THE EVOLUTION OF BIOLOGICAL & GEOCHEMICAL ELECTRON TRANSFER REACTIONS

BY JOHN DONGUN KIM

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Chemistry and Chemical Biology

Written under the direction of

Paul G. Falkowski

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2013

ABSTRACT OF THE DISSERTATION

The Evolution of Biological & Geochemical Electron Transfer Reactions

by JOHN DONGUN KIM

Dissertation Director: Paul G. Falkowski

The chemistry of life is characterized by a set of oxidation-reduction (redox) reactions far from thermodynamic equilibrium. In this thesis, I examine geological and biological electron transfer systems that would have enabled the existence of self-replicating and energy-transducing organic matter, the origin of life. In chapter 1, I give a brief overview of the topic, which is covered in greater detail in chapters 2,3 and 4. In chapter 2, I demonstrate that photo-induced abiological electron transfer reaction of iron-bearing minerals could have contributed to the oxidation of Earth's atmosphere in the Precambrian and provided reductants for early organisms. In chapter 3, I examine a probable evolutionary pathway of early metalloenzymes. Using bacterial ferredoxins as an example, I show how early protoenzymes, capable of electron transfer reactions, may have evolved from iron-sulfur mineral surfaces. In chapter 4, I reveal structural relationships across metal-binding protein folds that are capable of electron transfer reactions.

Prior Publications

Several sections of this dissertation have been published elsewhere, or are pending external publication. **Chapter 2** was published in its entirety and has the following citation: *Anoxic photochemical oxidation of siderite generates molecular hydrogen and iron oxides.* ***Proceedings of the National Academy of Sciences***, *110, 25, 10073-10077 (2013).* **Chapter 3** was published in its entirety and has the following citation: *Energetic selection of topology in ferredoxins.* ***PLoS Computational Biology***, *8(4) (2012).* **Chapter 4** was published in its entirety and has the following citation: *Discovering the electronic circuit diagram of life: Structural relationships among transition metal binding sites in oxidoreductases.* ***Philosophical Transactions of The Royal Society B***, *368, 1622, 20120257 (2013).*

Acknowledgements

Not knowing what was just about to happen, I mindlessly attended **Prof. Paul Falkowski**'s lecture in 2009, because it was mandatory for all first year graduate students. He talked about the Big Bang, possible alien lives, plate tectonics, and how his wiring diagram is like a London subway system and makes the world go around. I was instantly hooked to what he talked about, but then I felt the subject was too far-fetched and simply beyond my comfort zone. Later in summer 2010, I encountered Paul again at the Rutgers Energy Institute's annual symposium. That time I was brave enough to reach out to him – thanks to **Prof. Eric Garfunkel**, my undergraduate mentor, who told me how crazy (creative) Paul is. Fortunately, Paul agreed to accept me as his graduate student. Since then, I have strived to copy him as much as possible to become a "mini Paul". I ended up mimicking his Hemingway-like writing style, passion for work, even drinking a cup of espresso with cream and a pack of sugar (I never had an espresso before). I just enjoyed every single moment I spent in Paul's lab and publications just came along naturally as a result, which was the best part. Of course, Paul is not the only one to whom I am forever indebted. **Prof. Vikas Nanda** (CABM), who I initially met in **Prof. Joseph Marcotrigiano** (CABM)'s undergraduate biophysics course in 2008, taught me computational biology and biophysics which became indispensable tools of my research. **Prof. David Case** (Chemistry) and **Prof. G. Charles Dismukes (Chemistry)**, who kindly agreed to serve as thesis committee

members, taught me biophysics and inorganic chemistry, all of which are at the core of my research. I also thank **Prof. Nathan Yee** (Environmental Sciences), **Prof. Debashish Bhattacharya** (Evolution, Ecology and Natural Resources), **Prof. Yana Bromberg** (Biochemistry and MIcrobiology), **Prof. Joseph Marcotrigiano** (Chemistry and Center for Advanced Biotechnology and Medicine), **Prof. Silke Severman** (Marine Sciences), **Prof. Bob Kopp** (Earth and Planetary Sciences), **Prof. Kay Bidle** (Marine Science), **Prof. Lawrence Williams** (Chemistry), **Prof. Edward Castner** (Chemistry), **Prof. Leonard Feldman** (Physics/IAMDN), **Prof. Jing Li** (Chemistry) and **Alan Goldman** (Chemistry) for inspiring discussions and support. **Prof. Eric Garfunkel** (Chemistry) and **Prof. Ki-Bum Lee** (Chemistry) were my undergraduate research advisors who initially introduced and guided me to the exciting world of scientific research. I also thank my dear colleagues, especially **Dr. Agustina Rodriguez-Granillo**, **Dr. Stefan Senn**, **Dr. Arye Harel** and **Ben Jelen**, whose works are included in this thesis and rightfully deserve to share credit. I would also like to mention **Kevin Wyman** and **Beatrice Birrer** whose help was absolutely necessary to get the work done. Last but not least, I would like to thank all former and current members of Environmental Biophysics and Molecular Biology Lab (Falkowski) and the neighboring labs at Marine Science (Bidle, Gorbunov et al.), Computational Design and Molecular Evolution Lab (Nanda), Evolutionary/Comparative Genetics Lab (Bhattacharya), Bioinformatics of Protein Function Prediction, Genome Variation Analysis Lab (Bromberg), Environmental Geomicrobiology Lab (Yee), Nanobioscience lab (Lee) and Surface Chemistry Lab (Garfunkel).

Dedication

For my wife **Jinyoung Song** and my parents **Sangsoo Kim** and **Soonae Cho**

Table of Contents

List of Tables

List of Figures

Chapter 1

Order from Chaos

Living organisms emerged on our planet roughly 3.8 billion years ago (1). Many scientists have attempted to develop plausible hypotheses to explain how it happened, yet the origin of life remains elusive. While reconstructing the exact history of life is virtually impossible, there is a consensus that a set of connected events must have taken place in the beginning of life. Throughout the thesis, I attempt to describe a few of those important events that ultimately would have led to an ordered system that allowed for self-replication and is far from thermodynamic equilibrium; i.e. the origins of life.

Energy requirements for Life Energy is essential for all living organisms because non-equilibrium thermodynamic reactions that are essential to life are maintained by harnessing energy from the surrounding environment into the system (2). Without external energy, the entropy in the system readily increases and reach thermodynamic equilibrium with the surrounding environment. However, the vast majority of external energy sources on Earth are electromagnetic energy (Solar) from outer space and thermal energy from subsurface that are not directly available for living organisms. Instead, early microbes emerged with strategies to carry out enzymatic reactions that utilize environmental electrochemical energy that is constantly generated by the external energy sources (Fig. 1.1). Externally generated electromagnetic energy and internally produced thermal energy would have promoted prebiotic chemical reactions that enabled endogenous synthesis of building blocks of life and provided precursor substrates for microbial metabolism. Since most metabolic reactions are coupled to reduction-oxidation reactions, a supply of regenerated electron donors and acceptors would have been required in the environment to maintain sufficient, sustainable, redox gradients (Table

1).

Geothermal energy from subsurface   Geothermal heat is one of the two major external sources that supply energy to Earth's surface and the constant flux of external energy to the environment creates energy gradients. Radioactive decay of potassium (^{40}K), thorium (^{232}Th) and uranium (^{235}U, ^{238}U) create thermal energy responsible for virtually all geophysical processes on Earth. Owing to its local pH and temperature gradients, hydrothermal vents have gained special attention as a probable birthplace of life exemplified by an iron-sulfur hypothesis (6) and a serpentinization hypothesis (7). Both hypotheses rely on the fact that environmental electron donors and acceptors are generated and cycled by geophysical processes. For example, molecular hydrogen outgassed from mid-ocean ridges, hydrothermal vents, and areal/sub-areal volcanoes would have been the source of electrons for early metabolic reactions. Likewise, a carbon dioxide rich early atmosphere, generated by geophysical a carbon cycle, would have been a significant electron acceptor on a planetary scale. With coexisting reducing and oxidizing equivalents, early organisms readily harnessed energy by a series of redox reactions coupled to H_2 oxidation and CO_2 reduction. For example, Wood-Ljungdahl (W-L) pathway (8) is considered one of the early metabolic pathways to emerge that utilize hydrogen as an electron donor and carbon dioxide as an electron acceptor. The W-L pathway drives many extant Archea and Bacteria in the modern biosphere.

Electromagnetic energy from space   Earth's upper atmosphere receives approximately 170 petawatts (10^{15} watt) of electromagnetic energy (9) which would have been virtually unlimited source of energy for early microorganisms. However, how early microorganisms may have utilized such a significant amount of energy, besides anoxic
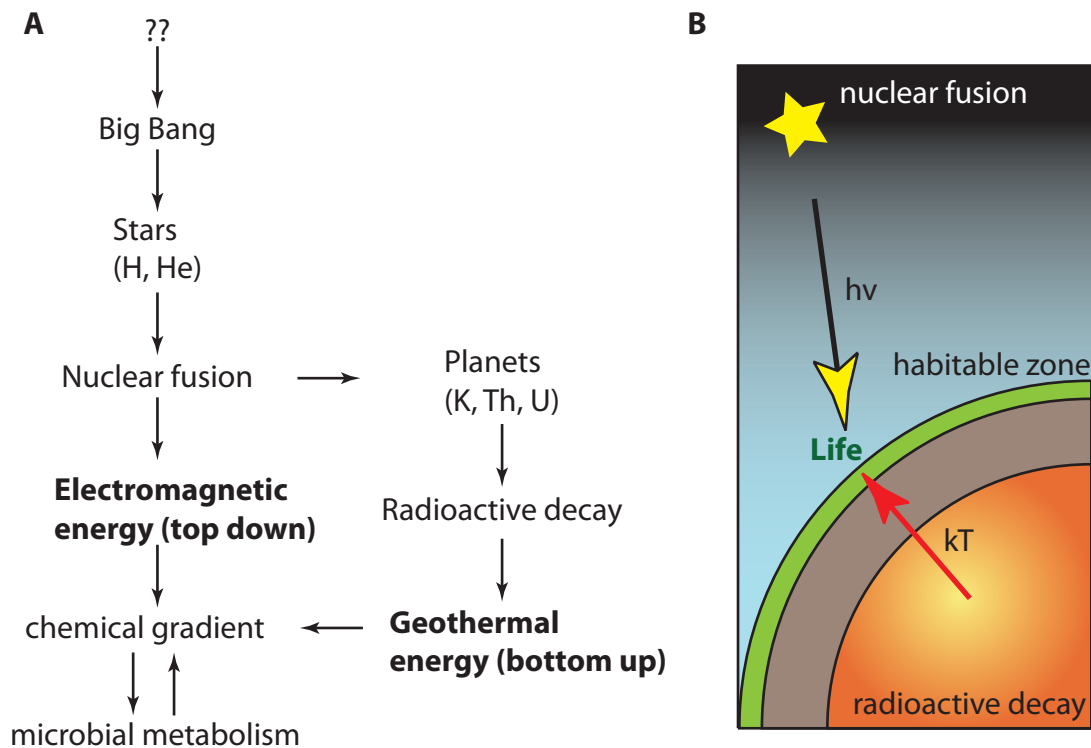
Figure 1.1: (A) The universe was born after the Big Bang, approximately 13.4 billion years ago (3). Through nucleosynthesis, elemental hydrogen and helium appeared and stars were created. Heavier elements subsequently were synthesized (Fig. 1.2A) *via* nuclear fusion in supernovae (4). The surface of the Earth cooled and allowed liquid water on a terrestrial surface, but the core remained molten due to heat generated by radioactive decay of potassium (^{40}K), thorium (^{232}Th) and uranium (^{235}U, ^{238}U). (B) The terrestrial surface on Earth would have been constantly stimulated by geothermal energy (bottom up) and electromagnetic energy (top down). With the constant supply of external energy, the chemical gradients were generated *via* photochemical reactions and geophysical processes. Bioenergetic pathways evolved to couple environmental chemical gradients to transduce energy. Microbial metabolic reactions, in turn, further shape the biosphere of the Earth through biogeochemical cycles primarily *via* gas exchange (5).

Electron donors	Source
H_2	subaerial and subaqueous volcanoes
H_2S	subaerial and subaqueous volcanoes
S^0	subaerial volcanoes (atmospheric reactions)
Fe^{2+}	subaqueous volcanoes, weathering
CH_4	subaqueous volcanoes
NH_4^+	subaqueous volcanoes
CH_2O	subaqueous volcanoes
Electron acceptors	Source
CO_2	subaerial and subaqueous volcanoes
CO	atmospheric reactions
SO_4^{2-}	subaerial volcanoes (atmospheric reactions)
S^0	subaerial volcanoes (atmospheric reactions)
NO (NO_3^- and NO_2^-)	lightening

Table 1.1: Virtually all Archean electron donors and acceptors are derived from the geophysical processes. Table is adapted from (10).

photosynthesis, is a largely unexplored area. Photochemical reactions, especially with high energy ultraviolet light, can make or break chemical bonds and significantly affect the geochemical inventories, similar to how geothermal energy produces electron donors and acceptors through geophysical processes. Photogeochemical reactions, if allowed to persisted over geological time scale, could have had an important consequences in the origins of life. In chapter 2, the potential photochemcial origin of H_2 gas in the early Archean upper ocean is discussed in detail. In particular, the study reveals that the photochemical oxidation of iron carbonate (siderite, $FeCO_3$) yields iron oxides as a by-product, which may suggest the mechanism by which enigmatic banded-iron formations were deposited. It is certainly possible that any minerals with the proper band gap can interact with electromagnetic waves, as long as the transition energy matches the energy of the incoming radiation (11). The photic zone of the ancient oceans would have been a place where countless number of chemical reactions could have taken place over billions of years, resulting in reactions condusive for early microbial life.

Metals in biology   Because of the unique chemical properties (Fig. 1.2B), transition metals play an outsized role in electron transfer reactions essential for energy transduction. The vast majority of proteins that conduct redox reactions catalyze single electron transfer reactions via prosthetic groups containing transition metals that undergo reversible changes in redox state. The transition states are, in turn, controlled by, and coordinated to, the ligands on the protein scaffold. Transition metals are specifically incorporated into the active sites via coordination bonds, in which the electronic states are three-dimensionally delocalized around the metal. Such a property is unique for d and f orbitals, allowing a high coordination number with multiple coordination geometries. In particular, d-orbitals are highly symmetrical and often capable of populating several orbital states, thereby conferring highly localized degenerate and pseudo-degenerate energy levels. This results in two types of magnetic properties (high/low spin), vibrant colors and multiple coordination numbers (12). Further, multiple oxidation states arising from the operational flexibility in d-orbital transitions allows the reduction-oxidation potential within a specific metal-ligand structure to be "tuned," or poised, for specific redox reactions. Such characteristics allow a spatiotemporal control of the directed electron flow, which is the hallmark of biological electron transfer reactions, exemplified in biological electron transport chains.

