substrate gating and component interactions." Proteins **29**(2): 141-152.

Rosenzweig, A. C., C. A. Frederick, S. J. Lippard, *et al.* (1993). "Crystal structure of a bacterial non-haem iron hydroxylase that catalyses the biological oxidation of methane." Nature **366**(6455): 537-543.

Sambrook, J., E. F. Fritsch and T. Maniatis, Eds. (1989). Molecular Cloning: A Laboratory Manual. Cold Spring Harbor, NY, Cold Spring Harbor Laboratory Press.

Sarma, G. N., C. Nickel, S. Rahlfs, *et al.* (2005). "Crystal structure of a novel *Plasmodium falciparum* 1-Cys peroxiredoxin." J Mol Biol **346**(4): 1021-1034.

Sayavedra-Soto, L. A., D. M. Doughty, E. G. Kurth, *et al.* (2005). "Product and product-independent induction of butane oxidation in *Pseudomonas butanovora*." FEMS Microbiol Lett **250**(1): 111-116.

Sayavedra-Soto, L. A., N. Hamamura, C. W. Liu, *et al.* (2011). "The membrane-associated monooxygenase in the butane-oxidizing Gram-positive bacterium *Nocardioides sp.* strain CF8 is a novel member of the AMO/PMO family." Environmental Microbiology Reports **3**(3): 390-396.

Sazinsky, M. H., J. Bard, A. Di Donato, *et al.* (2004). "Crystal structure of the toluene/*o*-xylene monooxygenase hydroxylase from *Pseudomonas stutzeri* OX1. Insight into the substrate specificity, substrate channeling, and active site tuning of multicomponent monooxygenases." J Biol Chem **279**(29): 30600-30610.

Sazinsky, M. H., P. W. Dunten, M. S. McCormick, *et al.* (2006). "X-ray structure of a hydroxylase-regulatory protein complex from a hydrocarbon-oxidizing multicomponent monooxygenase, *Pseudomonas* sp. OX1 phenol hydroxylase." Biochemistry **45**(51): 15392-15404.

Sazinsky, M. H. and S. J. Lippard (2005). "Product bound structures of the soluble methane monooxygenase hydroxylase from *Methylococcus capsulatus* (Bath): protein motion in the α-subunit." J Am Chem Soc **127**(16): 5814-5825.

Sazinsky, M. H. and S. J. Lippard (2005). "Product bound structures of the soluble methane monooxygenase hydroxylase from *Methylococcus capsulatus* (Bath): protein motion in the alpha-subunit." J Am Chem Soc **127**(16): 5814-5825.

Sazinsky, M. H. and S. J. Lippard (2006). "Correlating structure with function in bacterial multicomponent monooxygenases and related diiron proteins." Acc Chem Res **39**(8): 558-566.

Schirmer, A., M. A. Rude, X. Li, *et al.* (2010). "Microbial biosynthesis of alkanes." Science **329**(5991): 559-562.

Schlichting, I., C. Jung and H. Schulze (1997). "Crystal structure of cytochrome P-450cam complexed with the (1S)-camphor enantiomer." FEBS Lett **415**(3): 253-257.

Schneider, T. D. and R. M. Stephens (1990). "Sequence logos: a new way to display consensus sequences." Nucleic Acids Res **18**(20): 6097-6100.

Schrodinger, LLC (2010). The PyMOL Molecular Graphics System, Version 1.3r1.

Schwartz, J. K., P. P. Wei, K. H. Mitchell, *et al.* (2008). "Geometric and electronic structure studies of the binuclear nonheme ferrous active site of toluene-4-monooxygenase: parallels with methane monooxygenase and insight into the role of the effector proteins in O_2 activation." J Am Chem Soc **130**(22): 7098-7109.

Shanklin, J., C. Achim, H. Schmidt, *et al.* (1997). "Mossbauer studies of alkane ω-hydroxylase: evidence for a diiron cluster in an integral-membrane enzyme." Proc Natl Acad Sci U S A **94**(7): 2981-2986.

Shanklin, J., J. E. Guy, G. Mishra, *et al.* (2009). "Desaturases: Emerging Models for Understanding Functional Diversification of Diiron-containing Enzymes." J Biol Chem **284**(28): 18559-18563.

Shennan, J. L. (2006). "Utilisation of C2-C4 gaseous hydrocarbons and isoprene by microorganisms." J Chem Technol Biotechnol **81**(3): 237-256.

Shortle, D. and J. Sondek (1995). "The emerging role of insertions and deletions in protein engineering." Curr Opin Biotechnol **6**(4): 387-393.

Shu, L., J. C. Nesheim, K. Kauffmann, *et al.* (1997). "An $Fe_2^{IV}O_2$ diamond core structure for the key intermediate Q of methane monooxygenase." Science **275**(5299): 515-518.

Skulan, A. J., T. C. Brunold, J. Baldwin, *et al.* (2004). "Nature of the Peroxo Intermediate of the W48F/D84E Ribonucleotide Reductase Variant: Implications for O_2 Activation by Binuclear Non-Heme Iron Enzymes." J. Am. Chem. Soc. **126**(28): 8842-8855.

Sluis, M. K., L. A. Sayavedra-Soto and D. J. Arp (2002). "Molecular analysis of the soluble butane monooxygenase from *'Pseudomonas butanovora'*." Microbiology **148**(Pt 11): 3617-3629.

Smith, C. A., K. T. O'Reilly and M. R. Hyman (2003). "Characterization of the initial reactions during the cometabolic oxidation of methyl *tert*-butyl ether by propane-grown *Mycobacterium vaccae* JOB5." Appl Environ Microbiol **69**(2): 796-804.

Smith, T. J. and H. Dalton (2004). Biocatalysis by methane monooxygenase and its implications for the petroleum industry. Petroleum Biotechnology, Developments and Perspectives,. R. Vazquez-Duhalt and R. Qintero-Ramirez. Amsterdam, Elsevier**:** 177-192.

Smith, T. J. and J. C. Murrell (2011). "Mutagenesis of soluble methane monooxygenase." Methods Enzymol **495**: 135-147.

Soding, J. and A. N. Lupas (2003). "More than the sum of their parts: on the evolution of proteins from peptides." Bioessays **25**(9): 837-846.

Sondek, J. and D. Shortle (1992). "A general strategy for random insertion and substitution mutagenesis: substoichiometric coupling of trinucleotide phosphoramidites." Proc Natl Acad Sci U S A **89**(8): 3581-3585.

Song, B., N. J. Palleroni, L. J. Kerkhof, *et al.* (2001). "Characterization of halobenzoate-degrading, denitrifying *Azoarcus* and *Thauera* isolates and description of *Thauera chlorobenzoica* sp. nov." Int J Syst Evol Microbiol **51**(Pt 2): 589-602.

Stirewalt, V. L., C. B. Michalowski, W. Löffelhardt, *et al.* (1995). "Nucleotide sequence of the cyanelle genome from *Cyanophora paradoxa*." Plant Mol Biol Reptr **13**(4): 327-332.

Tabor, S. and C. C. Richardson (1985). "A bacteriophage T7 RNA polymerase/promoter system for controlled exclusive expression of specific genes." Proc Natl Acad Sci U S A **82**(4): 1074-1078.

Takahashi, J., Y. Ichikawa, H. Sagae, *et al.* (1980). "Isolation and identification of *n*-butane assimilating bacterium." Agric Biol Chem **44**: 1835–1840.

Thompson, J. D., D. G. Higgins and T. J. Gibson (1994). "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice." Nucleic Acids Res **22**(22): 4673-4680.

Tinberg, C. E. and S. J. Lippard (2011). "Dioxygen activation in soluble methane monooxygenase." Acc Chem Res **44**(4): 280-288.

Van Beeumen, J. J., G. Van Driessche, M. Y. Liu, *et al.* (1991). "The primary structure of rubrerythrin, a protein with inorganic pyrophosphatase activity from Desulfovibrio vulgaris. Comparison with hemerythrin and rubredoxin." J Biol Chem **266**(31): 20645-20653.

van Beilen, J. B. and E. G. Funhoff (2007). "Alkane hydroxylases involved in microbial alkane degradation." Appl Microbiol Biotechnol **74**(1): 13-21.

van Beilen, J. B., J. Kingma and B. Witholt (1994). "Substrate specificity of the alkane hydroxylase system of *Pseudomonas oleovorans* GPo1." Enzyme Microb. Technol. **16**: 904 - 911.

van Belkum, A., S. Scherer, L. van Alphen, *et al.* (1998). "Short-sequence DNA repeats in prokaryotic genomes." Microbiol Mol Biol Rev **62**(2): 275-293.

Vangnai, A. S. and D. J. Arp (2001). "An inducible 1-butanol dehydrogenase, a quinohaemoprotein, is involved in the oxidation of butane by *"Pseudomonas butanovora"*." Microbiology **147**(Pt 3): 745-756.

Vangnai, A. S., D. J. Arp and L. A. Sayavedra-Soto (2002). "Two distinct alcohol dehydrogenases participate in butane metabolism by *Pseudomonas butanovora*." J Bacteriol **184**(7): 1916-1924.

Vangnai, A. S., L. A. Sayavedra-Soto and D. J. Arp (2002). "Roles for the two 1-butanol dehydrogenases of *Pseudomonas butanovora* in butane and 1-butanol metabolism." J Bacteriol **184**(16): 4343-4350.

Velankar, S., C. Best, B. Beuth, *et al.* (2010). "PDBe: protein data bank in Europe." Nucleic acids research **38**(suppl 1): D308.

Voegtli, W. C., N. Khidekel, J. Baldwin, *et al.* (2000). "Crystal Structure of the Ribonucleotide Reductase R2 Mutant that Accumulates a μ-1, 2-Peroxodiiron (III) Intermediate during Oxygen Activation." J. Am. Chem. Soc. **122**(14): 3255-3261.

Wakagi, T. (2003). "Sulerythrin, the smallest member of the rubrerythrin family, from a strictly aerobic and thermoacidophilic archaeon, Sulfolobus tokodaii strain 7." FEMS Microbiol Lett **222**(1): 33-37.

Wallar, B. J. and J. D. Lipscomb (1996). "Dioxygen Activation by Enzymes Containing Binuclear Non-Heme Iron Clusters." Chem Rev **96**(7): 2625-2658.

Wallar, B. J. and J. D. Lipscomb (2001). "Methane monooxygenase component B mutants alter the kinetics of steps throughout the catalytic cycle." Biochemistry **40**(7): 2220-2233.

Walters, K. J., G. T. Gassner, S. J. Lippard, *et al.* (1999). "Structure of the soluble methane monooxygenase regulatory protein B." Proc Natl Acad Sci U S A **96**(14): 7877-7882.

Wang, G. and R. L. Dunbrack, Jr. (2003). "PISCES: a protein sequence culling server." Bioinformatics **19**(12): 1589-1591.

Weaver, T. M. (2000). "The π-helix translates structure into function." Protein Sci **9**(1): 201-206.

Werner, D. S., T. R. Lee and D. S. Lawrence (1996). "Is protein kinase substrate efficacy a reliable barometer for successful inhibitor design?" J Biol Chem **271**(1): 180-185.

Wiedenheft, B., J. Mosolf, D. Willits, *et al.* (2005). "An archaeal antioxidant: characterization of a Dps-like protein from *Sulfolobus solfataricus*." Proc Natl Acad Sci U S A **102**(30): 10551-10556.

Woodland, M. P. and H. Dalton (1984). "Purification and characterization of component A of the methane monooxygenase from Methylococcus capsulatus (Bath)." J Biol Chem **259**(1): 53-59.

Worth, C. L., S. Gong and T. L. Blundell (2009). "Structural and functional constraints in the evolution of protein families." Nat Rev Mol Cell Biol **10**(10): 709-720.

Xie, L. and W. A. van der Donk (2001). "Homemade cofactors: self-processing in galactose oxidase." Proc Natl Acad Sci U S A **98**(23): 12863-12865.

Xie, M., H. Alonso and A. Roujeinikova (2011). "An Improved Procedure for the Purification of Catalytically Active Alkane Hydroxylase from Pseudomonas putida GPo1." Appl Biochem Biotechnol.

Xu, H. and M. A. Freitas (2007). "A mass accuracy sensitive probability based scoring algorithm for database searching of tandem mass spectrometry data." BMC Bioinformatics **8**: 133.

Xu, H. and M. A. Freitas (2009). "MassMatrix: a database search program for rapid characterization of proteins and peptides from tandem mass spectrometry data." Proteomics **9**(6): 1548-1555.

Yamashita, A., S. K. Singh, T. Kawate, *et al.* (2005). "Crystal structure of a bacterial homologue of Na+/Cl--dependent neurotransmitter transporters." Nature **437**(7056): 215-223.

Zhao, W., Z. Ye and J. Zhao (2007). "RbrA, a cyanobacterial rubrerythrin, functions as a FNR-dependent peroxidase in heterocysts in protection of nitrogenase from damage by hydrogen peroxide in *Anabaena* sp. PCC 7120." Mol Microbiol **66**(5): 1219-1230.

Zheng, H. and J. D. Lipscomb (2006). "Regulation of methane monooxygenase catalysis based on size exclusion and quantum tunneling." Biochemistry **45**(6): 1685-1692.

Zhu, C. Z., B. Ouyang, J. Q. Wang, *et al.* (2007). "Photochemistry in the mixed aqueous solution of nitrobenzene and nitrous acid as initiated by the 355 nm UV light." Chemosphere **67**(5): 855-861.

Appendices

Appendix 1

Evolutionary origin of a secondary structure: π-helices as cryptic but widespread insertional variations of α-helices enhancing protein functionality - Supplementary Information

Richard B. Cooley, Daniel J. Arp and P. Andrew Karplus

Published in *Journal of Molecular Biology* (2010), 404(2), 232-246.

Table A1.1. α/π-helical homologous pairs for single π-helices. All 89 single π-helices in the dataset are listed. For the 73 having a homologous α-helical segment, the sequence alignment is shown with the sequence of the π-helical segment on the top and the sequence of the homologous α-helical segment on the bottom (with the gap position marked as a dot). Although the position of each insertion is identified from the FSSP database, this need not be the exact location the insertion took place given the mobility of π-helices within α-helices (e.g. Fig. 6). Bold residues are involved in a (i+5,i) H-bond. For underlined residues, the (i+5,i) interaction is the strongest of the (i+3,i), (i+4,i) and (i+5,i) interactions. The column Hb reports the number of π-type H-bonds in the π-helix. The remaining columns report the SCOP superfamily to which the homologous α/π-helical proteins belong, the Z-scores for structural similarity, the number of structurally aligned residues compared to the length of the shorter protein, and the percent sequence identity for the structurally aligned residues all as reported in the FSSP database. Here and in Table A1.2, the homologs shown are those having the highest Z-score.