Protein fold evolution   Proteins are one of four major biological macromolecules that comprise life. A set of proteins capable of transferring electrons is essential for life and is encoded by conserved core genes. Such proteins, given their importance, would be among the earliest to emerge in the ancient oceans. The majority of the
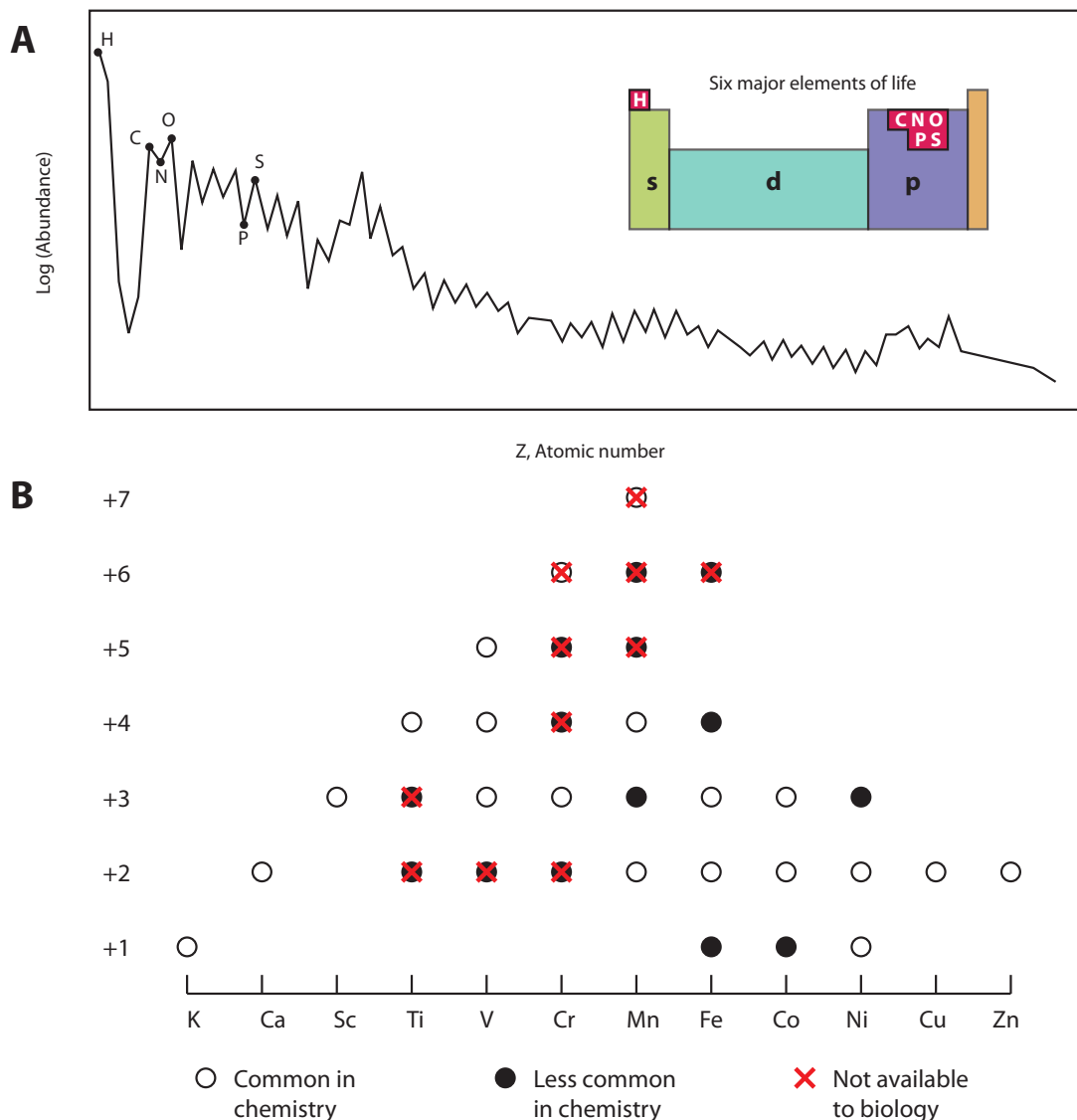
Figure 1.2: (A) Chemical elements were created in stars via nuclear fusion. Elemental abundance is unevenly distributed with hydrogen being the most abundant uranium being the scarce (13). The "big six" elements (H, C, N, O, P, S) make up virtually all biological molecules with a trace amount of metals as cofactors. Transition metals reside in the middle (green) of a periodic table, where the d-orbital chemistry dictates their chemical and physical properties. (B) Trace element metals have played a significant role in stabilizing protein structure, promoting electron transfer and performing catalysis in the early Archean. Such reactions are implicated in the origin of disequilibrium redox reactions. Provided with the building blocks of life, metals bound to short peptides could have functioned as protoenzymes, as is proposed by models of early protein evolution. Iron is by far the most common metal in biology with its two oxidation states, ferric and ferrous. Iron is found as an atomic co-factor or commonly a part of larger metal clusters, such as iron-sulfur cluster or Molybdenum-iron-sulfur cluster. Adapted from RJP Williams et. al. (14)

core enzymes contain metals as cofactors and prosthetic groups. The metals are bound by distinct structural units, protein folds that are specifically tailored to the cofactor. For example, nitrogenase (Fig. 1.3) is an enzyme that reduces N_2 gas to ammonium. Without nitrogenase, reduced nitrogen would not be available for biosynthesis biological macromolecules. Early redox enzymes must have evolved where transition metals are readily available (14). In a primordial soup model of the early Archean, metals in minerals, such as pyrite and mackinawite, may have played a significant role in adsorbing and concentrating organic molecules at the mineral-water interface and catalyzing electron transfer reactions. Revealing how metal-peptide interactions confer a specific geometry of the fold, a computational model with a heptapeptide demonstrated that a chiral topology was the outcome of intra-hydrogen bonding interactions, leaving only right-handed topology accessible for ferredoxins (Chapter 3). Hydrogen bonds play an important role in determining the protein folds around metals. Secondary structures in proteins are directly correlated to the hydrogen bonding forces, thereby providing a facile method for comparing a large number of protein folds with different sizes and metal types. The analyses indicate that simple folds, lacking alpha-helixes and beta-sheets, would be among the earliest fold to emerge with low degree of hydrogen bonding networks, followed by more complex folds with more alpha-helix or beta-sheet components in metalloenzyme folds (Chapter 4).
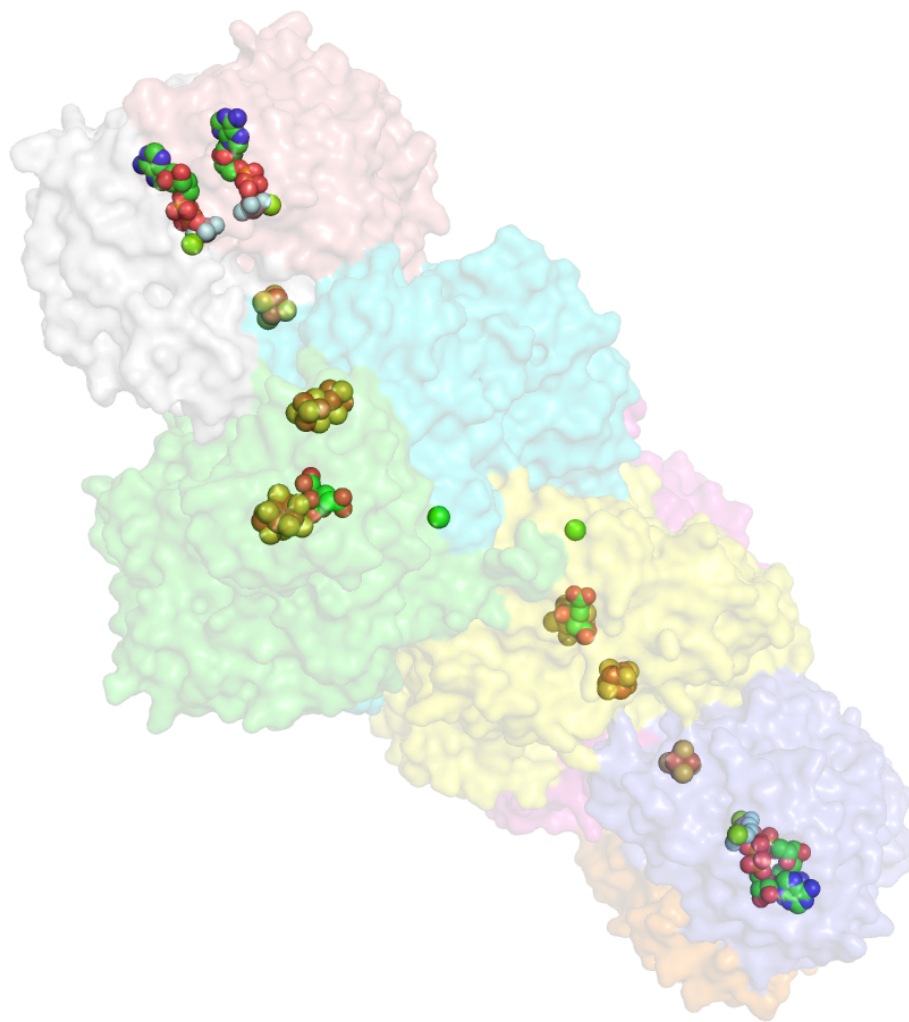
Figure 1.3: Nitrogenase is a hallmark of metalloenzymes with 38 iron atoms in the protein complex. The nitrogenase iron-protein, which encapsulates an iron sulfur cluster (Fe_4S_4) relays electron to MoFe protein, which contains a P-cluster (Fe_8S_7) and a FeMoCo cluster (Fe_7MoS_9C)(15), where dinitrogen molecule is reduced to ammonia.

References

1. Orgel LE (1998) The origin of life a review of facts and speculations. Trends in biochemical sciences 23(12):491-495.

2. Kauffman S (1993) The origins of order: Self organization and selection in evolution (Oxford University Press).

3. Lineweaver CH (1999) A younger age for the universe. Science 284(5419):1503-1507.

4. Spite M & Spite F (1978) Nucleosynthesis in the galaxy and the chemical composition of old halo stars. Astronomy and Astrophysics 67:23-31.

5. Vernadsky VI (1998) Living matter in the biosphere. The biosphere, (Springer), pp 56-60.

6. Wachtershauser G (1988) Pyrite formation, the first energy source for life: a hypothesis. Syst Appl Microbiol 10(3):207-210.

7. Martin W, Baross J, Kelley D, & Russell MJ (2008) Hydrothermal vents and the origin of life. Nature Reviews Microbiology 6(11):805-814.

8. Ragsdale SW & Pierce E (2008) Acetogenesis and the Wood-Ljungdahl pathway of CO_2 fixation. Biochimica et Biophysica Acta (BBA)-Proteins & Proteomics 1784(12):1873-1898.

9. Trenberth KE, Fasullo JT, & Kiehl J (2009) Earth's global energy budget. Bulletin of the American Meteorological Society 90(3):311-323.

10. Canfield DE, Rosing MT, & Bjerrum C (2006) Early anaerobic metabolisms. Philosophical Transactions of the Royal Society B: Biological Sciences 361(1474):1819-1836.

11. Sherman DM (2009) Electronic structures of siderite ($FeCO_3$) and rhodochrosite ($MnCO_3$): Oxygen K-edge spectroscopy and hybrid density functional theory. American Mineralogist 94(1):166-171.

12. Bersuker IB (2010) Electronic structure and properties of transition metal compounds: introduction to the theory (John Wiley & Sons).

13. Gamow G (1946) Expanding universe and the origin of elements. Physical Review 70(7-8):572-573.

14. Da Silva JF & Williams RJP (2001) The biological chemistry of the elements: the inorganic chemistry of life (Oxford University Press).

15. Spatzal T, et al. (2011) Evidence for interstitial carbon in nitrogenase FeMo cofactor. Science 334(6058):940-940.

Chapter 2

Anoxic Photochemical Oxidation of Siderite Generates Molecular Hydrogen and Iron Oxides

Photochemical reactions of minerals are under appreciated processes that can make or break chemical bonds. We report the photo-oxidation of siderite ($FeCO_3$) by ultraviolet radiation to produce hydrogen gas and iron oxides via a two photon reaction. The calculated quantum yield for the reaction suggests it would have been a significant source of molecular hydrogen for the first half of Earths history. Further, experimental results indicate this abiotic, photochemical process may have led to the formation of iron oxides under anoxic conditions. The reaction would have continued through the Archean to at least the early phases of the Great Oxidation Event, and provided a mechanism for oxidizing the atmosphere through the loss of hydrogen to space, while simultaneously providing a key reductant for microbial metabolism. We propose that the photochemistry of Earth abundant minerals with wide band gaps would potentially played a critical role in shaping the biogeochemical evolution of early Earth.

2.1   Introduction

In the early Archean eon, mineral surfaces have been hypothesized to have played a significant role in adsorbing and concentrating biologically relevant molecules and carrying out various chemical reactions as catalysts or reactants (1-4). Upon exposure to a flux of photons, minerals may undergo photo-induced electron excitation reactions, driving electron transitions, such as crystal field transitions, valence and conduction band transitions, and molecular orbital transitions as intervalence charge transfer (5). As a result, the photochemical reactions of minerals potentially can influence chemical reactions of basic elements such as H, C, N, O. Over geologic time scales, the fluxes of major chemical elements would have cycled through such photochemical reactions, prior

to biological chemistry of life on Earth (6). We hypothesize that emergent photogeo-chemical cycles may have played a critical role in shaping the early Earth environment. In this paper we examine the potential photochemical reactions of siderite in the pro-duction of molecular hydrogen. Although gaseous hydrogen is virtually non-existent in the contemporary atmosphere, in the early Archean the gas would have enhanced the greenhouse effect directly (7) and indirectly extending the lifetime of atmospheric methane. Both processes would have helped water to remain liquid on Earths surface despite significantly lower solar luminosity. Further, the existence of a hydrogen-rich environment would have enabled the spontaneous synthesis of key building blocks of life (1), and served as an electron donor in the early evolution of microbial metabolism (8). Despite its importance, the details of the early Archean hydrogen cycle are poorly un-derstood. Photochemical origins of molecular hydrogen have been proposed previously (9-12). The reactions suggested were based on the photo-oxidation of reduced metal species, such as hydrated Fe(II). Indeed, previous studies have confirmed this type of reaction; however, the ferrous iron species used (10-12) are unlikely to have been present in the upper ocean in the early Archean at neutral pH (13) because the precipitation of Fe(II) minerals would have been rapid. Hence, the major ferrous species in the early Archean upper ocean with a relatively low sulfide concentration would most likely have been associated with silicates and carbonate minerals such as siderite ($FeCO_3$), depend-ing primarily on the pCO_2 level (14). Models of Earths early atmosphere also suggest the presence of very high concentrations of CO_2 (15). One consequence of high concen-trations of CO_2 in contact with an anoxic, ferruginous ocean would be the spontaneous precipitation of siderite with an experimentally determined solubility product (K_{sp}) of $\sim 10^{-10}$ (16). In the absence of oxygen, siderite precipitates through the reaction of

ferrous iron with dissolved inorganic carbon (HCO_3^-, CO_3^{2-}); this reaction would have been thermodynamically highly favorable in the Archean. Indeed, siderite is found in rocks as old as 3.8 Ga, and its abundance suggests it was a prevalent ferrous species during Archean time (17-21). Although UV photo-oxidation of siderite is possible (22), and may have implications for the early hydrogen cycle, the rates and mechanisms of hydrogen production by this potential geochemical pathway have not been experimentally elucidated. Here we report a novel reaction of low energy UV photons with siderite that generates molecular hydrogen and iron oxides under anoxic conditions at neutral pH.

2.2   Results

Photochemical oxidation of siderite   To examine the potential photochemical reaction of siderite, we synthesized and suspended the mineral in anoxic aqueous phase and irradiated it with broadband light from a Xe source (see Methods section). Over a period of 24 h, the grayish white mineral became brown (Fig. 2.1), and upon further irradiation a black/dark brown product was formed. The product was magnetic (Fig. 2.2A) and XRD analysis confirmed it had a spinel structure concordant with either magnetite and/or maghemite (Fig. 2.2B). Measurements of the headspace revealed the presence of hydrogen gas. Mass spectroscopy and control experiments confirmed that the production of hydrogen resulted from a photochemical reaction with siderite. To determine the effective absorption cross-section of the reaction, we isolated the spectral irradiance with bandpass filters. We assumed the effective cross section of siderite photo-oxidation, $\Phi(\lambda)$, follows a Gaussian distribution with a function:

$$\Phi\left(\lambda\right) = e^{-\frac{(\lambda - \lambda_{max})^2}{\sigma^2}}$$

where λ_{max} is the wavelength at maximal effective absorption cross section. The irradiance spectrum from the Xe source through a filter, I(λ), multiplied by $\Phi(\lambda)$ yields an effective absorption spectrum. The integrated area under $I(\lambda) \times \Phi(\lambda)$ is proportional to the rate of hydrogen evolution, hence, λ_{max} and σ can be solved numerically with three bandpass filters. Coarse-grained parameter fitting method was used with a step size of 0.1 for λ_{max} and 1 for variance (σ^2) and the least square fit was optimal with λ_{max} =267 nm and (σ^2)=123. We determined the quantum yield for the reaction as the ratio of the number of hydrogen molecules produced to the number of photons absorbed by the siderite and the relative quantum yield follows the effective absorption cross-section of the reaction (Fig. 2.3A). To further elucidate the detailed reaction mechanism, we measured the amount of hydrogen produced under different light intensities. The experimental results (Fig. 2.3B) fit with a power function ($x^{2.08}$), revealing the rate of production of H$_2$ is proportional to the square of light intensity. The linear increase in quantum yield with increasing flux (Fig. 2.3B, inset) strongly suggests the photochemical production of hydrogen requires two photons that drive two independent photochemical reactions.
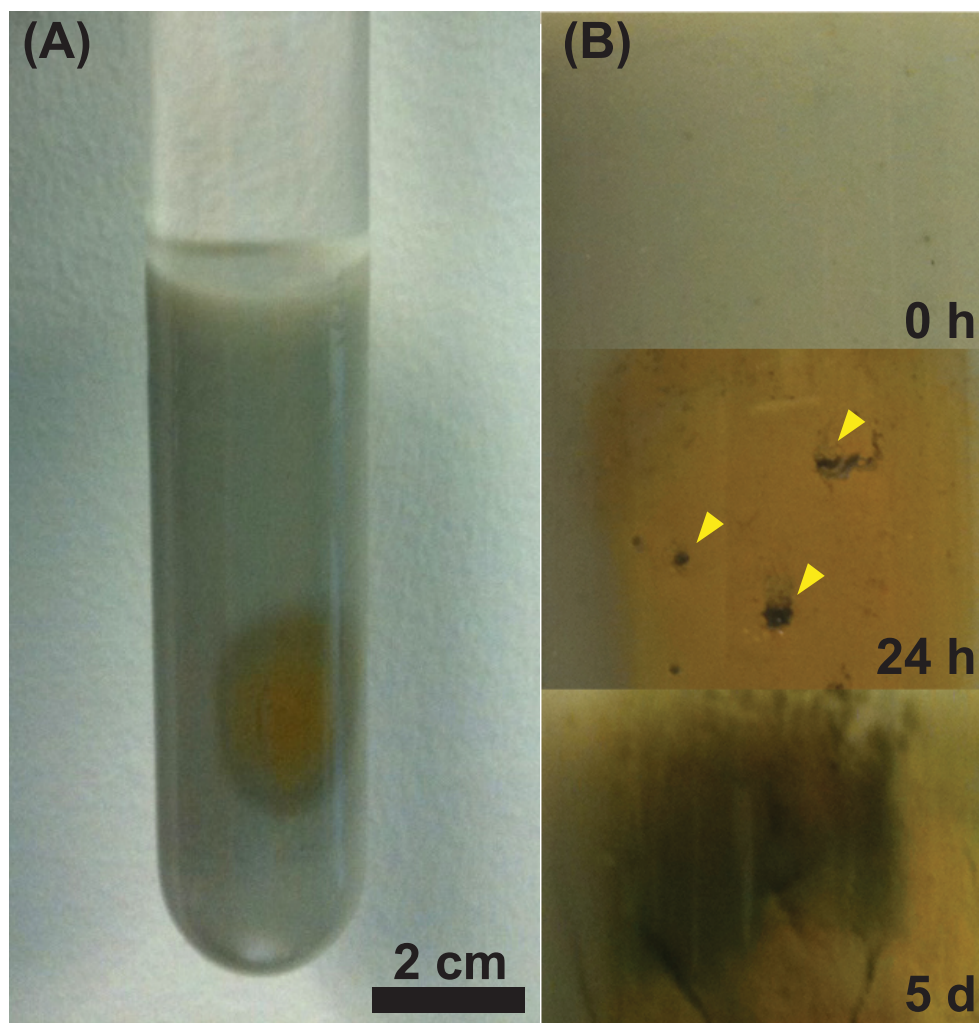
Figure 2.1: (A) Siderite was placed in a quartz tube and irradiated with a Xe lamp under anoxic conditions at pH 7.5-8.0. (B) Siderite oxidation is observed as the color changes upon UV irradiation over time (from top to bottom). The mineral turns brown (oxyhydroxide) and subsequently black (magnetite). Trapped hydrogen bubbles are marked with yellow wedges.
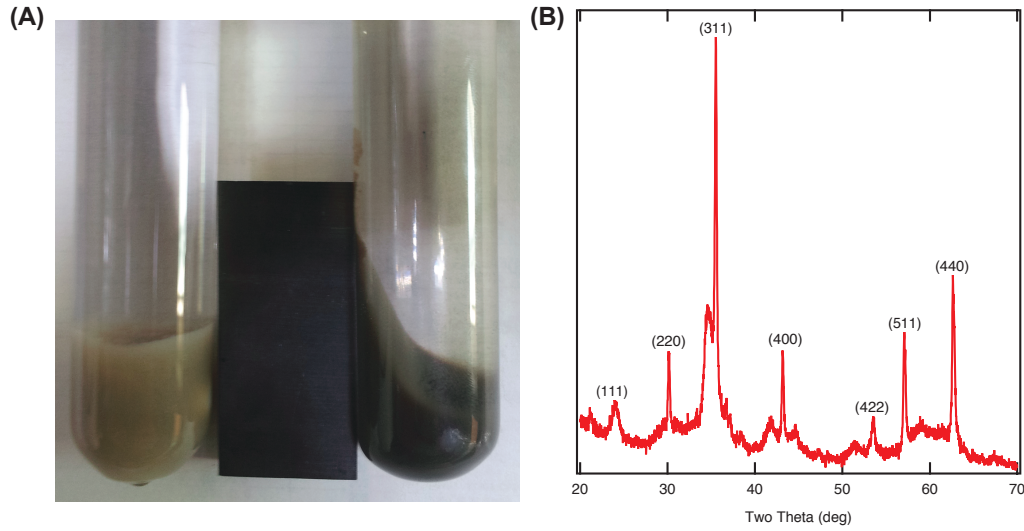
Figure 2.2: (A) The photochemical oxidation of siderite (left) leads to a magnetic product after 1 day of illumination under 600 W (right). Upon exposure to oxygen, siderite becomes non-magnetic goethite (see Methods section). (B) Powder XRD spectra of the magnetic product reveals it is maghemite/magnetite under anoxic conditions.

Estimated Quantum Yield of the Photochemical Reaction under Projected Archean Solar Flux In the early Archean, the integrated solar flux was $\sim 30\%$ lower than present but the relative contribution from UV radiation was proportionally higher and was not blocked by ozone (15). To estimate the solar UV flux, we used an estimated value of 0.1 W/m^2nm (23). The UV photon flux from 245 to 290 nm, boundaries defined by the Gaussian curve (Fig. 2.3A), can then be calculated by (24):

$$\frac{Q(quanta \cdot m^{-2}sec^{-1})}{W(watt \cdot m^{-2})} = \frac{\lambda}{hc} = \lambda(nm) \times 0.5035 \times 10^{16}$$

which can be arranged:

$$Q(\lambda) = 0.5035 \times 10^{15} \times \lambda$$

The integrated area under Q(λ) $\times\Phi(\lambda)$ yields the total number of effective UV photons per unit time (s) and area (m^2):

$$\int_{245}^{290} 0.5035 \times 10^{15} \times \lambda \times e^{-\frac{(\lambda-\lambda_{max})^2}{\sigma^2}} d\lambda$$

where λ_{max} is 267 nm and the variance (σ^2) is 123. The projected effective UV photon flux is 2.6×10^{18} quanta/m^{2}s, or 4.4 μmol quanta/m^{2}s. Because the photochemical oxidation of siderite is a two photon reaction, the linear increase in quantum yield with increasing flux suggests the effective quantum yield at the Earths surface to have been $\sim 1.8 \times 10^{-4}$ (Fig. 2.3B).

Hydrogen Production Capacity in the Archean "Photic Zome" The attenuation coefficient at 270 nm through a water column is $\sim 0.5m^{-1}$ (25). Therefore the flux at a given depth is reduced by a factor of e$^{-0.5x}$, where x is depth in meters. Based on the linear dependence of quantum yield to the photon flux density, (Fig. 2.3B), we can estimate the H$_2$ production rate from a siderite surface at a given depth. Assuming a near-surface effective UV flux of 4.4 μmol quanta/m^{2}s, the flux under x meters from the surface is $4.4 \times$ e$^{-0.5x}\mu$mol quanta/m^{2}sec. Since the reaction quantum yield is linearly proportional to the flux in μmol quanta/m^{2}sec with a slope of 4×10^{-5} (Fig. 2.3B), the H$_2$ production rate at x meters can be obtained as:

$$4.4e^{-0.5x} \times 4.4e^{-0.5x} \times 4x10^{-5} \mu mol H_2/m^2 s$$

Assuming the upper 1% of the ocean ($\sim$40 m) was the photic zone, we can estimate the hydrogen production in μmol/m^{2}s based on the estimated quantum yield and UV
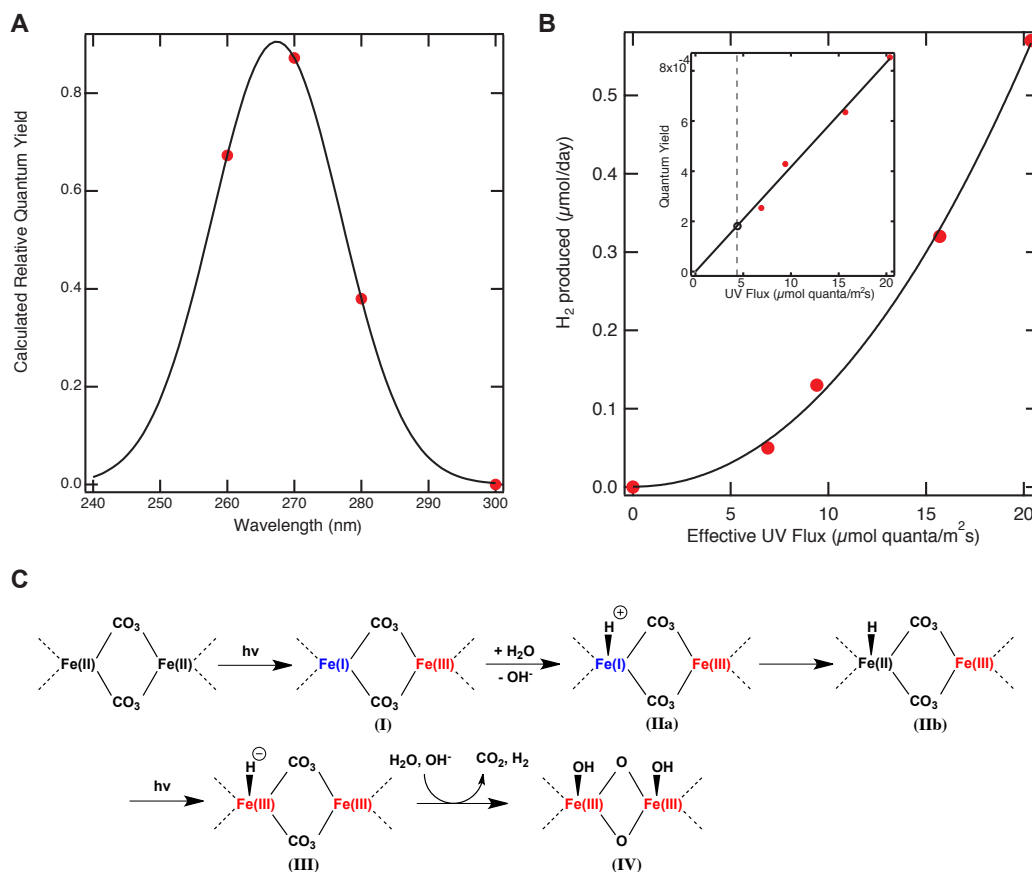
Figure 2.3: (A) Plot of the relative quantum yield of the photochemical oxidation of siderite as a function of wavelength. The maximal cross section is at 267 nm, which closely corresponds to the energy gap between the Fe(3d) ground state and a C-O antibonding orbital excited state. Four measurements were used to fit to a Gaussian function. (B) Plot of the hydrogen production (μmol) in 24 hours as a function of the photon flux (μmol quanta/m^{2}s). The experimental results revealed the rate of production of H_2 is proportional to the square of light intensity, and hence the yield increases linearly as a function of photon flux density. This analysis strongly suggests the overall reaction requires two photons to drive a two-electron transfer reaction mechanism. The estimated effective solar UV flux is $\sim$4.4 μmol quanta/m^{2}s at the ocean surface (dotted line) and the corresponding reaction yield based on the linear fit is $\sim$1.8 $\times$ 10^{-4}. (C) Proposed mechanism of the two-photon oxidation of siderite. Upon UV irradiation, an electron from Fe(3d) state is transferred to a neighboring iron atom generating charge separation **(I)**. This process is probably assisted by intermediate product **(IIa & IIb)** by simultaneous protonation of surface Fe(I). A second photon oxidize the intermediate product to hydrido-Fe(III) and this highly reactive intermediate product **(III)** reacts with a proton to form a molecular hydrogen. The two photon reaction yields an Fe(III)-oxyhydroxide precursor **(IV)** after losing CO_2 and adding hydroxyl groups. Since two photons are involved in the reaction, the probability of H_2 production scales linearly with increasing flux.

photon flux, both in μmol/m^{2}s.

Hence, the estimated maximum hydrogen production capacity in the upper 1% of the Archean ocean was $\sim$ 7.7 $\times 10^{-4} \mu$mol H$_2$/m^2s. With 510,100,000 km^2 surface area on Earth, of which 40% (80% ocean, of which half is exposed to the sun at any given time) can produce iron oxides, the estimated maximum iron oxide production capacity is estimated to be $\sim$ 10^{13} mol/yr. However, the actual hydrogen flux from the photooxidation would almost certainly be limited by the amount of reduced iron species.

Global Significance of Hydrogen and Iron Oxide Production from Photochemical Reaction The main sources of iron to the Archean ocean were hydrothermal vents and subaerial volcanoes. Given a flux of water through hydrothermal systems of 3 $\times$ 10^{13} kg/year in the contemporary ocean (26), and provided the hydrothermal flux was approximately 3 times higher in the Archean (27), the Archean hydrothermal water flux is estimated to be $\sim$9 $\times$ 10^{13} kg/yr. Further, assuming an Fe to H$_2$S ratio of $\sim$12:1 and an iron concentration of $\sim$50 mmol Fe/kg (28, 29), we estimate the iron flux to the entire Archean ocean to have been $\sim$4.5 $\times$ 10^{12} mol Fe/yr. For heuristic purposes, we assume the iron was uniformly dispersed in the ocean, and therefore only $\sim$1% of the flux would have been exposed to UV radiation. Hence, the photooxidative capacity in the photic zone of the Archean ocean would have exceeded the flux of iron by three orders of magnitude. This analysis strongly suggests that virtually all iron in the photic zone was photo-oxidized during the entire Archean eon. Since two iron atoms are oxidized for each hydrogen molecule, we can further estimate the global hydrogen production capacity in the early Archean upper ocean to be $\sim$2 $\times$ 10^{10} mol

H_2/yr, which is comparable to the estimated volcanic production rate of $\sim 10^{11}$ mol/yr (30). Our experimental results suggest that the photo-oxidation of siderite during the Archean potentially supplied significant amounts of molecular hydrogen in the upper ocean and to the atmosphere. This reaction would have had important consequences during the first half of Earths history.

2.3 Discussion

Proposed Reaction Mechanism The results of this study clearly establish that siderite can be readily photo-oxidized by UV light to generate molecular hydrogen and ferri-oxyhydroxides under anoxic conditions in an aqueous phase. From oxygen K-edge spectroscopy and density functional theory (31), the projected density of states based on hybrid function (B3LYP) revealed an energy gap of 4.6 eV (i.e., 270 nm), corresponding to either the energy between Fe(3d) and a C-O antibonding orbital or, alternatively, between O^{2-} and an Fe^{2+} excitation band gap. We propose that the first photon could produce a charge separation via electron excitation from Fe(3d) state to a C-O antibonding orbital, converting Fe(II)-Fe(II) into Fe(I)-Fe(III). This process would be promoted by simultaneous protonation of the surface Fe(I). The second photon would also produce a charge separation, forming highly reactive hydrido-Fe(III) which subsequently reacts with a proton from solution, giving H_2 and second Fe(III). The production of a H_2 molecule accompanies oxidation and hydrolysis of two Fe(III) atoms, resulting an iron oxyhydroxide mineral phase. The oxyhydroxide spontaneously is transformed to a magnetic assemblage of magnetite and maghemite. The solid-state transformation from the oxyhydroxide phase to magnetite/maghemite is accompanied by a loss of water molecules and recombination of ferrous iron with ferric iron.

Hydrogen Fueled Early Archean Microbial Community   The photochemically generated flux of hydrogen would have provided a local source of reductant for H_2-utilizating microbial communities in the photic zone of the Archean ocean. This source of hydrogen would have favored the evolution of anoxygenic phototrophs that used the gas as an electron donor for photosynthesis. Indeed, thin, siderite-bearing, carbonaceous laminations preserved in shallow-water facies of the 3.4 Ga Buck Reef Chert, South Africa, have been interpreted to represent some of the oldest-known mats constructed by photosynthetic microbes (32, 33) and hydrogen was the most probable electron donor to this ancient microbial community (34, 35). Furthermore, a photochemical source of hydrogen would have provided vast amount of energy for early chemolithotrophs, such as acetogens and methanogens. Ancient prokaryotic lineages of bacteria and archea that harbored the reductive acetyl-CoA pathway (36), considered one of the earliest bioenergetic metabolisms (37), would have flourished in the photic zone and may have been key members of the microbial photoautotrophic community in the surface waters of the Archean ocean (34, 35).

Solar Radiation Cycles Iron in the Upper Archean Ocean   The photo-oxidation of siderite in the upper ocean potentially alters our conceptual models of the Archean iron cycle (17). Ubiquitous carbonate minerals, such as diagenetic and ferroan dolomite in the terrestrial and shallow water settings and siderite in deep water settings suggest a carbonate rich Archean ocean. The paucity of siderite in shallow water sequences could be the result of low iron flux from deep thermal vents to the upper ocean, but the existence of shallow water banded-iron formations suggests there must have been a significant source of iron to the upper ocean. With the estimated

UV flux through a water column and the corresponding reaction quantum yield, the maximum iron oxidation capacity in the photic zone is $\sim 10^{13}$ mol Fe/yr, which is in excess of the optimistic estimation of the total Archean iron flux of $\sim 5 \times 10^{12}$ mol/yr into the total ocean from hydrothermal sources. The calculations imply the top 1% of the Archean ocean (i.e., the photic zone), has a much greater capacity for iron oxidation compared to the net iron flux from hydrothermal vents. We suggest that siderite in the photic zone of the Archean ocean was destroyed by UV radiation, leaving little trace except for iron oxides.

Photochemical formation of Iron Oxides The deposition of iron oxides is a hallmark of sedimentary sequences throughout the Archean and Proterozoic eons (38). The iron oxides originate at 3760 ± 70 Ma up to the late Precambrian, however the process(es) responsible for their formation remain enigmatic (39). The primary mineral product of UV photo-oxidized siderite under anoxic conditions is magnetite, suggesting solar UV radiation may have supplied iron oxides abiologically. Indeed, the deposition of ferric iron is potentially indicative of the UV flux during the Archean.