Table A1.1 (continued)

PDB	Position	Sequence	Hb	SCOP superfamily	Z-Score	Aligned Residues/Total Residues	Seq. ID (%)
1a8e 1lst	124-130 120-125	**RS**AGW**NI** STQ.EAY	2	Periplasmic binding protein-like II	10.2	203/239	14
1a8i 1gz5	488-495 A229-235	**PRR**WL**VLC** PKEIA.KQ	3	UDP-Glycosyltransferase /glycogen phosphorylase	17.6	378/456	13
1b16 2nm0	A104-110 A94-99	**IA**INF**TG** ETN.LTG	2	NAD(P)-binding Rossmann-fold domains	18.9	198/221	21
1b25 1aor	A479-485 A483-488	**RL**RGG**LF** DLT.ALI	2	Aldehyde ferredoxin oxidoreductase, C-term. domain	56.3	589/605	39
1b5q 1s3b	A74-80 A70-75	**PI**VNS**TL** RLAK.EL	2	FAD/NAD(P)-binding domain	31.8	417/494	18
1bdb Not yet known	112-118	**FH**INV**KG**	2	NAD(P)-binding Rossmann-fold domains			
1bdm 1a5z	A217-223 224-229	**WY**EKV**FI** ILE.NFA	2	NAD(P)-binding Rossmann-fold domains	32.6	289/312	22
1bg6 2hdh	297-303 A258-263	**DV**STG**LV** TKF.IVD	2	6-P-gluconate dehydrogenase C-terminal domain-like	14.6	233/293	11
1bxk 2vrb	A98-104 A80-85	**IE**TNI**VG** TLL.IVQ	2	NAD(P)-binding Rossmann-fold domains	19.8	243/284	12
1c3p 1woh	A97-103 A92-97	**YA**MFT**GS** EP.QLAH	2	Arginase/deacetylase	11.6	212/303	10
1c3w 1f88	A213-219 B297-302	**VS**AKV**GF** T.SAVYN	2	Family A G protein-coupled receptor-like	11.2	184/305	11
1c7s 1o7a	A801-807 A528-533	**IL**GQR**EL** RLT.RHR	2	(Trans)glycosidases	36.4	452/483	23
1cb8 1hn0	A267-274 A541-547	**YYR**DS**YLK** AMV.SAWI	3	Chondroitin AC/alginate lyase	43.2	625/971	19
1coj Not yet known	A26-33	**PHF**EA**HYK**	3	Fe,Mn superoxide dismutase (SOD), N-terminal domain			
1cxp Not yet known	C291-297	**IT**YRD**YL**	2	Heme-dependent peroxidases			
1cyd Not yet known	A104-110	**FS**VNL**RS**	2	NAD(P)-binding Rossmann-fold domains			
1d3g 1f76	A37-43 A2-6	**FY**AEH**LM** YYP.FVR	2	FMN-linked oxidoreductases	46.7	319/336	40
1d3y 2zbk	A253-260 G265-271	**YGY**LN**IYR** YGW.YIFS	3	DNA topoisomerase IV, alpha subunit	28.4	277/347	34
1d8d 1dce	A343-349 A410-415	**IL**AKE**KD** TLK.AVD	2	Protein prenylyltransferase	17.1	284/567	21
1dc1 Not yet known	A98-104	**LI**ENF**LE**	2	Restriction endonuclease-like			
1dek 1shk	A136-145 A45-54	**QALGTDLIVN** VADVVAAE.G	5	P-loop containing nucleoside triphosphate hydrolases	7.5	125/158	13
1dj0 1vmb	A81-87 A79-84	**AA**WTL**GV** QELEN.F	2	Pseudouridine synthase	7.3	83/107	7
1dk8 1ezt	A242-249 A126-132	**DIY**LE**YTR** RFYL.DLT	3	Regulator of G-protein signaling, RGS	16.0	126/129	31
1dqa 1qax	A732-740 A252-259	**NINK**N**LVGS** YAFA.AVDP	4	Substrate-binding domain of HMG-CoA reductase	29.6	316/425	20
1dqs 1xag	A142-148 A127-132	**LA**MVD**SS** LAH.DSS	2	Dehydroquinate synthase-like	44.6	341/353	32
1dqz 1c4x	A97-103 A90-95	**TF**LTR**EM** EQIL.GL	2	alpha/beta-Hydrolases	16.1	215/281	14
1dxr Not yet known	C277-283	**LN**MNY**LA**	2	Multiheme cytochromes			
1dxr 1eys	H27-33 H26-31	**TVV**LLY**LR** GLII.YLR	3	Photosystem II reaction centre, subunit H	26.0	235/238	53
1dxr 1jb0	L129-135 609-614	**CV**LQV**FR** VIFHF.S	2	Photosystem II reaction centre, L/M subunits	7.9	172/740	8
1dxr 1jb0	M156-162 609-614	**LC**IGC**IH** VIFHF.S	2	Photosystem II reaction centre, L/M subunits	8.1	168/740	7

Table A1.1 (continued)

PDB	Position	Sequence	Hb	SCOP superfamily	Z-Score	Aligned Residues/Total Residues	Seq. ID (%)
1dys 1qk0	A112-118 A195-200	**YK**NEY**VN** YKN.YID	2	Glycosyl hydrolases family 6, cellulases	48.2	338/363	36
1dz4 1jpz	A150-156 B141-146	**FT**EDY**AE** VPE.DMT	2	Cytochrome P450	30.2	372/458	16
1e3a Not yet known	A138-144	**FV**GTM**AN**	2	N-terminal nucleophile aminohydrolases			
1ea5 1f0n	A396-402 A230-235	**GD**HNV**IC** EN.FVRS	2	alpha/beta-Hydrolases	13.6	223/284	9
1ea5 1mx9	A522-528 1545-1550	**VF**WNQ**FL** AFW.TNL	2	alpha/beta-Hydrolases	56.0	433/531	37
1egu 3e7j	A292-298 A148-153	**WD**YEI**GT** RGI.GLF	2	Chondroitin AC/alginate lyase	17.4	274/743	5
1egu 3e7j	A441-447 A288-293	**WI**DKS**FA** ILY.DAI	2	Chondroitin AC/alginate lyase	17.4	274/743	5
1ei5 1e25	A177-183 A148-153	**LS**ERI**FA** LHD.YIQ	2	beta-lactamase/transpeptidase-like	14.3	210/278	12
1ek6 2p4h	A105-111 X101-107	Y**YR**VN**LTG** TKR.TVDG	2	NAD(P)-binding Rossmann-fold domains	26.2	285/310	18
1el4 1uhi	A44-52 A38-45	**SKAS**D**DICA** YKAS.DIVI	4	EF-hand	31.1	187/191	70
1elk 1jpl	A95-101 A90-95	**VE**SVL**VR** LNE.LIK	2	ENTH/VHS domain	19.8	144/161	34
1eok 1waw	A111-118 A124-130	**KSI**VI**DKW** IRFLR.KY	3	(Trans)glycosidases	19.1	229/366	15
1evy 2d3t	A257-263 B656-661	**GL**AGL**GD** MG.LVYG	2	6-P-gluconate dehydrogenase C-terminal domain-like	13.7	162/708	9
1ewf 2obd	A181-186 A190-195	**SV**SSE**LQ** ISN.IMA	2	Bactericidal permeability-increasing protein	25.7	421/472	19
1eyz 1b6s	A119-125 B85-90	**RL**AAE**EL** QLFD.KL	2	Glutathione synthetase ATP-binding domain-like	33.7	324/355	23
1f24 1izo	A140-146 A134-139	**LV**KEF**AL** LFE.EAK	2	Cytochrome P450	35.6	366/411	15
1f24 1n40	A214-220 A206-211	**LC**TEQ**VK** LSR.LRK	2	Cytochrome P450	40.6	374/394	26
1f3a 1okt	A126-132 A127-132	**RT**KNR**YL** DLP.KWS	2	GST C-terminal domain-like	24.0	198/211	32
1fds 2nm0	111-117 A97-102	**LD**VNV**VG** ET.NLTG	2	NAD(P)-binding Rossmann-fold domains	22.8	193/221	30
1fp3 2gz6	A273-279 A264-269	**ID**TFL**LL** VDV.VLN	2	Six-hairpin glycosidases	52.7	370/372	38
1fqa 1urg	A279-285 A305-310	**FL**ENY**LL** LVQ.ALT	2	Periplasmic binding protein-like II	39.6	357/373	30
1frp 2hhm	A276-282 A217-222	**RL**LYE**CN** HC.WDVA	2	Carbohydrate phosphatase	19.4	219/272	12
1fsw 1g6a	A174-180 A148-153	**MQ**TRV**FQ** VTD.FLR	2	beta-lactamase/transpeptidase-like	16.0	212/266	15
1fur 1j3u	A155-161 A156-161	**AL**RKQ**LI** LLN.QLI	2	L-aspartase-like	44.8	455/462	40
1fur 1i0a	A383-389 A346-351	**FN**KHC**AV** ATG.VIS	2	L-aspartase-like	29.8	390/453	17
1g8k 1kqg	A181-187 A179-184	**KL**MFS**AI** KFAR.SL	2	Formate dehydrogenase/DMSO reductase, domains 1-3	28.8	352/982	19
1g8k 2nap	A242-248 A184-189	**YF**LNH**WL** LFR.RIA	2	Formate dehydrogenase/DMSO reductase, domains 1-3	33.6	645/720	23
1gai 1ayx	150-156 181-186	**AA**TEI**VW** STE.DIY	2	Six-hairpin glycosidases	45.3	420/492	37
1hfe Not yet known	S71-77	**LY**KSY**LE**	2	Fe-only hydrogenase smaller subunit			
1hvb 1e25	A183-189 A148-153	**YQ**NRI**FT** LHDY.IQ	2	beta-lactamase/transpeptidase-like	17.5	216/278	18