Photochemical Reactions Alter Planetary Redox State Hydrogen production from the photo-oxidation of siderite, and the subsequent escape of the gas from Earths atmosphere would have altered the oxidation state of the planet (40, 41). This reaction ultimately was attenuated but may not have been entirely quenched by the photobiological production and accumulation of ozone in Earths atmosphere following the Great Oxidation Event (GOE) at ~ 2.3 Ga. The generation of ozone clearly would have blocked shorter wavelength UV radiation, and quenched the mass-independent fractionation of sulfur isotopes in SO_2 (42). However, up until that time, the photochemical

oxidation of siderite and the loss of hydrogen almost certainly continued unabated. Depending on the ozone concentration, the photochemical oxidation of siderite could have continued for several hundred million years after the GOE, and in the process accelerated the long-term oxidation of Earths crust and atmosphere.

2.4   Concluding Remarks

The photochemistry of siderite and potentially other Earth abundant minerals in shaping the biogeochemistry of Earth is a poorly explored area of experimental geochemistry, which we call photogeochemistry. Our experimental results strongly suggest that the photochemical oxidation of an abundant mineral potentially generated a significant source of hydrogen in shallow waters of the Archean oceans for several hundred million years, if not longer. The reaction would have continued through the Archean to at least the early phases of the GOE, and provided a mechanism for oxidizing the planet through the abiotic production of hydrogen and its loss of hydrogen to space, while simultaneously providing a key reductant for microbial metabolism. This process may have operated on Mars when it was a wet planet, and should be operating on other terrestrial planets with liquid water in the habitable zone, as long as they do not have gases that greatly attenuate UV photons from their stars.

2.5   Methods

Siderite Synthesis   Siderite was synthesized by mixing solutions of $FeSO_4$ and Na_2CO_3 under anoxic condition (1% H_2/N_2 gas mixture) in a glovebox. Sulfate and sodium ions in the siderite precipitate solution were removed by repeated washing with deoxygenated Milli-Q water. The siderite was filter dried under N_2, transferred to a
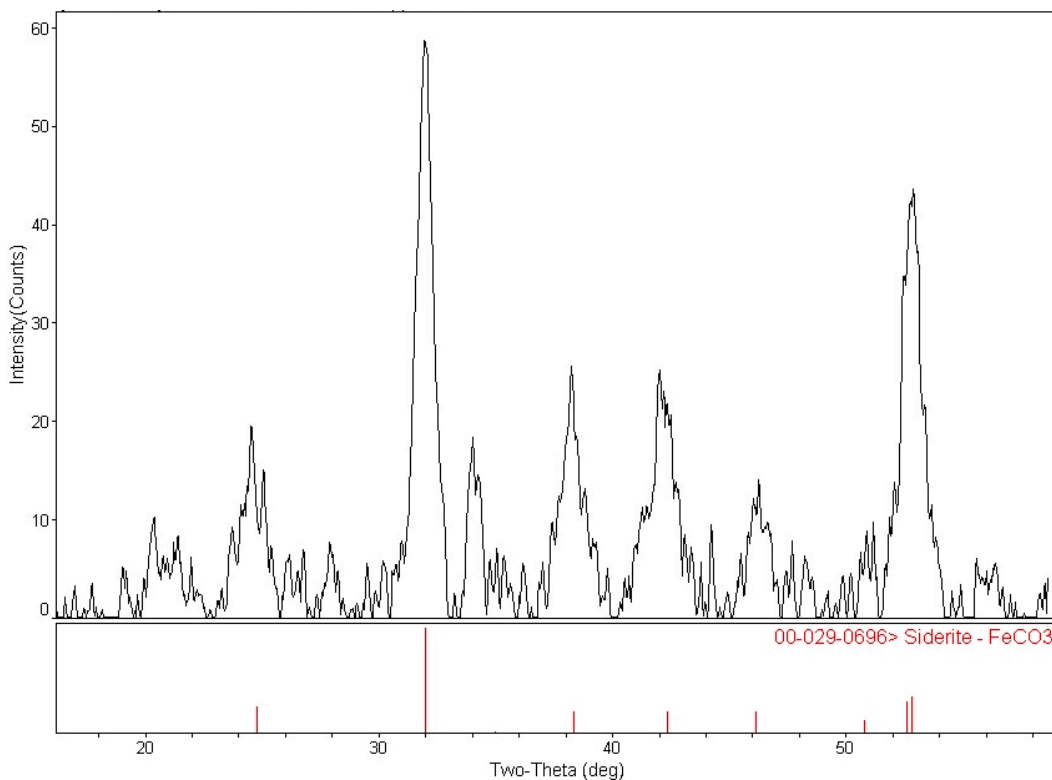
Figure 2.4: The synthetic siderite sample was filtered and dried inside an anaerobic chamber. The dried product was placed in a capillary tube and sealed with Bee's wax to prevent oxidation from oxygen. The sample was characterized with a transmission XRD technique (Brucker/Siemens HiStar diffractometer)

capillary tube and sealed for analysis with a transmission XRD technique to confirm the product (Fig. 2.4). Upon exposure to oxygen, siderite becomes the non-magnetic mineral, goethite (Fig. 2.5).

Optical setup   All components were installed on a Newport optical bench. A 1000 W Xe source (Oriel/Newport) was used throughout the experiments. A water filter was installed and water was circulated to prevent heat transfer from the source and a collimator was attached at the end of the water filter. Three bandpass filters were used to isolate the UV source with centers at 260, 270 and 280 nm. A fiber optics spectrometer (Ocean Optics USB 4000) was used to characterize the spectral irradiance
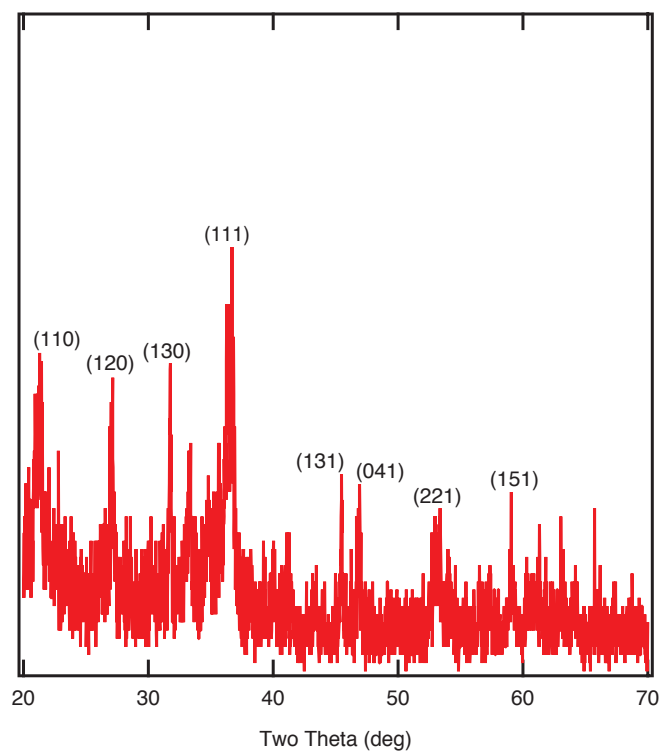
Figure 2.5: Siderite oxidation with oxygen forms goethite.
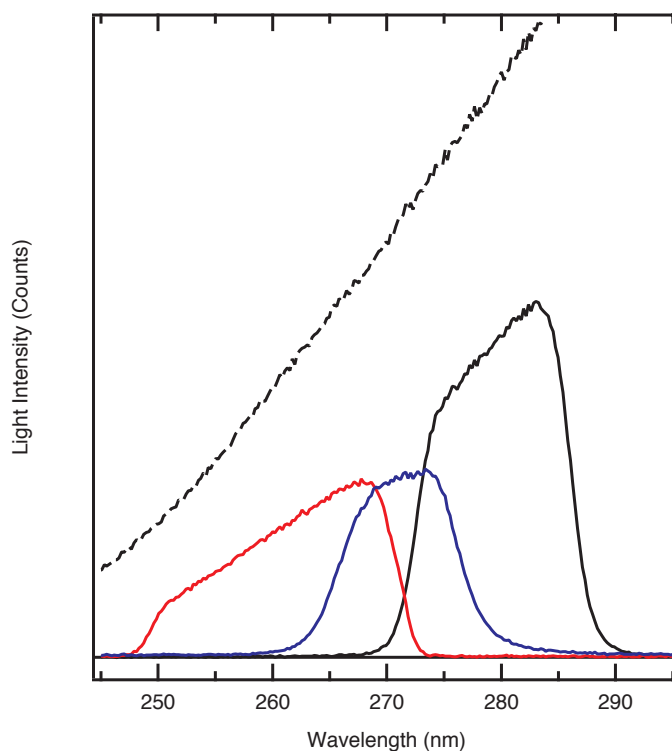
Figure 2.6: Three bandpass filters create monochromatic light. Unfiltered Xe arc light source spectrum (Black dotted line), Xe arc light through 260 nm (red), 270 nm (blue) and 280 nm (black).

of the source (from 200 to 800 nm at 0.22 nm resolution, Fig. 2.6).

Siderite photo-oxidation and analysis   A quartz reaction vessel with stress-relief gratings was fused with a serum bottle for the experiment. The final internal volume was ~75 ml and 3 mL of siderite suspension ($\sim 44\mu$mol FeCO$_3$) and 22 ml of boiled, anoxic Milli-Q water was transferred to the custom-made quartz reaction vessel in a glovebox. The sealed quartz tube was purged with ultra high purity N$_2$ gas for 10 min to ensure the absence of hydrogen or oxygen inside the tube. The headspace (50 ml) was analyzed by gas chromatography (SRI instruments) with a thermal conductivity detector (Fig. 2.7). To confirm that the gas produced by UV photo-oxidation was
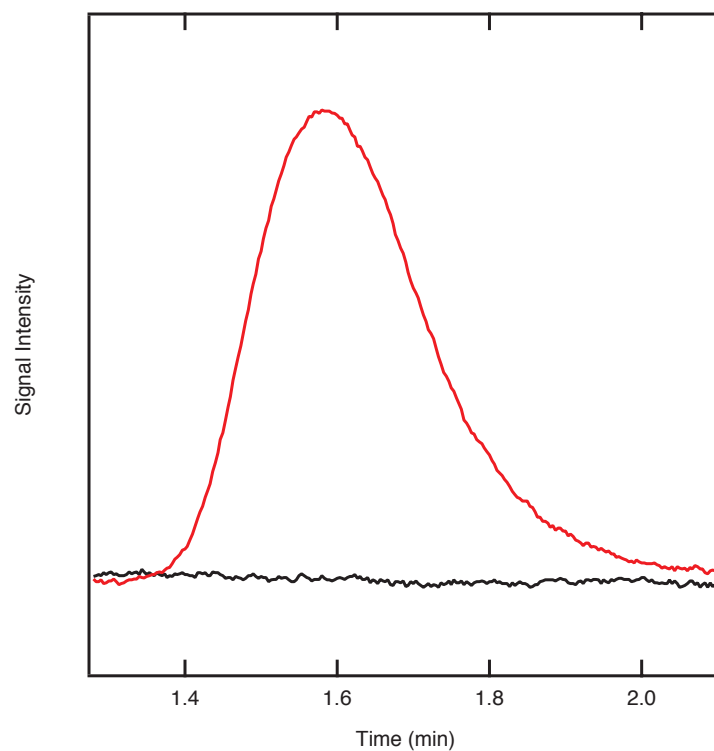
Figure 2.7: UV illuminated siderite yields a strong signal around 1.6 minutes (hydrogen), whereas a dark control experiment does not.

hydrogen, a membrane-inlet quadrupole mass spectrometer (MIMS) was used and the signal for m/z of 2 was recorded(Fig. 2.8).

Ferreioxalate chemical actinometry   To measure the number of incident photons, we used a ferrioxalate chemical actinometer (43).
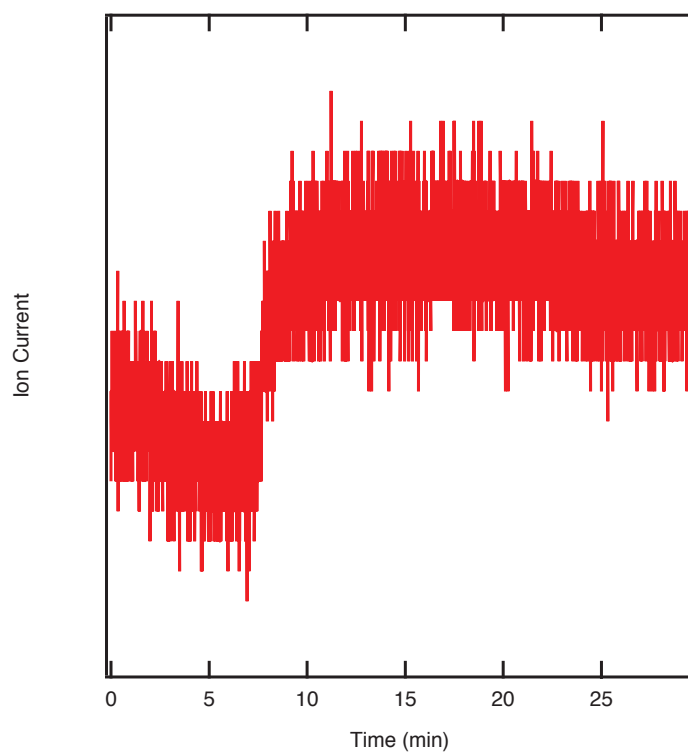
Figure 2.8: Headspace gas generated from siderite photooxidation has a m/z of 2.

Data analysis Spectra from the Xenon arc lamp through a set of filters, $I(\lambda)$, is then subject to the light absorption spectrum to obtain the true effective cross section $\Phi(\lambda)$, where:

$$\Phi_{260}(\lambda) = A(\lambda) \times I_{260}(\lambda)$$

$$\Phi_{270}(\lambda) = A(\lambda) \times I_{270}(\lambda)$$

$$\Phi_{280}(\lambda) = A(\lambda) \times I_{280}(\lambda)$$
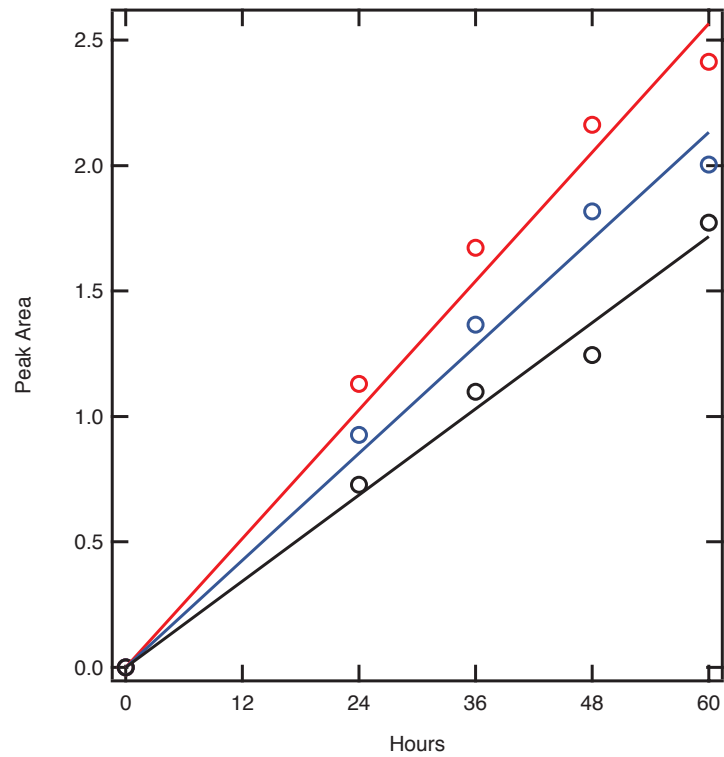
Filter	Linear Fit	R^2	Photon flux
260 nm	$y = 0.0432x$	0.98315	0.49 μmol quanta/min
270 nm	$y = 0.0359x$	0.98405	0.32 μmol quanta/min
280 nm	$y = 0.0287x$	0.98502	0.56 μmol quanta/min

Since the rate of hydrogen gas production is proportional to the integrated area under the effective cross section, the ratios of the hydrogen production rates ($k_{260}, k_{270}, k_{280}$)

correspond to the ratios of area under effective absorption cross section with different filters. Therefore, the following approximation would solve for the unknown constants, λ_{max} and σ.

$$k_{260} : k_{270} : k_{280} \simeq \int \Phi_{260}(\lambda)d\lambda : \int \Phi_{270}(\lambda)d\lambda : \int \Phi_{280}(\lambda)d\lambda$$

This can be rearranged to:

$$\frac{k_{260}}{k_{280}} : \frac{k_{270}}{k_{280}} \simeq \frac{\int \Phi_{260}(\lambda)d\lambda}{\int \Phi_{280}(\lambda)d\lambda} : \frac{\int \Phi_{270}(\lambda)d\lambda}{\int \Phi_{280}(\lambda)d\lambda}$$

To find λ_{max} and σ, the above equation is further rearranged and empirically determined. The measured H_2 production rates are then plugged into the equation resulting,

$$f(\lambda_{max}, \sigma) = \left| 1.505 - \frac{\int \Phi_{260}(\lambda)d\lambda}{\int \Phi_{280}(\lambda)d\lambda} \right| + \left| 1.251 - \frac{\int \Phi_{270}(\lambda)d\lambda}{\int \Phi_{280}(\lambda)d\lambda} \right|$$

where $\Phi_\lambda(\lambda) = e^{-\frac{(\lambda - \lambda_{max})^2}{\sigma^2}} \times I_\lambda(\lambda)$.

A coarse-grained search with a stepsize of 0.1 for λ_{max} and 1 for σ^2 yields a best fit with $\lambda_{max} = 266.9$ and $\sigma^2 = 117$.

2.6 Acknowledgement

discussions and technical assistance. We also thank anonymous reviewers for constructive comments.

2.7   References

1. Oparin AI (1957) The origin of life on the earth. The origin of life on the Earth. (3rd Ed).

2. Kim JD, Rodriguez-Granillo A, Case DA, Nanda V, Falkowski PG (2012) Energetic selection of topology in ferredoxins. PLoS Computational Biology 8(4):e1002463.

3. Hazen RM, Sverjensky DA (2010) Mineral surfaces, geochemical complexities, and the origins of life. Cold Spring Harbor perspectives in biology 2(5).

4. Da Silva JF, Williams RJP (2001) The biological chemistry of the elements: the inorganic chemistry of life (Oxford University Press).

5. Marcus RA (1965) On the theory of electron transfer reactions. VI. Unified treatment for homogeneous and electrode Reactions. The Journal of Chemical Physics 43:679.

6. Falkowski PG, Fenchel T, Delong EF (2008) The microbial engines that drive Earth's biogeochemical cycles. Science 320(5879):1034-1039.

7. Wordsworth R, Pierrehumbert R (2013) Hydrogen-nitrogen greenhouse warming in Earth's early atmosphere. Science 339(6115):64-67.

8. Canfield DE, Rosing MT, Bjerrum C (2006) Early anaerobic metabolisms. Philos T R Soc B 361(1474):1819-1834.

9. Borowska Z, Mauzerall D (1986) Formation of hydrogen on irradiation of aqueous ferrous ion with UV-light at neutral pH. Origins of Life and Evolution of the Biosphere 16(3-4):194-195.

10. Borowska ZK, Mauzerall DC (1987) Efficient near ultraviolet-light induced formation of hydrogen by ferrous hydroxide. Origins of Life and Evolution of the Biosphere 17(3-4):251-259.

11. Braterman P, Cairns-Smith AG, Sloper RW (1983) Photo-oxidation of hydrated Fe^{2+} – significance for banded iron formations. Nature 303:163-164.

12. Cairns-Smith AG (1978) Precambrian solution photochemistry, inverse segregation, and banded iron formations. Nature 76:807-808.

13. Holland HD (2003) The geological history of seawater. Treatise on Geochemistry, ed. HD Holland, KK Turekian Oxford: Elsevier:583-625.

14. Konhauser KO, et al. (2007) Decoupling photochemical Fe(II) oxidation from shallow-water BIF deposition. Earth Planet Sc Lett 258(1-2):87-100.

15. Kasting JF (1993) Earths early atmosphere. Science 259(5097):920-926.

16. Jensen DL, Boddum JK, Tjell JC, Christensen TH (2002) The solubility of rhodochrosite ($MnCO_3$) and siderite ($FeCO_3$) in anaerobic aquatic environments. Applied geochemistry 17(4):503-511.

17. Klein C (2005) Some Precambrian banded iron-formations (BIFs) from around the world: Their age, geologic setting, mineralogy, metamorphism, geochemistry, and origin. Am Mineral 90(10):1473-1499.

18. Klein C, Beukes NJ (1989) Geochemistry and sedimentology of a facies transition from limestone to iron-formation deposition in the early Proterozoic Transvaal Supergroup, South-Africa. Econ Geol 84(7):1733-1774.

19. Hessler AM, Lowe DR, Jones RL, Bird DK (2004) A lower limit for atmospheric carbon dioxide levels 3.2 billion years ago. Nature 428(6984):736-738.

20. Kaufman AJ, Hayes JM, Klein C (1990) Primary and diagenetic controls of isotopic compositions of iron-formation carbonates. Geochim Cosmochim Ac 54(12):3461-3473.

21. Ohmoto H, Watanabe Y, Kumazawa K (2004) Evidence from massive siderite beds for a CO_2-rich atmosphere before, 1.8 billion years ago. Nature 429(6990):395-399.

22. Joe H, Kuma K, Paplawsky W, Rea B, Arrhenius G (1986) Abiotic photosynthesis from ferrous carbonate (siderite) and water. Origins of Life and Evolution of the Biosphere 16(3-4):369-370.

23. Cockell CS, Raven JA (2007) Ozone and life on the Archaean earth. Philos T R Soc A 365(1856):1889-1901.

24. Morel A, Smith RC (1974) Relation between total quanta and total energy for aquatic photosynthesis. Limnol Oceanogr 19(4):591-600.

25. Smith RC, Baker KS (1981) Optical properties of the clearest natural waters(200-800 nm). Appl Optics 20(2):177-184.

26. Elderfield H, Schultz A (1996) Mid-ocean ridge hydrothermal fluxes and the chemical composition of the ocean. Annu Rev Earth Pl Sc 24:191-224.

27. Isley AE (1995) Hydrothermal plumes and the delivery of iron to banded iron formation. The Journal of Geology:169-185.

28. Kump LR, Seyfried WE (2005) Hydrothermal Fe fluxes during the Precambrian: Effect of low oceanic sulfate concentrations and low hydrostatic pressure on the composition of black smokers. Earth Planet Sc Lett 235(3):654-662.

29. Shibuya T, Komiya T, Nakamura K, Takai K, Maruyama S (2010) Highly alkaline, high-temperature hydrothermal fluids in the early Archean ocean. Precambrian Res 182(3):230-238.

30. Tian F, Toon OB, Pavlov AA, De Sterck H (2005) A hydrogen-rich early Earth atmosphere. Science 308(5724):1014-1017.

31. Sherman DM (2009) Electronic structures of siderite ($FeCO_3$) and rhodochrosite ($MnCO_3$): Oxygen K-edge spectroscopy and hybrid density functional theory. Am Mineral 94(1):166-171.

32. Tice MM, Lowe DR (2004) Photosynthetic microbial mats in the 3,416-Myr-old ocean. Nature 431(7008):549-552.

33. Tice MM, Lowe DR (2006) Hydrogen-based carbon fixation in the earliest known photosynthetic organisms. Geology 34(1):37-40.

34. Lovley DR, Phillips EJP, Lonergan DJ (1989) Hydrogen and formate oxidation coupled to dissimilatory reduction of iron or manganese by Alteromonas Putrefaciens. Appl Environ Microb 55(3):700-706.

35. Mojzsis SJ, et al. (1996) Evidence for life on Earth before 3,800 million years ago. Nature 384(6604):55-59.

36. Ljungdahl LG (1986) The autotrophic pathway of acetate synthesis in acetogenic bacteria. Annu Rev Microbiol 40:415-450.

37. Braakman R, Smith E (2012) The emergence and early evolution of biological carbon-fixation. PLoS Computational Biology 8(4):e1002455.

38. Kump LR, Holland HD (1992) Iron in Precambrian rocks - Implications for the global oxygen budget of the ancient Earth. Geochim Cosmochim Ac 56(8):3217-3223.

39. Dymek RF, Klein C (1988) Chemistry, petrology and origin of banded iron-formation lithologies from the 3800-Ma Isua Supracrustal Belt, West Greenland. Precambrian Res 39(4):247-302.

40. Walker JCG (1977) Evolution of the atmosphere.

41. Catling DC, Zahnle KJ, McKay CP (2001) Biogenic methane, hydrogen escape, and the irreversible oxidation of early Earth. Science 293(5531):839-843.

42. Farquhar J, Bao H, Thiemens M (2000) Atmospheric influence of Earth's earliest sulfur cycle. Science 289(5480):756-758.

43. Hatchard CG, Parker CA (1956) A new sensitive chemical actinometer. II. Potassium ferrioxalate as a standard chemical actinometer. Proceedings of The Royal Society Of London A 235:518-536.

Chapter 3

Energetic Selection of Topology in Ferredoxins

Models of early protein evolution posit the existence of short peptides that bound metals and ions and served as transporters, membranes or catalysts. The Cys-X-X-Cys-X-X-Cys heptapeptide located within bacterial ferredoxins, enclosing an Fe_4S_4 metal center, is an attractive candidate for such an early peptide. Ferredoxins are ancient proteins and the simple $\alpha+\beta$ fold is found alone or as a domain in larger proteins throughout all three kingdoms of life. Previous analyses of the heptapeptide conformation in experimentally determined ferredoxin structures revealed a pervasive right-handed topology, despite the fact that the Fe_4S_4 cluster is achiral. Conformational enumeration of a model CGGCGGC heptapeptide bound to a cubane iron-sulfur cluster indicates both left-handed and right-handed folds could exist and have comparable stabilities. However, only the natural ferredoxin topology provides a significant network of backbone-to-cluster hydrogen bonds that would stabilize the metal-peptide complex. The optimal peptide configuration (alternating α_L,α_R) is that of an α-sheet, providing an additional mechanism where oligomerization could stabilize the peptide and facilitate iron-sulfur cluster binding.

3.1   Introduction

Metals in proteins play important roles in stabilizing structure, promoting electron transfer and performing catalysis. Whole-genome analyses of phylogenetically diverse microorganisms suggest the earliest proteins incorporated metals and that metal usage over biological history evolved to match the availability of inorganic components in the environment (1,2,3). The mechanisms by which the ligand environment modulates metal affinity and specificity are of significant interest in the study of metalloprotein

evolution, function and design. Geometric requirements of metal coordination are predicted to impose specific constraints on the structure and topology of a bound polypeptide chain. In this study, we computationally model the accessible conformations of a ferredoxin-like peptide bound to an Fe_4S_4 cubane cluster in order to better understand how a putative early metalloprotein may have evolved. It has been proposed that a set of core genes encode proteins that carry out key redox reactions essential for promoting life and driving biogeochemical cycles (4). These proteins would be among the earliest to emerge in the ancient oceans. Identifying members of this set of core genes is an important step in understanding the evolution of microbial metabolism and emergent biogeochemical cycles. A number of features of ferredoxins make them an attractive as key players in the evolution of redox active proteins. Sequence analysis suggests that ferredoxins evolved very early in the origins of biological catalysis of redox reactions (5,6). All ferredoxins have a simple, conserved fold that binds two Fe_4S_4 clusters and is composed of fifty to sixty amino acids. Sequence and structural symmetry suggest it may have evolved from a gene duplication event of a thirty amino acid sequence, each capable of binding one iron-sulfur cluster (7,8,9,10). An early study of the ferredoxin sequence by Eck and Dayhoff in 1961 revealed even shorter repeats of four amino-acids (5), suggesting a prebiotic "protoferredoxin" was potentially composed of a primeval subset of the twenty amino acids (11,12). Midpoint potentials (-700 to -300 mV) of ferredoxins are lower than most other proteins, consistent with the mildly reducing early oceans (13,14). It has been speculated that the iron-sulfur cluster utilized in many redox proteins (15) may be an evolutionary relic of prebiotic chemistry catalyzed by mineral surfaces. Mineral surfaces can effectively adsorb and concentrate organic molecules and catalyze various chemical reactions implicated in the origin of
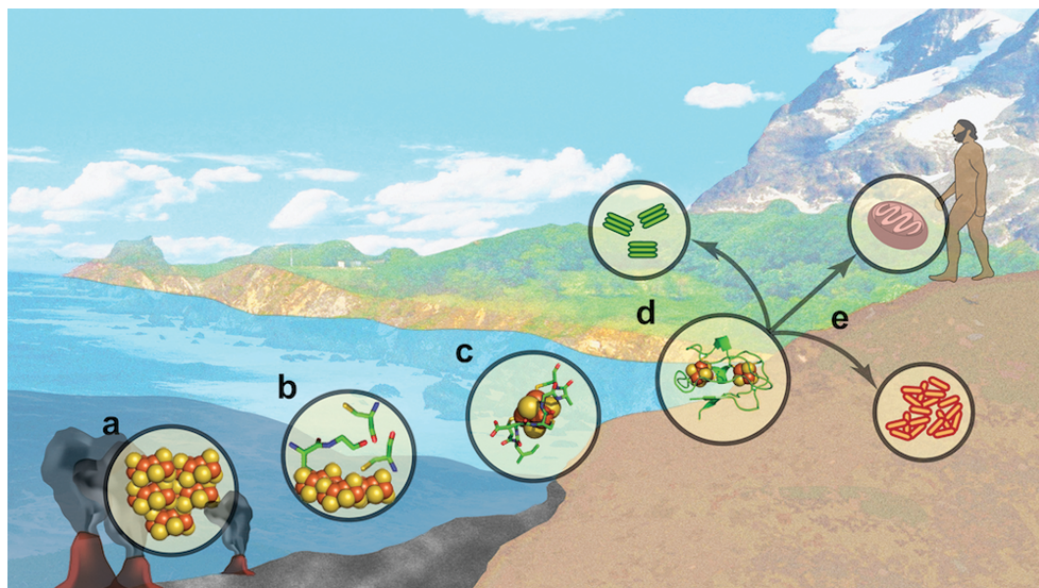
Figure 3.1: Hypothesized progression of iron-sulfur clusters from hydrothermal vents to life.

non-equilibrium redox reactions. Chiral mineral surfaces can selectively interact with chiral amino acids, and thus have been extensively studied as a potential origin of life on Earth (16). Iron-sulfur mineral surfaces especially have gained much attention in the context of deep-sea iron-sulfur rich hydrothermal vents where the earliest biologically relevant redox reactions are postulated to have occurred (17,18).

Assuming ferredoxin is one of the select core genes that originated from a mineral surface catalyst - what might intermediates in this progression from mineral to protein look like? (Fig. 3.1): (A) Iron-sulfur minerals such as pyrite and mackinawite can spontaneously catalyze carbon fixation to generate essential organic molecules for life (19,20,21,22), (B) The regular mineral concentrates amino acids (23), permitting new chemistry or enhancing existing reactions. (C) Condensation of small polypeptides occurs at the water-mineral interface (24). These polypeptides could have sequences similar to Dayhoff's proposed tetrapeptides (25) and would be capable of stabilizing

specific oxidation states of bound iron-sulfur fragments. (D) Small polypeptides are used as components of ferredoxin-like proteins. This is the transition from prebiotic chemistry to life and could occur within the context of models for such a transition such as an RNA-world where peptides are co-opted by small RNA hairpins (26). (E) Ferredoxin is retained in all kingdoms and becomes a domain of larger proteins that include many of the core redox genes of life. Although each of these stages is poorly understood and arguably controversial, this conceptual framework allows the design of specific simulations and experiments to explore the feasibility of ferredoxin evolution from a mineral precursor. The structural properties of a putative proto-ferredoxin peptide in Stage C have implications beyond origins of life models to metalloprotein design. Although several iron-sulfur binding sites have been designed into existing proteins (27,28) and de novo folds (29,30,31), very few have shown any significant stability to cycles of oxidation-reduction, diminishing their utility in catalysis or bioenergy applications (32,33). By elucidating the geometric and energetic constraints on a polypeptide bound to an iron-sulfur cluster, one can potentially understand the physical rules governing biological redox reactions and the designing novel protein structures. In the ferredoxin fold, iron-sulfur cluster has a quasi-tetrahedral structure with four coordination sites, which are most commonly occupied by four cysteine thiolates. The iron-sulfur cluster itself is achiral and the protein topology is mainly dependent on how the cysteine groups from a peptide chain are linked with four iron atoms in the cluster (34). Topologically, two different modes of protein-cluster interactions, right-handed or left-handed, are possible (Fig. 3.2). These two topological states cannot be superimposed onto each other by bending or stretching the representative molecular graphs (34). Previous studies analyzing iron-sulfur proteins in the Protein Data Bank (PDB)

reported that all redox active proteins had a right-handed fold; although left-handed configurations existed for redox inactive proteins (35). Herein, we present the work that elucidates why a right-handed heptapeptide topology may have evolved in the context of metal-protein energetics.

3.2 Results/Discussion

Definition of handedness in protein folds The achiral iron-sulfur (Fe_4S_4) cluster has a D_{2d} point group symmetry and is generally bonded to four cysteine thiolate groups (36,37). Three of the coordination sites are occupied by cysteine thiolates from a conserved heptapeptide sequence motif (CXXCXXC) and the remaining fourth coordination site is occupied by an outlier cysteine, which is most frequently followed by a proline (CP) (38). This particular binding motif accounts for approximately 25 % (36 out of 137) of iron-sulfur binding motifs from 104 crystal structures available from PDB (Table S1). Among the CXXCXXC motifs, about 85 % (31 out of 36) have a ferredoxin fold and approximately 15 % have globin-like folds and others as defined by Structural Classification of Proteins (SCOP) (39).