Table A1.1 (continued)

PDB	Position	Sequence	Hb	SCOP superfamily	Z-Score	Aligned Residues/Total Residues	Seq. ID (%)
1i0h Not yet known	A25-32	**HH**TK**HHQ**	3	Insulin-like			
1kve Not yet known	B177-183	**LA**KNG**FK**	2	Yeast killer toxins			
1lml 1slm	155-161 112-117	**IL**VKH**LI** AVD.SAV	2	Metalloproteases ("zincins"), catalytic domain	8.4	122/226	16
1lst 1wdn	126-132 A124-129	**AN**DNW**RT** AKANI.K	2	Periplasmic binding protein-like II	29.9	222/224	30
1lst 1wdn	165-171 A161-166	**AS**EGF**LK** ILY.FIK	2	Periplasmic binding protein-like II	29.9	222/224	30
1mty 1t0r	B296-305 A239-247	**DLYYNCLGD** PSLK.ILVE	5	Ferritin-like	22.7	313/391	12
1mty 1h0o	D185-191 A208-218	**KR**VFS**DG** ADWA.LR	2	Ferritin-like	19.9	250/288	12
1muc 1jpd	A70-76 X61-66	**NI**DAH**LA** QIM.SVV	2	Enolase N-terminal domain-like	35.4	313/318	25
1one 1yey	A67-73 C66-71	**NV**NDV**IA** AVA.ALA	2	Enolase N-terminal domain-like	20.2	293/435	17
1phn 1jeb	A107-113 D81-86	**MD**EYL**IA** LKGT.FA	2	Globin-like	9.5	119/146	14
1qgw 1uc3	C105-111 A94-99	**LE**DRC**LN** MSM.LKR	2	Globin-like	9.5	118/149	7
1qh3 1a8t	A154-161 A187-193	**KAL**LE**VLG** KTLDK.VK	3	Metallo-hydrolase/oxidoreductase	16.8	183/230	20
1qh8 1mio	A63-72 D25-33	**AGSKGVVFG** VGAM.YAAL	5	"Helical backbone" metal receptor	30.6	406/457	18
1qlm Not yet known	A86-94	**AVST**LA**AQK**	4	Methenyltetrahydromethanopterin cyclohydrolase			
1qoy 2nrj	A25-32 A27-33	**DLY**NK**YLD** QKA.GLFA	3	Bacterial hemolysins	13.7	238/338	6
1smd 1mxg	27-33 A33-38	**EA**ERY**LA** KIP.EQY	2	(Trans)glycosidases	28.9	364/435	17
1sur 1j1z	131-137 A92-97	**ND**INK**VE** RP.LIAK	2	Adenine nucleotide alpha hydrolases-like	6.7	117/390	15
1svf Not yet known	A171-177	**HI**NSV**VS**	2	Virus ectodomain			
1thg 1din	424-430 A176-181	**SD**MLF**QS** APS.RQL	2	alpha/beta-Hydrolases	12.2	190/232	14
1uok 2zic	393-399 A377-382	**KE**KVM**ER** KEAFT.N	2	(Trans)glycosidases	53.5	531/536	53
1yac 2r8y	A114-120 F125-130	**VT**EVC**VA** DDLI.DW	2	Isochorismatase-like hydrolases	7.2	126/172	8
1yge 2p0m	261-267 B177-182	**SL**SQI**VQ** ASLA.WG	2	Lipoxigenase	37.8	638/662	22
1yge Not yet known	684-690	**WI**ASA**LH**	2	Lipoxigenase			
2hmq Not yet known	A100-107	**NHI**KT**IDF**	3	Hemerythrin-like			
2olb Not yet known	A301-308	**IIV**NK**VKN**	3	Periplasmic binding protein-like II			
2scp 1nya	A56-62 A59-64	**VW**DNF**LT** LFD.YLA	2	EF-hand	13.0	165/176	28
4pah Not yet known	325-331	**YW**FTV**EF**	2	Aromatic aminoacid monoxygenases, catalytic and oligomerization domains			
5csm 1ecm	A233-239 B76-81	**IY**KEI**VI** LFQ.LII	2	Chorismate mutase II	6.4	92/95	18
7a3h 1g0c	A146-152 A390-395	**TW**GNQ**IK** GWE.AVK	2	(Trans)glycosidases	45.8	297/357	44

Table A1.2. α/π-helical homologous pairs for overlapping π-helices. All 17 multiple overlapping π-helices in the dataset are listed. For the 15 having a homologous α-helical segment, the sequence alignment is shown with both a pure α-helical homolog and additional homologs representing the stepwise intermediates in the evolutionary history. The contents of each column are as described in Table A1.1.