Topologically, the CXXCXXC heptapeptide motif can interact with an iron-sulfur cluster in two different ways, right-handed or left-handed (Fig. 3.2). For the discussion of these topological states, we quantitatively describe the handedness of the folding using a topology angle, θ, aligning the outlier cysteine on a z-axis of an internal coordinate frame (Fig. 3.3). Once the outlier cysteine is specified, handedness in this study is defined relative to the N- to C-terminus chain direction, either proceeding clockwise (right-handed: $0° < \theta < 90°$) or counterclockwise (left-handed: $90° < \theta < 180°$) around the cluster (Fig. 3.4). The outlier cysteine residue can be located before or after the
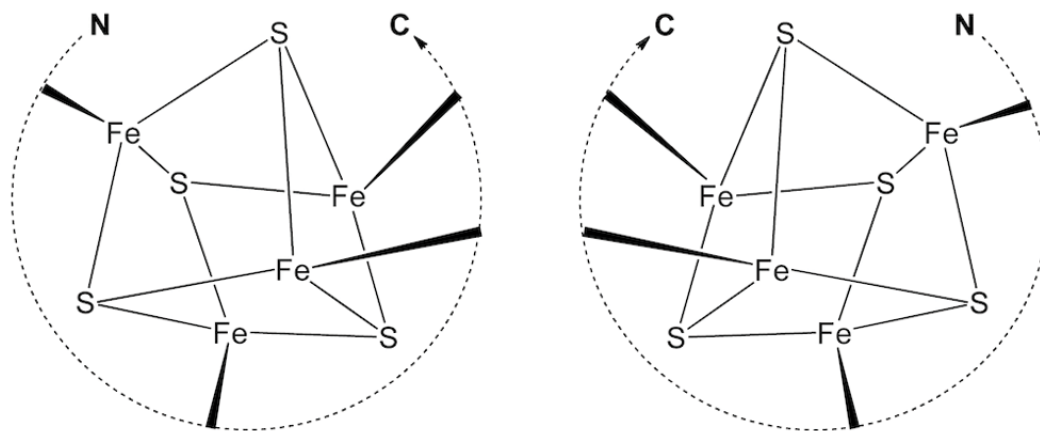
Figure 3.2: Two topological states of peptide-Fe_4S_4 cluster.

CXXCXXC motif (CPCXXCXXC or CXXCXXCCP).

Analysis of experimentally determined protein structures from PDB   Since the initial analysis on protein structure database (35), the number of solved protein structures has increased at an exponential rate. A non-redundant subset (30 % sequence similarity filter) of the PDB was searched for structures with an iron-sulfur (Fe_4S_4) cluster coordinated by a CXXCXXC sequence. The topology angle, θ, was calculated from the PDB coordinates (Fig. 3.4). A histogram of the topology angles reveals that only right-handed folds are involved in an iron-sulfur cluster binding (Fig. 3.5). The CXXCXXC motif always has a topology angle around 75°.

Computational simulation with protCAD & AMBER   Left-handed configurations of CXXCXXC were not observed, leading us to examine whether such configurations were energetically plausible. An ensemble of CGGCGGC polypeptide configurations was generated. Glycine was chosen for non-Cys positions due to its high backbone
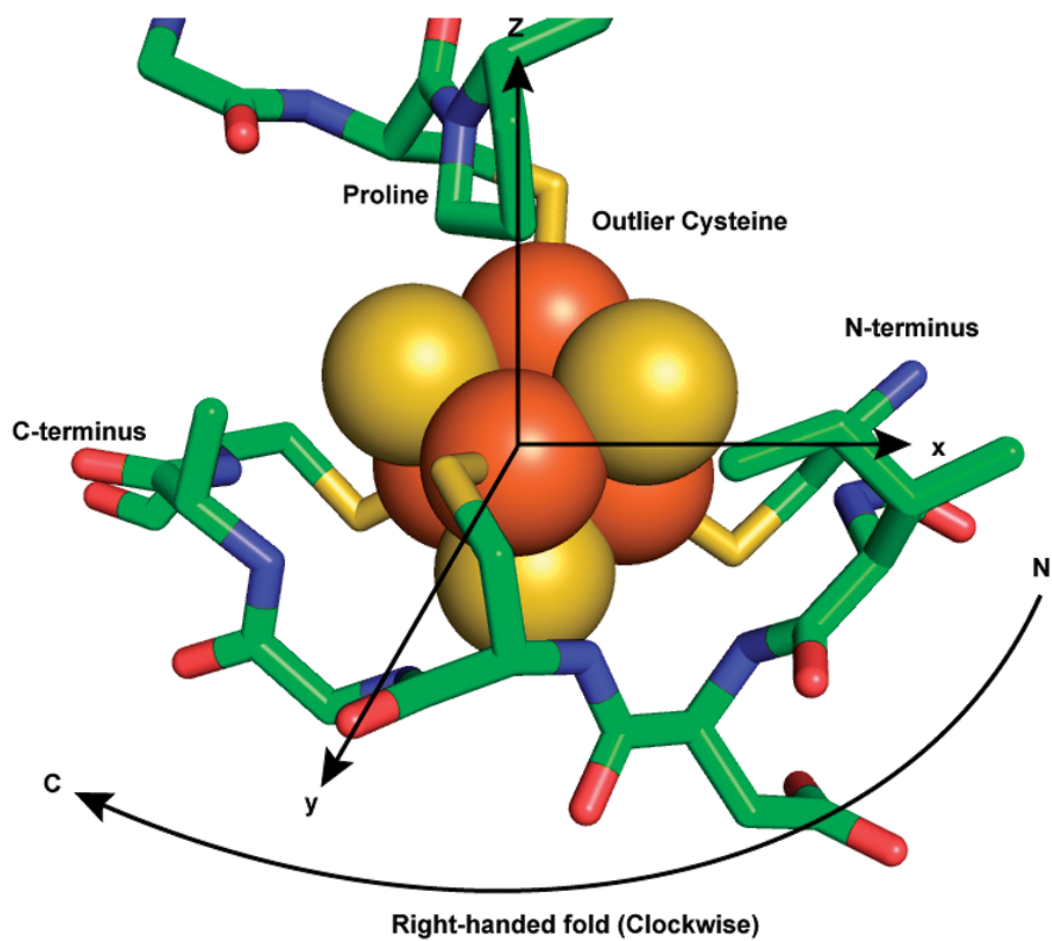
Figure 3.3: Right/Left fold configuration can be defined with an outlier, by orienting the outlier cysteine along the z-axis and iron-sulfur cluster being at the origin. A ferredoxin fold, with a conserved sequence CxxCxxC with an outlier cysteine, can create either right or left topological configuration. Right-handed fold is shown.
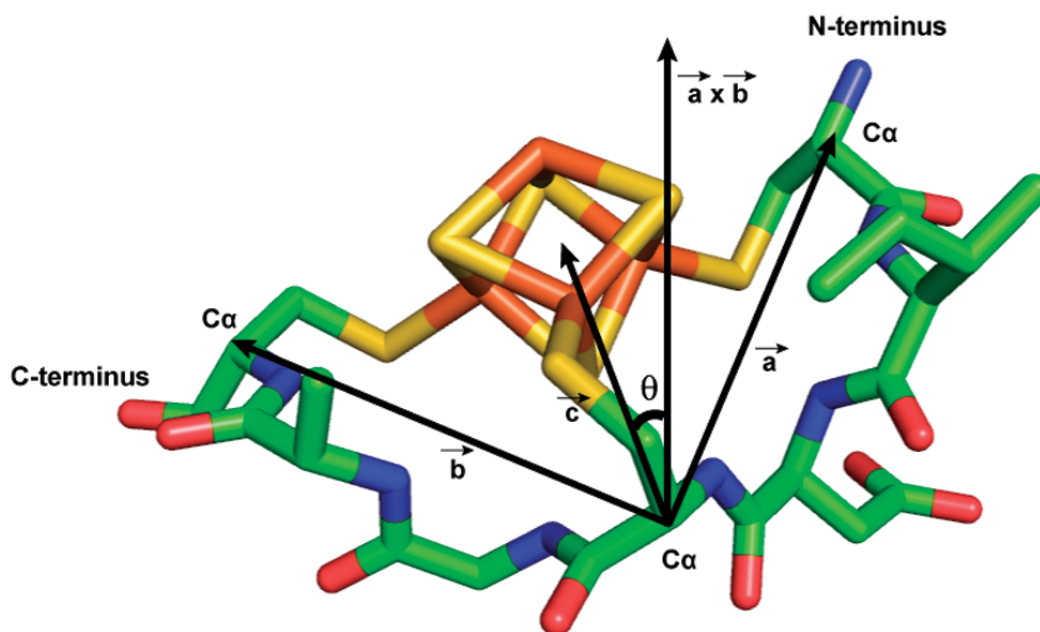
Figure 3.4: An arbitrary plane was defined with three cysteine carbon alpha coordinates. Three dimensional vector calculations were done to determine the topology angle of the protein fold.

flexibility, ensuring the primary conformational constraints came from metal-peptide interactions. The protCAD software platform(protein Computer Assisted Design) (29,40) was used to exhaustively enumerate all combinations of backbone and sidechain torsions in 60° intervals for ψ,ϕ and 120° intervals for the cysteine χ_1 rotamer (Fig. 3.7 and 3.6). Out of $5.8\text{x}10^{10}$ ($3^3\text{x}6^{12}$) configurations, 232 exhibited net-favorable van der Waals interactions (less than 0 kcal/mol), $\text{Fe}_{cluster}$ $\cdots$ Sγ distances (< 3 Å) and $\text{C}_\beta\text{-S}_\gamma$ $\cdots$ $\text{Fe}_{cluster}$ angles (120° to 180°) that would permit binding to an iron-sulfur cluster. The protein structures were then minimized in AMBER to reduce strain from distortions caused by discrete conformation sampling (41).

Topology angles of the computationally generated dataset clustered into two distinct populations - right and left-handed folds - suggesting the CGGCGGC heptapeptide could bind to the iron-sulfur cluster with either topology (Fig. 3.8). In fact, the
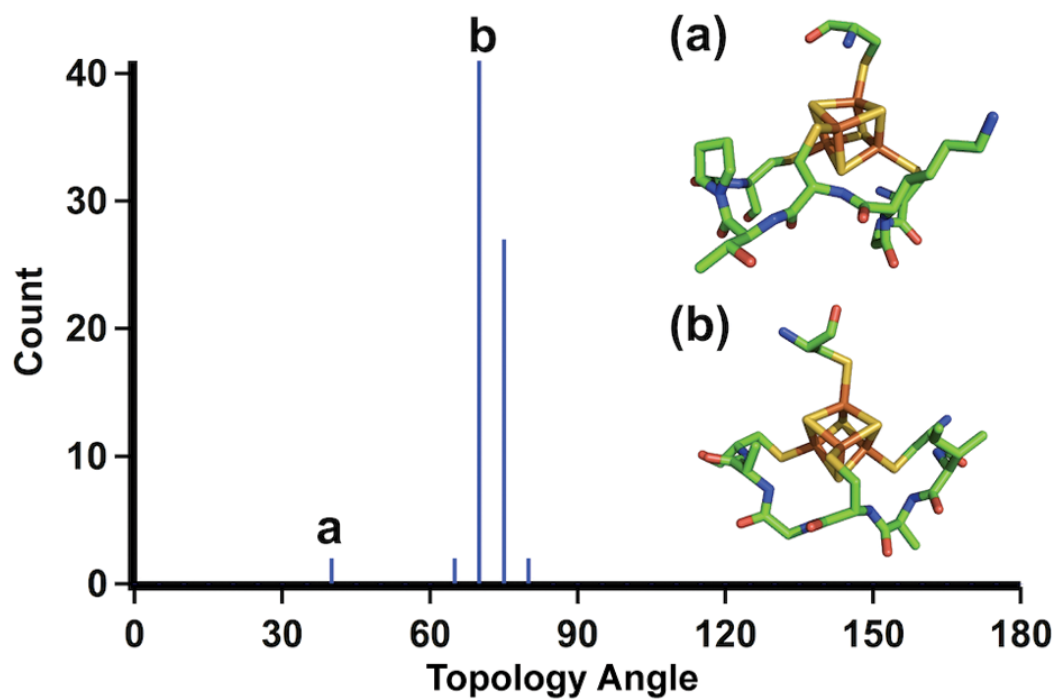
Figure 3.5: Experimentally determined protein structures from PDB (Protein Data Bank) were analyzed. The absence of peaks between 90 to 180 degrees suggests that the left-handed fold conformation does not exist in the known structures archived in the PDB.
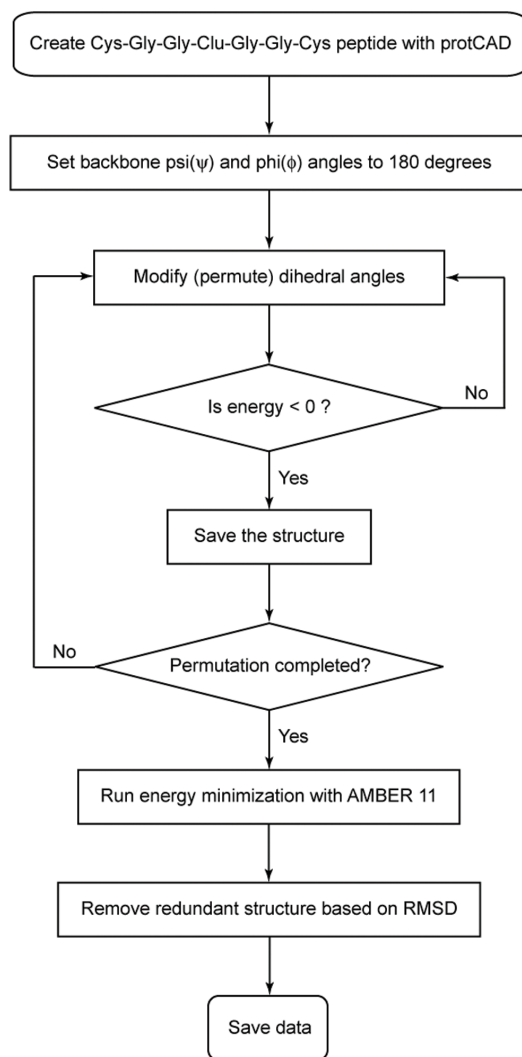
Figure 3.6: Cys-Gly-Gly-Clu-Gly-Gly-Cys peptide is created with protCAD. All possible structures are explored by permuting 17 rotatable dihedral angles of the peptide from -180 to 180 with a step size of 60 degrees.
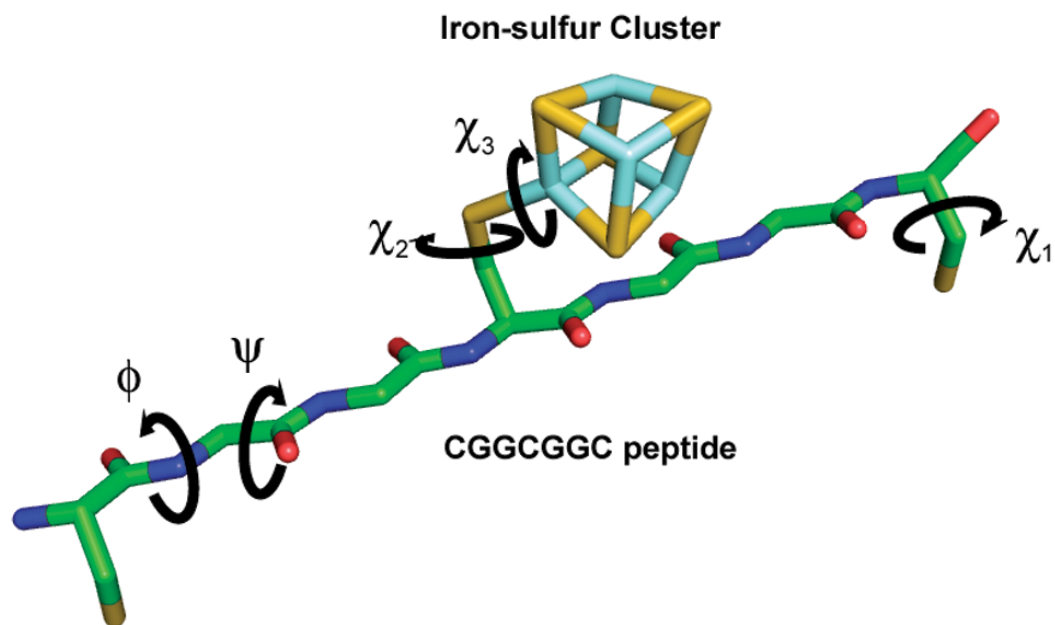
Figure 3.7: Protein ensemble generated by modifying psi, phi and chi dihedral angles. For a model heptapeptide-cluster complex, CGGCGGC fused to an iron-sulfur cluster, there are total 6 ψ angles, 6 ϕ angles, 3 χ_1 angles, and one each for χ_2 and χ_3 angles. The permutations are carried out by 60 degrees step size for ϕ and ψ and 120 degrees step size for χ angles.
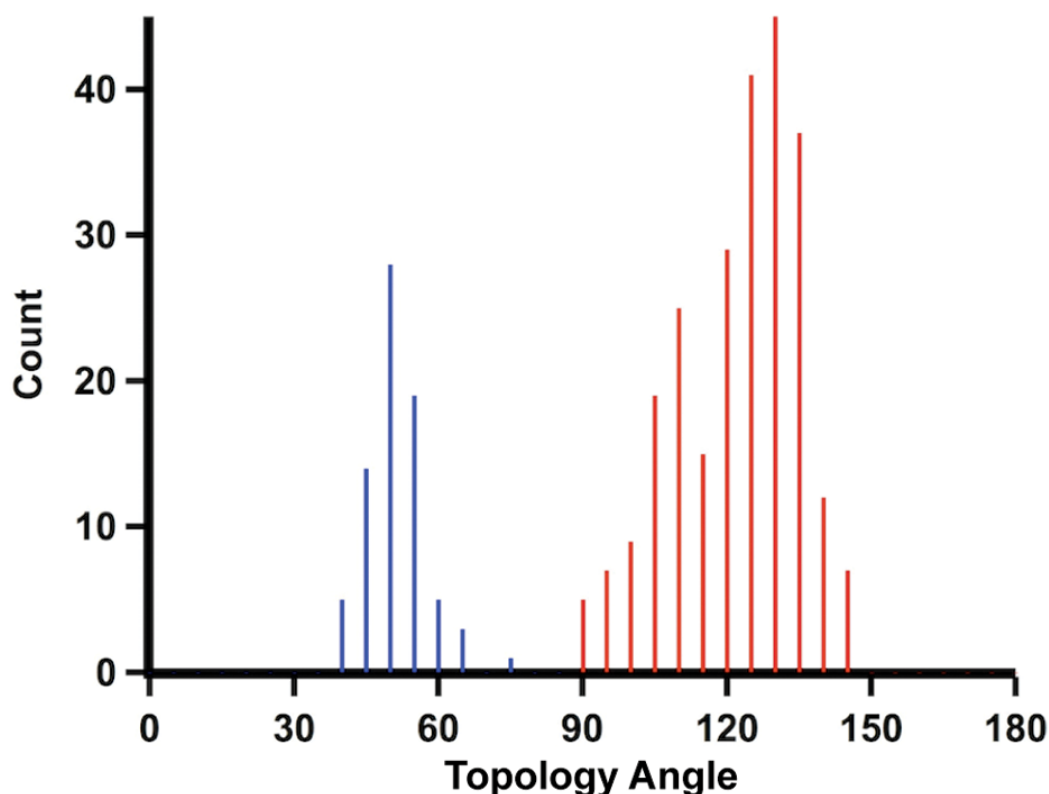
Figure 3.8: Cys-Gly-Gly-Clu-Gly-Gly-Cys hepeptide model has 232 structural entactic states, either right-handed (blue, 75 out of 232) or left-handed (red, 157 out of 232). Despite the inexistence of left-handed topological state in nature, model peptide suggests that left-handed structure can also properly interact with an iron-sulfur cluster