PDB	Position	Sequence	Hb	SCOP superfamily	Z-Score	Aligned Residues/Total Residues	Seq. ID (%)
1qmg	A349-355	**YK**NTV**EC**	2	6-P-gluconate dehydrogenase C-terminal domain-like			
	A355-361	**CI**TGV**IS**	2				
1np3	B224-235	**YF**ECL**HE**L.KLIV			30.0	317/327	27
3d1l	B203-214	.MLPLDETAR.KV			14.9	230/259	11
1qmg	A490-496	**IN**ESV**IE**	2	6-P-gluconate dehydrogenase C-terminal domain-like			
	A496-502	**EA**VDS**LN**	2				
1yrl	A387-398	**YY**ESL**HE**LP.LIA			35.7	432/487	22
1ie3	A205-215	RIQN.AGTE.VVE			6.2	148/312	12
1mro	A313-321	**MLYD**Q**IWLG**	4	Methyl-coenzyme M reductase alpha and beta chain C-terminal domain			
	A318-324	**IW**LGS**YM**	2				
1e6v	B218-228	TLQQ.**TA**MFE**MG**			23.9	375/436	14
1e6y	E5214-5223	IYE.QSGI.FEM			24.6	371/433	14
1mty	B140-146	**WR**DEF**IN**	2	Ferritin-like			
	B144-150	**FI**NRY**WG**	2				
2inp	C115-124	TQ.K**QL**LRL**LV**			31.8	309/318	14
1t0r	A89-97	WV.STMQ.LHF			22.5	313/491	12
1mty	B247-256	SAF**SV**HAV**YD**A	2	Ferritin-like			
	B252-258	**YD**ALF**GQ**	2				
	B258-266	FG**QFVR**R**EFFQ**	4				
1syy	A184-191	**NLV**GY**YII**ME			19.1	273/317	8
1syy	A191-205	**IM**EGI**FF**YSGF.VMIL					
1uzr	A155-175	RKVAS.TL**LE**SFL**FY**SGF.YLPM			18.3	258/282	7
1bfr	A89-105	MLRSD.LALEL.DGAKNL.REAI			9.3	143/158	13
1mty	D201-210	**SLNLQLVGEA**	5	Ferritin-like			
	D208-214	**GE**ACF**TN**PLIV	2				
1uzr	A156-172	KVAST.LLES**FLF**YS**GFY**			19.0	254/282	11
1bfr	A87-98	LRSDL.ALEL.DGA			10.2	141/158	13
1mty	D306-314	TW**DRWV**Y**EDWG**	4	Ferritin-like			
	D311-318	**EDW**GG**IWI**	3				
2inp	A294-307	**AWEV**Y**YEQN**G.GALF			36.2	459/494	20
1h0o	A323-333	MKQYIEF.VADR.LM			19.5	254/288	12
1yge	494-502	**HQLM**S**HWLN**	4	Lipoxigenase			
	500-506	**WL**NTH**AA**	2				
Not yet known							

Table A1.3. π-helices found in the Protein Data Bank using different qualifying criteria

No. π-type H-bonds [a]	25%, WW [b]	25%, MW [b]	25%, MM [b]	25%, I [b]	90%, WW [b]
2 HB	827	822	721	35	2418
3 HB	138	138	131	7	445
4 HB	26	26	26	1	57
5 HB	20	20	20	1	40
6 HB	3	3	3	0	4
7 HB	1	1	1	0	3
Total	1015	1010	902	44	2967
No. Chains analyzed	5620	5620	5620	5620	14,967

[a] The number of qualifying (i+5) → i hydrogen bonds in the π-helix
[b] The maximum percent identity between any two PDB chains and the minimal criteria needed for a potential π-helix to be counted. These criteria are designated WW (at least two sequential π-type hydrogen bonds with strengths ≤ -0.5 kcal/mol), MW (at least two sequential π-type hydrogen bonds where the strength of one is ≤ -1.0 kcal/mol and the other is ≤ -0.5 kcal/mol), MM (at least two sequential π-type hydrogen bonds with strengths ≤ -1.0 kcal/mol) and I (only those formally designated by DSSP as π-helical).

Table A1.4. Number of π-helices found per protein chain using 90% sequence identity cutoff and the WW criteria from structures determined at or better than 2.5 Å resolution.

Number of π-helices per PDB chain	no. PDB chains found	% of total	% of chains with π-helix
0	12523	84	-
1	2052	14	84
2	314	2	13
3	54	0.3	2
4	7	0.05	0.3
5	10	0.07	0.4
6	3	0.02	0.1
7	3	0.02	0.1
8	1	0.01	0.04
Total	**14967**	**100**	**100**

Table A1.5: Specific π-helix positions of insertions πA through πM in a set of structurally known representatives of the ferritin-like superfamily. Abbreviations used are RNR: R2 subunit of ribonucleotide reductase, PH: phenol hydroxylase, ArH: aromatic hydroxylase, and MMOH: soluble methane monooxygenase hydroxylase. The latter three are divided further into their respective α and β chains.

Insertion	Eukaryotic Ferritin	Rubrerythrin	RNR Ia/Ib	RNR Ic	PH		ArH		MMOH	
					α	β	α	β	α	β
πA_1	-	96-102	-	-	-	-	-	-	-	-
πA_2	-	-	-	-	-	-	-	-	-	-
πB	-	-	161-168	191-197	196-202	228-235	196-202	231-238	208-214	252-258
πC	-	-	74-80	95-101	-	-	-	-	-	-
πD	-	-	-	185-194	191-199	216-223	188-195	219-227	201-210	247-253
πE	-	-	-	-	295-302	-	294-300	-	305-314	-
πF	-	-	-	-	-	-	176-182	-	185-191	-
πG	-	-	-	-	-	-	-	-	311-318	-
πH	-	-	-	-	-	-	-	-	81-87	-
πI	-	-	-	-	-	118-124	-	120-126	-	143-150
πJ	-	-	-	-	-	-	-	-	-	258-266
πK	-	-	-	-	-	-	-	-	-	143-150
πL	-	-	-	-	-	-	-	-	-	296-305
πM	133-139	-	-	-	-	-	-	-	-	-
PDB code	**1lb3**	**1lkm**	**1uzr**	**1syy**	**2inp**		**1t0r**		**1mty**	

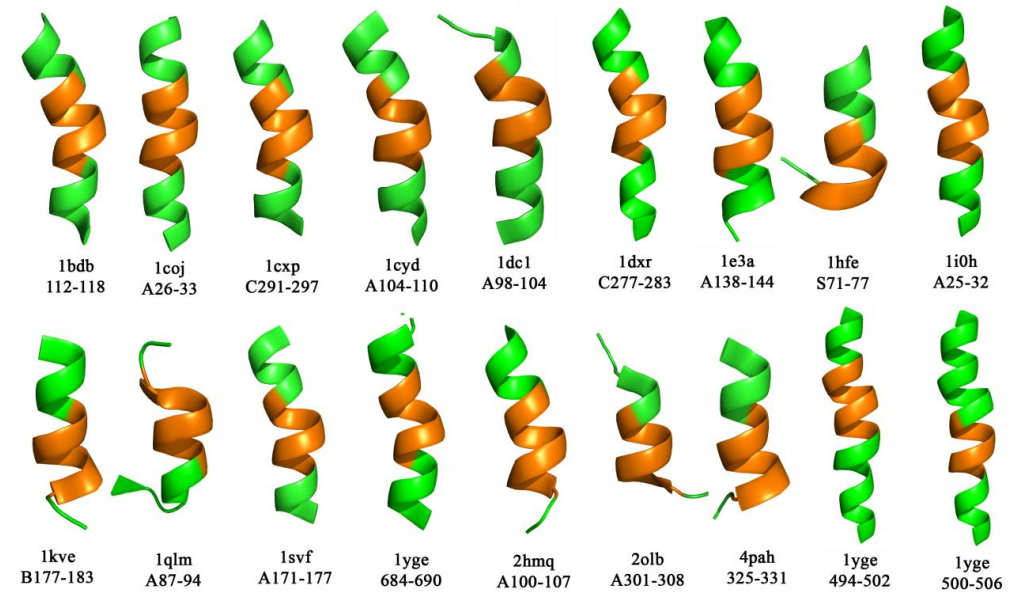


Figure A1.1. Of the π-helices in the dataset without an identified homolog containing an equivalent α-helix, all 18 are internal to α-helices. For each π-helix, the context (green) as well as the π-helix (orange) are shown. PDB codes and the π-helical residue ranges are written directly below each structure. The first 16 are non-overlapping π-helices from Table A1.1, and the last two are individual but overlapping π-helices from Table A1.2.

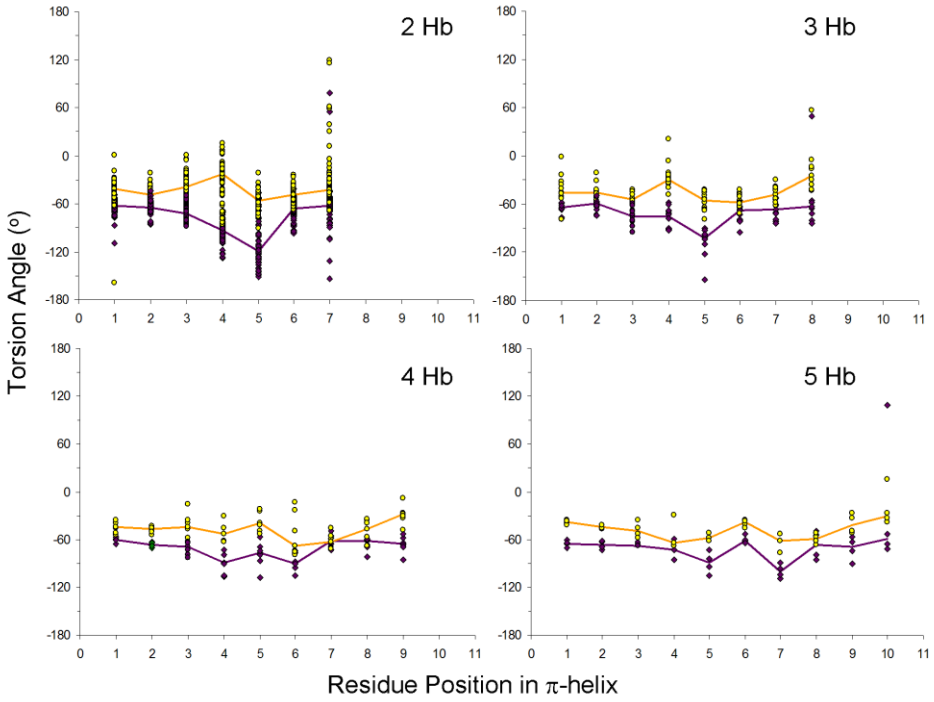


Figure A1.2. The range of ϕ,ψ-torsion angles for 2-, 3-, 4-, and 5-H-bonded π-helices demonstrates the variety of backbone conformations that can result from the accommodation of a single residue insertion. ϕ (purple), ψ (yellow) are plotted for each residue position in the π-helix with position 1 being the first residue in the π-helix. The lines connect the medians of each set of angles. The 2-H-bond π-helices show the typical (ϕ,ψ) pattern of an α-aneurism (Keefe *et al.* 1993), and as such do not possess a single characteristic (ϕ,ψ) value. Rather, the torsion angles of positions 4 and 5 are distorted relative to those at positions 1-3 and 6-7. This makes characterizing π-helices with a single (ϕ,ψ) value highly misleading. For the helices with 3-, 4-, and 5-H-bonds, the distortion is spread over more residues so that individual ϕ,ψ angles are less different from these of a pure α-helix.

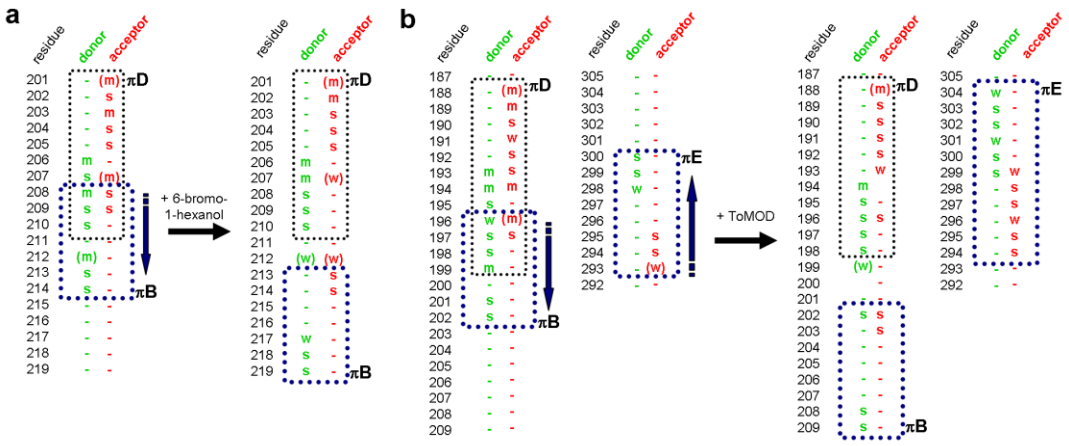


Figure A1.3. π-type H-bond patterns showing the π-helical peristaltic shifts seen in MMOH and toluene-4-monooxygenase (ToMO). (a) The π-type H-bonding patterns of MMOH show the active site contains two distinct, but overlapping π-helices (structures shown in Fig. 4.6a). Upon binding of 6-bromohexan-1-ol, the C-terminal π-helix (blue box) shifts in a peristaltic-like manner in the C-terminal direction so that these two π-helices are no longer overlapping. (b) The π-type H-bond patterns in the active site of ToMO upon regulatory subunit (ToMOD) binding show the same movement as MMOH in panel a, and also reveal that a third π-helix on residues 293-300 simultaneously shifts and expands in the C-terminal direction to occupy 294-304 (structures shown in Fig. 4.6b). It is interesting to note that this 6-H-bond π-helix and the 7-H-bond π-helix seen for residues 188-199 in the left side of panel b are derived from insertions πE and πD of Figure 7a, respectively, demonstrating that a π-helix based on a single insertion can even extend to 6- and 7-H-bonds. The π-helical conformation observed in the active site of the related phenol hydroxylase (PH) complexed with its regulatory component was reported to also be similar to that of 6-bromohexan-1-ol bound MMOH (right side of panel a) (Sazinsky *et al.* 2006). However, closer inspection shows the π-type H-bond positions in the PH-regulatory subunit complex match those of holo MMOH (left side of panel a). This could indicate that the energetics of structural changes of PH upon regulatory subunit binding may be somewhat different. Blue arrows show the directions of π-helix movement. The red and green letters indicate the strengths of donor and acceptor (i+5,i) interactions, respectively, as defined in Fig. 4.2.