simulation identified more left-handed structures (67 %) than right-handed structures (32 %), indicating left-handed topologies were entropically favorable. Conducting the same simulation on CAACAAC resulted in 54 % left-handed and 46 % right-handed structures, suggesting that the steric hindrance of amino acid side chains itself is not sufficient to discriminate the handedness of the topological state. A histogram of the energy distributions for left and right-handed topologies show no significant difference (Fig. 3.9), indicating intrinsic stability of the fold alone is unlikely to account for evolution of a unique topology.
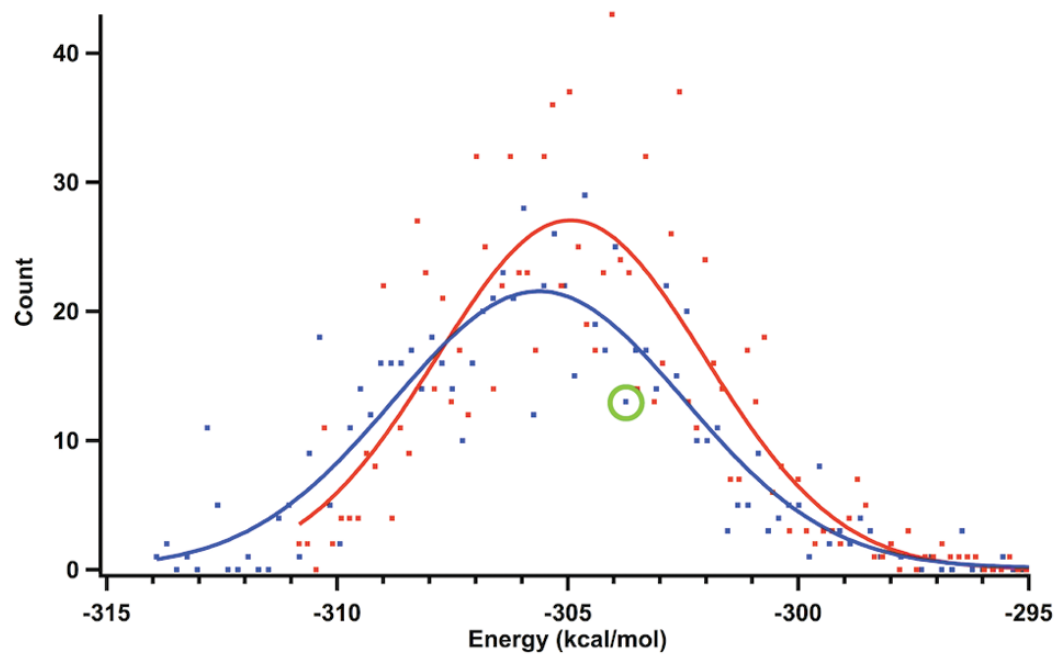
Figure 3.9: The energy distributions of right (blue) and left-handed (red) structures and Gaussian fits are very similar, which suggests that the natural selection was not influenced by the energetic stability alone. The energy corresponding to the ensemble that has the lowest RMSD to the experimentally determined ferredoxin structure (PDB: 2FDN)- green circle
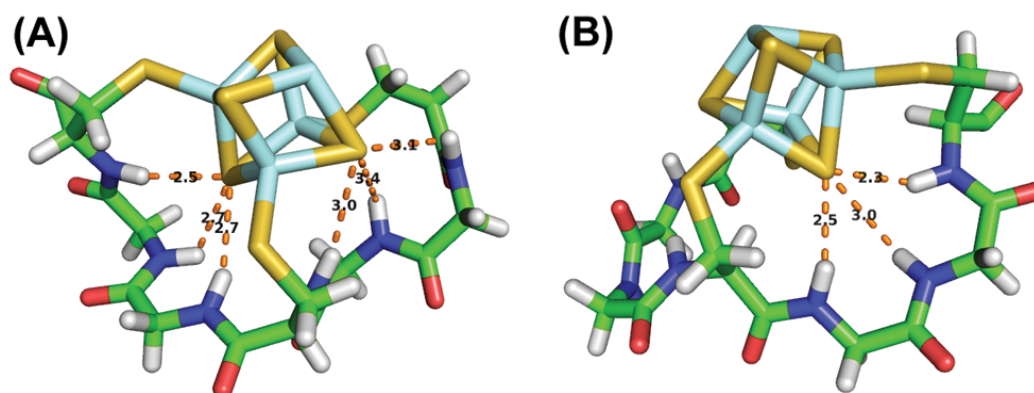
Figure 3.10: Computationally generated entactic states of the model heptapeptide with optimal peptide-cluster interaction energies. (A) Right-handed fold can form six hydrogen bonds, whereas (B) left-handed fold can only contribute three hydrogen bonds

Hydrogen bonds in iron-sulfur proteins The reduced state of the iron-sulfur cluster can be stabilized by hydrogen bonds contributed by nearby backbone amides (42). The number of hydrogen bonds around the iron-sulfur cluster is also related to the solvent accessibility to the cluster, thereby tuning the midpoint potential (43,44). A typical ferredoxin fold exhibits six such interactions with backbone amides directing the proton toward the cluster. Hydrogen bond formation is at the expense of unfavorable backbone dihedral angles, particularly the positive ϕ values at X_2 and X_3 positions (Table S2). For the analysis of the hydrogen bonding environment of computationally generated structures, interactions were counted based on discrete distance and angular cutoffs: a hydrogen-sulfur distance less than 3.5 Å and N-H $\cdots$ S angles between 120 ° to 180 ° (45). The number of hydrogen bonds between nitrogen and sulfur were counted based on cutoffs: 3.8 Å and 110° to 180°. Right-handed folds could accommodate six hydrogen bonds, but a maximum of three hydrogen bonds were found in structures with left-handed folds (Fig. 3.10).

Energetics of Peptide-Cluster Interactions   The electrostatic stabilization of a bound cluster by proximal backbone amides was estimated by comparing the total energies of charged versus uncharged clusters in the context of a coordinating peptide. The net contribution of hydrogen bonds can represented several ways: the average of pairwise distances between hydrogen and sulfur atoms (Fig. 3.11A) and discrete number of hydrogen bonds plotted against the peptide-cluster interaction energies (Fig. 3.11B). The interaction energy improves as the distances between sulfur atoms to hydrogen atoms are reduced. The result also indicates that the right-handed peptide-cluster interaction can have a stabilization effect up to -80 kcal/mol, whereas a left-handed fold can only achieve -50 kcal/mol. For comparison, we generated a CGGCGGC peptide using coordinates from experimental ferredoxin structures, including proteins with non-ferredoxin fold (Fig. 3.11A inset). The right-handed topology in natural ferredoxin and non-Fd proteins presents a network of stabilizing backbone amides that interact strongly with the Fe_4S_4 cluster. The result shows the best right-handed structure contributes more stabilizing hydrogen bonds than the best left-handed structure. Additionally, the inset to (Fig. 3.11) reveals tightly clustered experimental results, all which cluster around the same right-handed configuration and present six hydrogen bonds, suggesting the right-handed heptapeptide topology is a unique entactic state.

3.3   Conclusion

A microscopic analysis of the Fe_4S_4 binding region of ferredoxin provides some insights into the predicted features of an ancient, short proto-ferredoxin. The right-handed topology observed in redox-active iron-sulfur proteins is not dictated by the peptide chain. In fact, left-handed chain topologies are entropically favored and have slightly
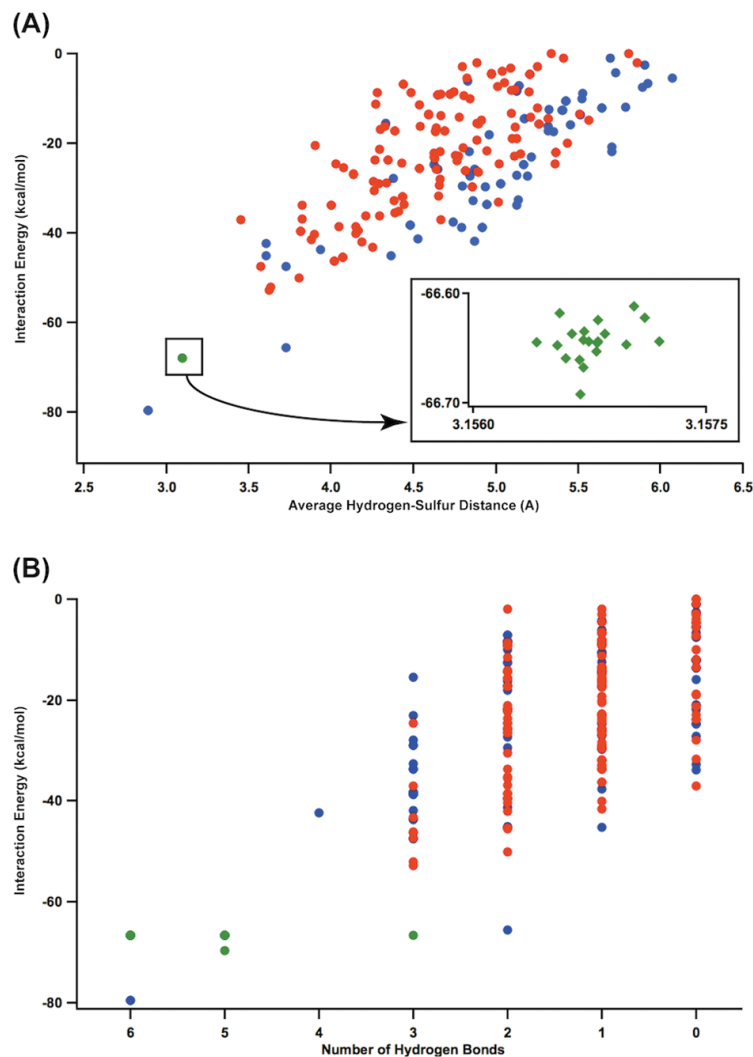
Figure 3.11: Hydrogen bonding environment of the 232 left- and right-handed heptapeptide-cluster conformations. (A) Interaction energy vs. average H-S distance of left (red), right-handed (blue) complexes. Experimentally determined ferredoxin structures (green) and non-ferredoxin redox active proteins (purple) show nearly identical bond geometries and calculated interaction energies. (B). The same dataset presented as the number of hydrogen bonds versus interaction energy. Only one simulated peptide in the ensemble contributes six hydrogen bonds, corresponding to the best interaction energy. This is equivalent to the natural right-handed fold.

improved stabilities in the absence of the cluster. Only when considering electrostatic interactions with the cofactor is the natural right-handed topology the optimal solution. Thus short CxxCxxC peptides alone are unlikely to serve as early redox active species without additional external stabilizing interactions. These may have taken the form of longer sequences with super-secondary structure such as those in designed peptide maquettes (38,46). It is interesting to note that the model conformation with the best peptide-cluster interaction energy and the ferredoxin-like conformations are both an α-sheet, characterized by residues in alternating α_L and β_R conformations. This motif was first described by Pauling and Corey as the pleated sheet (47). α-sheets are thought to be intermediates in a number of protein aggregation disorders (48,49). The conformation is also implicated in early peptides due to their anion binding properties (50). It is possible that stabilization of α-sheets provides the entactic state required for favorable cluster binding. The identification of a specific iron-sulfur binding topology may point the way to a mechanism by which the first core metalloproteins evolved.

3.4   Summary

The ferredoxin fold is one of the oldest structures capable of catalyzing electron transfer reactions. In nature, only a right-handed topology exists in the ferredoxin fold. To understand how a specific fold-handedness was selected, we analyzed the structural motif using the tools of de novo protein design, searching in an unbiased fashion for backbone geometries that can favorably interact with the tetrahedral iron-sulfur cluster. In silico, we found both left-handed and right-handed folds can be formed, however the right-handed folds provide up to six hydrogen bonds that can stabilize the reduced iron-sulfur cluster, whereas left-handed folds at most form three hydrogen bonds. The difference

in electrostatic conformational energy may have influenced selection of topology early in the evolution of iron-sulfur cluster containing proteins. This observation led us to establish a fundamental protein design principle that only right-handed peptide folds can properly interact while maintain redox function. Our results provide guidance in the creation of artificial proteins capable of carrying out redox reactions.

3.5   Methods and Materials

Topology angle   To have a quantitative measure for the fold-handedness, an arbitrary plane was defined with two vectors, which were defined by Cα coordinates from three cysteine residues. The topology angle, a quantitative measure of fold-handedness, was then defined as the angle between a normal vector of the arbitrary plane and a vector from the middle cysteine Cα to the cluster. By definition, the quantitative measurement of fold-handedness (topology angle) can take any numeric value from 0° to 180°.