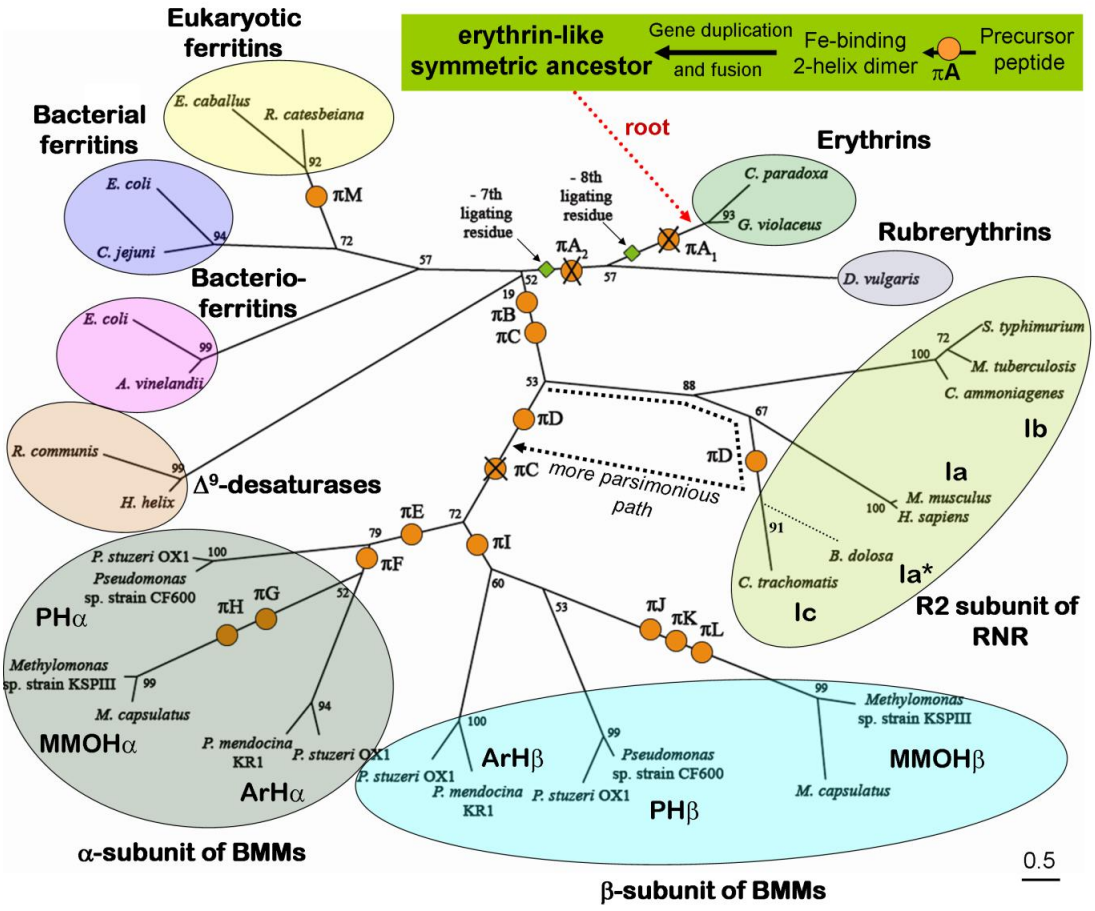


Figure A1.4. Maximum likelihood phylogenetic tree identical to that of Fig. 4.7a from the main text, but enlarged and containing the organism names from which the ferritin-like protein sequences were selected. Orange circles indicate insertion events that resulted in a π-helix while crossed out orange circles represent a deletion event that resulted in the loss of a π-helix. The length of the bar (bottom right) corresponds to 0.5 substitutions/site. Abbreviations are identical that of Fig. 4.7 from the main text.

Appendix 2

A diiron protein autogenerates a Valine-Phenylalanine crosslink - Supplementary Information

Richard B. Cooley, Timothy W. Rhoads, Daniel J. Arp and P. Andrew Karplus

Published in *Science* (2011), **332**(6032), 929.

Materials and Methods

Purification and structure determination: Briefly, ORF180 from the cyanelle genome of *C. paradoxa* (NCBI Gene ID: 801647) was purchased from GenScript (Piscataway, NJ) and ligated into the plasmid pT7-7. Expression was induced by IPTG in *E. coli* BL21(DE3) grown in LB media and 100 μg/ml ampicillin. Upon induction, the media was supplemented with 150 μM $FeSO_4$. Purification used DEAE anion exchange and Superdex S-75 size exclusion chromatography. Crosslinked symerythrin required an additional phenyl-sepharose hydrophobic interaction column. Protein presence and purity were assessed by SDS-PAGE with bands identified by LC-MS/MS. Yields were ~5 mg/l culture non-crosslinked and ~0.5 mg/l culture crosslinked protein. Proteins were stored frozen at ~30 mg/ml in 10 mM MOPS pH 7.2 (Buffer1). Protein concentrations were determined by BCA assay (Pierce) using a BSA standard. A summary of crystallographic information is in Table S1; coordinates and structure factors are deposited as Protein Data Bank entry 3qhb. Further details of the purification and crystallography will be published elsewhere.

Iron reconstitution: Non-crosslinked symerythrin purified with ~0.4-0.5 Fe/monomer (by ferrozine assay (Percival 1991)). Iron reconstitution involved anaerobic incubation in 25 mM MOPS pH 7.2, 150 mM NaCl pH 7.2 (Buffer2) for 3 h at room temperature with a 5-fold excess $Fe(NH_4)_2(SO_4)_2$ and 10-fold excess dithionite followed by dialysis into Buffer1 to leave 1.6-1.8 Fe/monomer. Crosslinked symerythrin purified with 1.7-1.8 Fe/monomer.

Crosslink formation. Spinach $NADP^+$-ferredoxin oxidoreductase (0.1 μM), spinach ferredoxin (1 μM), bovine catalase (0.1 mg/ml) and non-crosslinked Fe-reconstituted symerythrin (20 μM) were incubated in Buffer2 with 1 mM NADPH at 30 ^oC. Time point aliquots were quenched with 2% SDS and analyzed by SDS-PAGE.

Iron chelation. Symerythrin (20 μM) in Buffer2 was reduced with 5 mM dithionite and 0.1 mM methylviologen under argon. Reactions were initiated by the addition of 1 mM 1,10-phenanthroline. The accumulation of $[Fe(o\text{-phen})_3]^{2+}$ was monitored spectrophotometrically at 505 nm.

Mass Spectrometry. For proteolytic digestions, TCEP and iodoacetamide were used to reduce and block cysteine thiols. In-gel and in-solution digests were performed at 37 ^{o}C in 50 mM Tris pH 8.0 with 15 ng/μl modified porcine trypsin and 0.05% ProteaseMax (Promega) for 3 h. After trypsin inactivation with 5-fold excess TLCK (Sigma), digestion continued overnight after adding 2.5 mM $ZnSO_4$ and 10 ng/μl AspN (New England Biolabs). Reactions were quenched with 0.5% TFA. Digests (2 μL) were analyzed on an LTQ-FT Ultra mass spectrometer (ThermoFisher, San Jose, CA) with an IonMax source. Peptides were separated on a C_{18} column (Agilent Zorbax 300SB-C_{18}, 250x0.3 mm, 5 μm) using a linear gradient from 10% to 30% acetonitrile in 1% formic acid at a flow of 4 μl/min. The mass spectrometer alternated between full FT-MS scans (m/z 350-2000) and CID MS/MS (helium gas, normalized collision energy 35%, activation time 30 ms) scans performed in parallel on the five most abundant ions. Data acquisition was controlled by Xcalibur (ThermoFisher). Mascot Generic Format files generated by Proteome Discoverer (ThermoFisher) were used to search the Uniprot Plant protein database using MassMatrix (v. 1.0.1, Chicago, IL) (Xu and Freitas 2007; Xu and Freitas 2009).

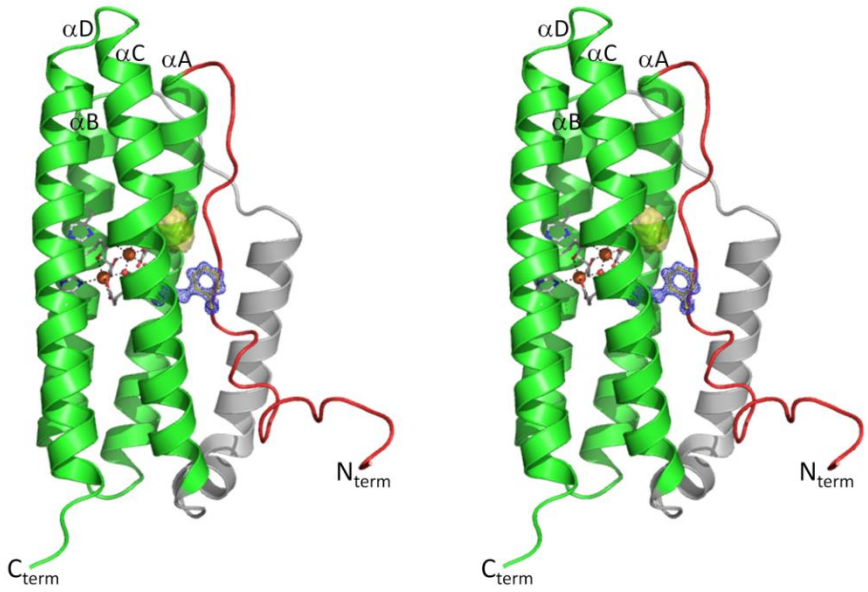


Figure A2.1. Stereo view of the symerythrin four-helix bundle showing the diiron center, the Val-Phe crosslink with $2F_o$-F_c density contoured at 2.2 ρ_{rms}, the N-terminal tail and the putative substrate binding pocket. Helices αA-αD are labeled. Coloring is the same as Fig. 5.1A.

Table A2.1. Crystallographic information of diferric symerythrin [a]

Data collection [b]	
Space group	$P3_2$
Unit cell axes (Å)	a=b=81.47, c=46.26
Resolution Limits (Å)	30.57-1.20 (1.25-1.20)
Unique Observations	107,444 (15,684)
Multiplicity	17.8 (10.7)
Average I/σ	23.4 (6.3)
R_{meas} [c] (%)	8.1 (35.2)
Completeness (%)	100.0 (100.0)
Refinement	
R_{cryst} / R_{free} (%)	10.1/12.5
No. protein molecules	2
No. protein residues	358
No. water molecules	555
Total number atoms	3543
rmsd bond angles (^{o})	3.6
rmsd bond lengths (Å)	0.019
<B> protein (Å^2)	12.9
<B> water (Å^2)	26.2
(ϕ,ψ)-Favored/Outliers (%) [d]	97.2/0.0
PDB code	3qhb

[a] Crystals were grown in hanging drops with a 2:1 mixture of protein stock (8 mg/ml in 10 mM Tris pH 7.2) to reservoir (22-28% PEG 3350, 0.1 M Bis-Tris pH 5.5, 0.35-0.4 M NH_4Cl) at 6 ^{o}C. Crystals (400 µm x 75 µm x 75 µm) were observed in 4-10 days.
[b] Numbers in parentheses correspond to values in the highest resolution bin
[c] R_{meas} is the multiplicity-weighted merging R-factor (Diederichs and Karplus 1997)
[d] Ramachandran plot generated using Molprobity (Davis *et al.* 2007)

Appendix 3

Symerythrin structures at atomic resolution and the origins of rubrerythrins and the ferritin-like superfamily - Supplemental Information

Richard B. Cooley, Daniel J. Arp and P. Andrew Karplus

Submitted to *Journal of Molecular Biology*

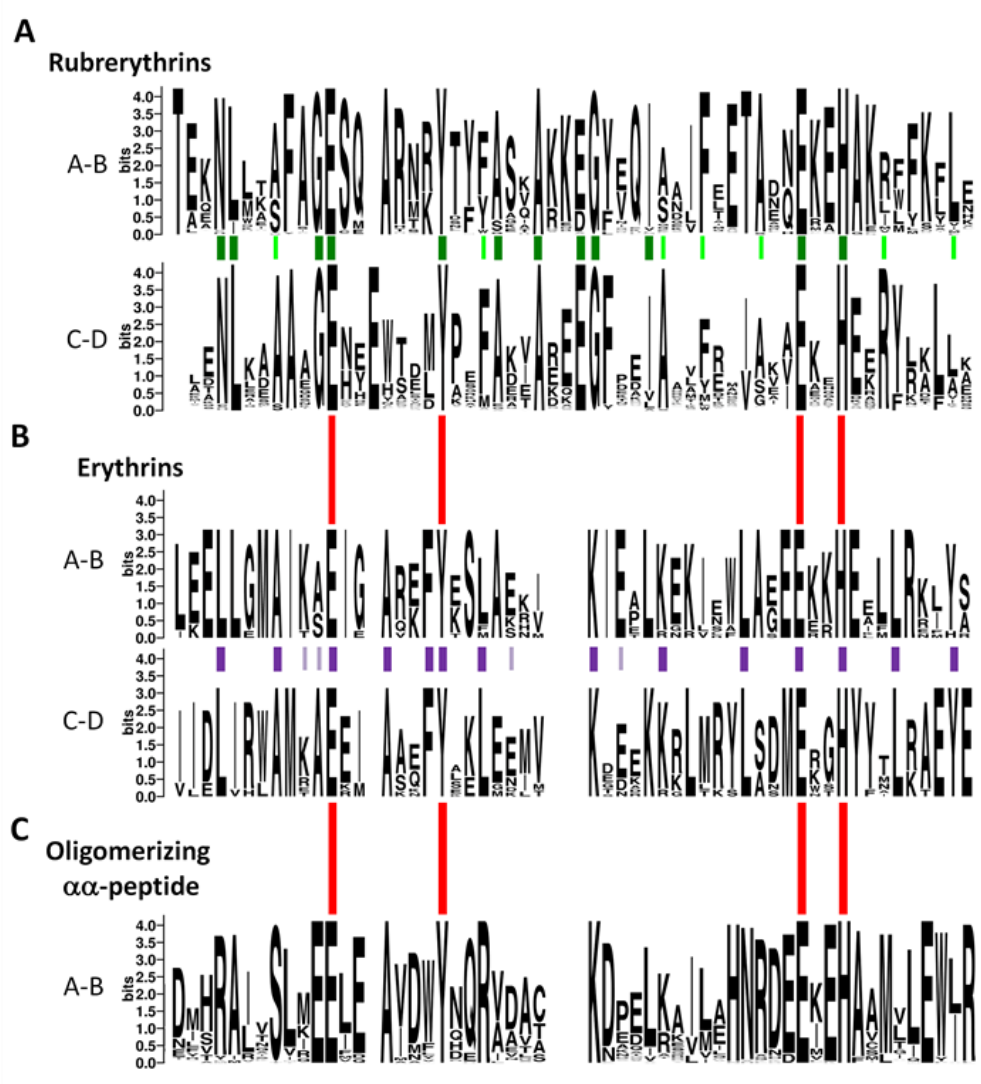


Figure A3.1. Internal alignments of conservation patterns of helix pairs A-B and C-D. Shown are SeqLogo (Schneider and Stephens 1990) plots generated from the alignment of multiple sequences belonging to the (a) rubrerythrin, (b) erythrin, and (c) oligomerizing αα-peptide families. Red lines connect metallocenter residues across families. Thick green and purple lines indicate highly conserved residues that are internally symmetric within the rubrerythrin and erythrin families, respectively. Thin green and purple lines designate symmetric pairs within the rubrerythrin and erythrin families, respectively, for which one residue is moderately conserved and one is either moderately or highly conserved. How the multiple sequence alignments were generated and the criteria for determining highly and moderately conserved residues for generating the consensus sequences of each family are as described in the *Materials and Methods*.