PERL script

```
#!/usr/bin/perl -w

# Last updated May20 2010

#report distances within range


use Math::Trig;

use warnings;


@fileList = glob("*.pdb");
```

```perl
foreach $file (@fileList)

{

print "loading $file \n";

open(INPUT, "$file") or die("cannot open the file\n");

#reads specified file only at the moment

open(OUTPUT, ">>amide.txt")

or die("cannot write the file for writing\n");

#save as report.txt


@atomarray=();

@SF4 = ();

@N = ();

@H = ();


while ($data = <INPUT>)

{


if ($data =~ m/^ATOM/)

{

push @atomarray, $data;

}

}


foreach $line (@atomarray)
```

```perl
{

$nitrogen=substr($line, 13, 1);

if ($nitrogen =~ m/^N/)

{

$backamide=substr($line, 14, 1);

if ($backamide =~ m/ /)

{

push @N, $line;

}

}

}

foreach $line (@atomarray)

{

$hydrogen=substr($line, 13, 1);

if ($hydrogen =~ m/^H/)

{

$hydrogen=substr($line, 14, 1);

if ($hydrogen =~ m/ /)

{

push @H, $line;

}

}

}

foreach $line (@atomarray)
```

```perl
{

$clus=substr($line, 13, 1);

if ($clus =~ m/^F/)

{

{

push @SF4, $line;

}

}

}

# Determine center of mass

@refcoords = ();

@refcoords = getCenter($SF4[0], $SF4[1], $SF4[2], $SF4[3]);


# Get distance from center of mass to hydrogen

@dist = ();


for $a (0 .. $#H)

{

$distance = getDist($H[$a]);

push @dist, $distance;

}


#Get angles between N-H and N-Cluster

@ang = ();
```

```perl
@num = ();


for $b (0 .. $#H)

{

$angle = getAngle($H[$b], $N[$b]);

push @ang, $angle;

$num = getNumber($angle);

push @num, $num;

}


# print OUTPUT "$refcoords[0] \t $refcoords[1] \t $refcoords[2]\n";

$total = 0;

$distbin = 0;

for $i (1 .. $#ang)

{

if ($dist[$i] < 3.5)

{

# print OUTPUT "$file\t$dist[$i]\t$ang[$i]\t$num[$i]\tHydrogen$i\n";

$total = $num[$i] + $total;


}

$distbin = $dist[$i] + $distbin;

}
```

```perl
$avedist = $distbin/6;

if ($total >0.1)

{

print OUTPUT "$file \t $total bonds \t $avedist \n";

}

$total = 0;

$distbin = 0;

@refcoords =();

@dist = ();

@ang = ();


close(INPUT);

close(OUTPUT);

}
#-----------------------------

##Sub Routines ####

#-----------------------------


sub getDistance

{

$Ax = substr($_[0], 30, 8) + 0;

$Ay = substr($_[0], 38, 8) + 0;

$Az = substr($_[0], 46, 8) + 0;
```

```perl
$Bx = substr($_[1], 30, 8) + 0;

$By = substr($_[1], 38, 8) + 0;

$Bz = substr($_[1], 46, 8) + 0;


$distance =

sqrt(($Ax-$Bx)*($Ax-$Bx)

+($Ay-$By)*($Ay-$By)+($Az-$Bz)*($Az-$Bz));


return $distance;

}

sub getCenter

{

@coords=();

$x1 = substr($_[0], 30, 8) + 0;

$x2 = substr($_[1], 30, 8) + 0;

$x3 = substr($_[2], 30, 8) + 0;

$x4 = substr($_[3], 30, 8) + 0;


$y1 = substr($_[0], 38, 8) + 0;

$y2 = substr($_[1], 38, 8) + 0;

$y3 = substr($_[2], 38, 8) + 0;

$y4 = substr($_[3], 38, 8) + 0;


$z1 = substr($_[0], 46, 8) + 0;
```

```
$z2 = substr($_[1], 46, 8) + 0;

$z3 = substr($_[2], 46, 8) + 0;

$z4 = substr($_[3], 46, 8) + 0;


$refx = ($x1+$x2+$x3+$x4)/4;

$refy = ($y1+$y2+$y3+$y4)/4;

$refz = ($z1+$z2+$z3+$z4)/4;


push @coords, $refx;

push @coords, $refy;

push @coords, $refz;

return @coords;


}


sub getDist

{

$Ax = substr($_[0], 30, 8) + 0;

$Ay = substr($_[0], 38, 8) + 0;

$Az = substr($_[0], 46, 8) + 0;


$Bx = $coords[0] + 0;

$By = $coords[1] + 0;

$Bz = $coords[2] + 0;
```

```perl
$distance =

sqrt(($Ax-$Bx)*($Ax-$Bx)+($Ay-$By)*

($Ay-$By)+($Az-$Bz)*($Az-$Bz));

return $distance;

}


sub getAngle

{

#hydrogen(A)

$Ax = substr($_[0], 30, 8) + 0;

$Ay = substr($_[0], 38, 8) + 0;

$Az = substr($_[0], 46, 8) + 0;


#nitrogen(B)

$Bx = substr($_[1], 30, 8) + 0;

$By = substr($_[1], 38, 8) + 0;

$Bz = substr($_[1], 46, 8) + 0;


#cluster(C)

$Cx = $coords[0] + 0;

$Cy = $coords[1] + 0;

$Cz = $coords[2] + 0;
```

```
$BAx = $Bx - $Ax;

$BAy = $By - $Ay;

$BAz = $Bz - $Az;


$CAx = $Cx - $Ax;

$CAy = $Cy - $Ay;

$CAz = $Cz - $Az;


$dot = $BAx*$CAx+$BAy*$CAy+$BAz*$CAz;

$mag = sqrt($BAx*$BAx+$BAy*$BAy+$BAz*$BAz)*

    sqrt($CAx*$CAx+$CAy*$CAy+$CAz*$CAz);

#$angle = sqrt((($dot/$mag)+1)/2); #angle is cos(theta) value

$angle = (180/3.14159)* acos($dot/$mag);


return $angle;

}

sub getNumber

{

if ($angle < 180 and $angle > 120)

{

$num = 1;

}

else

{
```

```
$num = 0;

}


return $num;

}
```

protCAD Iron-sulfur cluster coordinates were extracted from the PDB file, 2FDN. We created a hybrid artificial amino acid residue (Clu) by linking an iron-sulfur cluster to a cysteine residue. The artificial amino acid was added to the amino acid library of protCAD. Initially a peptide ensemble (Cys-Gly-Gly-Cys-Gly-Gly-Cys) was created and subsequently the central Cys was substituted to Clu. For a given ensemble, there are six ϕ (C-N-C$_\alpha$-C), six ψ (N-C$_\alpha$-C-N). For each cysteine residue, there three χ_1 (N-C$_\alpha$-C$_\beta$-S$_\gamma$) dihedral angles. For the central iron-sulfur cluster fused cysteine residue, there are additional dihedral angles, which are χ_2 (C$_\alpha$-C$_\beta$-S$_\gamma$-Fe$_{Clu}$) and χ_3 (C$_\beta$-S$_\gamma$-Fe$_{Clu}$-S$_{Clu}$). All ϕ and ψ dihedral angles were increased by a step size of 60° and all χ dihedral angles were set at -180°, -60°, or 60°. The entire protein structural space was searched by the permutations of seventeen dihedral angles. Plausible protein structures were then determined by geometric parameters, such as a distance from S$_\gamma$ to Fe$_{Clu}$ with a cutoff (< 3.0Å). Energy parameters calculated based on a Lennard-Jones equation (45) was also used to detect feasible structures (total energy < 0 kcal/mol).

protCAD code (in C++)

```
#include "typedef.h"

#include "ensemble.h"

#include "PDBInterface.h"
```

```
void setBackBone(protein* _prot, UInt _position, double _step);

void checkSulfurGeometry(protein* _prot);


int main (int argc, char* argv[])

{

enum aminoAcid {A,R,N,D,C,Q,E,G,H,I,L,K,M,F,P,S,T,W,Y,V};


string inFile = argv[1];


double step;

sscanf(argv[2], "\%lf", &step);


//Input file should be four amino acids


PDBInterface* thePDB = new PDBInterface(inFile);

ensemble* theEnsemble = thePDB->getEnsemblePointer();

molecule* theMol = theEnsemble->getMoleculePointer(0);

protein* prot = static_cast<protein*>(theMol);


prot->silenceMessages();

prot->activateAllForRepacking(0);


//Modify sequence to CGGC
```

```
prot->mutate(0,0,C);

prot->mutate(0,1,G);

prot->mutate(0,2,G);

prot->mutate(0,3,C);


for (double psi = -180.0; psi < 180.0; psi += step)

{

prot->setPsi(0,0,psi);

setBackBone(prot, 1, step);

}

return 0;

}



void setBackBone(protein* _prot, UInt _pos, double _step)

{

if (_pos + 1 <  _prot->getNumResidues(0))

{

for (double phi = -179.0; phi < 180.0; phi += _step)

for (double psi = -179.0; psi < 180.0; psi += _step)

{

if (_pos != 0) _prot->setPhi(0,_pos,phi);

_prot->setPsi(0,_pos,psi);

setBackBone(_prot, _pos+1, _step);
```

```
}

return;

}

else

{

for (double phi = -179.0; phi < 180.0; phi += _step)

{

_prot->setPhi(0,_pos,phi);

checkSulfurGeometry(_prot);

}

return;

}

}


void checkSulfurGeometry (protein* _prot)

{

double psi0 = _prot->getPsi(0,0);

double phi1 = _prot->getPhi(0,1);

double psi1 = _prot->getPsi(0,1);

double phi2 = _prot->getPhi(0,2);

double psi2 = _prot->getPsi(0,2);

double phi3 = _prot->getPhi(0,3);


cout << psi0 << " " << phi1 << "," << psi1 << " " << phi2
```

```
<< "," << psi2 << " " << phi3 << " energy "

<< _prot->intraEnergy() <<  endl;


return;

}
```

AMBER 11   The structures obtained from the ProtCAD simulations were subjected to energy minimization calculations using Amber 11 (51), with a generalized Born solvent model (52,53). Protein atoms were described with the parm99SB (54,55,56) force field parameterization. The atomic charges were modified so that an oxidized $[Fe_4S_4]^{+2}$ cluster bound to 3 Cys had a net charge of -1, yielding the following charges: qFe = 0.6518 e, qS (cluster) = -0.5552 e, qSG (cysteine) = -0.6042 e. The maximum number of minimization cycles was set to 105, and the structures were considered minimized when the root-mean-square of the Cartesian elements of the gradient was less than 10^{-4} kcal/mol-Å. To compare the degree of electrostatic stabilization of the cluster in the different peptide models, the charge of the S atoms of the Fe_4S_4 cluster was set to zero, and a single point energy calculation was performed. A number of structures converged to an identical structure after the energy minimization process. The redundant structures were then removed by MMTSB (Multiscale Modeling Tools in Structural Biology) k-clustering algorithm (57).

3.6 Acknowledgement

3.7 References

1. Dupont CL, Yang S, Palenik B, Bourne PE (2006) Modern proteomes contain putative imprints of ancient shifts in trace metal geochemistry. Proceedings of the National Academy of Sciences of the United States of America 103: 17822-17827.

2. Caetano-Anolles G, Caetano-Anolles D, Kim KM, Mittenthal JE (2011) Proteome evolution and the metabolic origins of translation and cellular life. Journal of Molecular Evolution 72: 14-33.

3. Caetano-Anolles G, Kim HS, Mittenthal JE (2007) The origin of modern metabolic networks inferred from phylogenomic analysis of protein architecture. Proceedings of the National Academy of Sciences of the United States of America 104: 9358-9363.

4. Falkowski PG, Fenchel T, Delong EF (2008) The microbial engines that drive Earth's biogeochemical cycles. Science 320: 1034-1039.

5. Eck RV, Dayhoff MO (1966) Evolution of the structure of ferredoxin based on living relics of primitive amino acid sequences. Science 152: 363-366.

6. Zuckerkandl E, Pauling L (1964) Molecules as documents of evolutionary history. Journal of Theoretical Biology 8: 357-366.

7. Darimont B, Sterner R (1994) Sequence, assembly and evolution of a primordial ferredoxin from *Thermotoga Maritima*. Embo Journal 13: 1772-1781.

8. Otaka E, Ooi T (1987) Examination of protein-sequence homologies: IV. Twenty-seven bacterial ferredoxins. Journal of Molecular Evolution 26: 257-267.

9. Bruschi M, Guerlesquin F (1988) Structure, function and evolution of bacterial ferredoxins. Fems Microbiology Reviews 54: 155-175.

10. Beinert H (1990) Recent developments in the field of iron-sulfur proteins. Faseb Journal 4: 2483-2491.

11. Miller S (1953) A production of amino acids under possible primitive Earth conditions. Science 117: 528-529.

12. Bada JL, Johnson AP, Cleaves HJ, Dworkin JP, Glavin DP, et al. (2008) The Miller volcanic spark discharge experiment. Science 322: 404-404.

13. Holland HD (1984) The chemical evolution of the atmosphere and oceans. Princeton, N.J.: Princeton University Press. xii, 582 p. p.

14. Anbar AD, Knoll AH (2002) Proterozoic ocean chemistry and evolution: A bioinorganic bridge? Science 297: 1137-1142.

15. Williams RJP (1990) Overview of biological electron-transfer. Advances in Chemistry Series: 3-23.

16. Hazen RM, Filley TR, Goodfriend GA (2001) Selective adsorption of L- and D-amino acids on calcite: Implications for biochemical homochirality. Proceedings of the National Academy of Sciences of the United States of America 98: 5487-5490.

17. Wachtershauser G (1988) Before enzymes and templates - Theory of surface metabolism. Microbiological Reviews 52: 452-484.

18. Wachtershauser G (1988) Pyrite formation, the 1st energy-source for life - a hypothesis. Systematic and Applied Microbiology 10: 207-210.

19. Hazen RM, Sverjensky DA (2010) Mineral surfaces, geochemical complexities, and the origins of life. Cold Spring Harbor Perspectives in Biology 2.

20. Wachtershauser G, Huber C (1998) Peptides by activation of amino acids with CO on (Ni,Fe)S surfaces: Implications for the origin of life. Science 281: 670-672.

21. Huber C, Wachtershauser G (1997) Activated acetic acid by carbon fixation on (Fe,Ni)S under primordial conditions. Science 276: 245-247.

22. Nakajima T, Yabushita Y, Tabushi I (1975) Amino acid synthesis through biogenetic-type CO_2 fixation. Nature: 60-61.

23. Boehme C, Marx D (2003) Glycine on a wet pyrite surface at extreme conditions. Journal of the American Chemical Society 125: 13362-13363.

24. Smith JV (1998) Biochemical evolution. I. Polymerization on internal, organophilic silica surfaces of dealuminated zeolites and feldspars. Proceedings of the National Academy of Sciences of the United States of America 95: 3370-3375.

25. Milner-White EJ, Russell MJ (2008) Predicting the conformations of peptides and proteins in early evolution. A review article, Biology Direct. Biology Direct 3.

26. Copley SD, Smith E, Morowitz HJ (2005) A mechanism for the association of amino acids with their codons and the origin of the genetic code. Proceedings of the National Academy of Sciences of the United States of America 102: 4442-4447.

27. Coldren CD, Hellinga HW, Caradonna JP (1997) The rational design and construction of a cuboidal iron-sulfur protein. Proceedings of the National Academy of Sciences of the United States of America 94: 6635-6640.

28. Hellinga HW, Benson DE, Wisz MS, Liu WT (1998) Construction of a novel redox protein by rational design: Conversion of a disulfide bridge into a mononuclear iron-sulfur center. Biochemistry 37: 7070-7076.

29. Grzyb J, Xu F, Weiner L, Reijerse EJ, Lubitz W, et al. (2010) De novo design of a non-natural fold for an iron-sulfur protein: Alpha-helical coiled-coil with a four-iron four-sulfur cluster binding site in its central core. Biochimica Et Biophysica Acta-Bioenergetics 1797: 406-413.

30. DeGrado WF, Nanda V, Rosenblatt MM, Osyczka A, Kono H, et al. (2005) De novo design of a redox-active minimal rubredoxin mimic. Journal of the American Chemical Society 127: 5804-5805.

31. Noy D, Grzyb J, Xu F, Weiner L, Reijerse EJ, et al. (2010) De novo design of a non-natural fold for an iron-sulfur protein: Alpha-helical coiled-coil with a four-iron four-sulfur cluster binding site in its central core. Biochimica Et Biophysica Acta-Bioenergetics 1797: 406-413.

32. Noy D, Cohen-Ofri I, van Gastel M, Grzyb J, Brandis A, et al. (2011) Zinc-Bacteriochlorophyllide dimers in de Novo Designed Four-Helix Bundle Proteins. A Model System for Natural Light Energy Harvesting and Dissipation. Journal of the American Chemical Society 133: 9526-9535.

33. Noy D, Dutton PL (2002) De novo designed bacteriochlorophyll-protein maquettes based on bacterial light-harvesting complexes. Biophysical Journal 82: 321A-321A.

34. Liang CZ, Mislow K (1994) Topological Chirality of Proteins. Journal of the American Chemical Society 116: 3588-3592.

35. Liang CZ, Mislow K (1997) Topological chirality of iron-sulfur proteins. Biopolymers 42: 411-414.

36. Blonk HL, Vanderlinden JGM, Steggerda JJ, Geleyn RP, Smits JMM, et al. (1992) Synthesis of the Iron-Sulfur Cluster Compounds $[Fe_4S_4(Mecp)_4](PF_6)_Y$ (Y = 0-2) - X-Ray Structure Determinations of $Fe_4(Mu-3-S)_4(Mecp)_4$ and $[Fe_4(Mu-3-S)_4(Mecp)_4](PF_6)$. Inorganic Chemistry 31: 957-962.

37. Beinert H, Holm RH, Munck E (1997) Iron-sulfur clusters: Nature's modular, multipurpose structures. Science 277: 653-659.

38. Mulholland SE, Gibney BR, Rabanal F, Dutton PL (1999) Determination of nonligand amino acids critical to $[4\,Fe-4S]^{2+,/+}$ assembly in ferredoxin maquettes. Biochemistry 38: 10442-10448.

39. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) Scop - a Structural Classification of Proteins Database for the Investigation of Sequences and Structures. Journal of Molecular Biology 247: 536-540.

40. Summa CM (2002) Computational methods and their applications for de novo functional protein design and membrane protein solubilization. Doctoral Thesis University of Pennsylvania School of Medicine, Philadelphia.

41. Case DA, Cheatham TE, Darden T, Gohlke H, Luo R, et al. (2005) The Amber biomolecular simulation programs. Journal of Computational Chemistry 26: 1668-1688.

42. Blake PR, Lee B, Summers MF, Adams MWW, Park JB, et al. (1992) Quantitative Measurement of Small through-Hydrogen-Bond and through-Space H-1-Cd-113 and H-1-Hg-199 J-Couplings in Metal-Substituted Rubredoxin from Pyrococcus-Furiosus. Journal of Biomolecular Nmr 2: 527-533.

43. Stephens PJ, Jollie DR, Warshel A (1996) Protein control of redox potentials of iron-sulfur proteins. Chemical Reviews 96: 2491-2513.

44. Dey A, Francis EJ, Adams MWW, Babini E, Takahashi Y, et al. (2007) Solvent tuning of electrochemical potentials in the active sites of HiPIP versus ferredoxin. Science 318: 1464-1468.

45. Gordon DB, Marshall SA, Mayo SL (1999) Energy functions for protein design. Current Opinion in Structural Biology 9: 509-513.

46. Hoppe A, Pandelia ME, Gartner W, Lubitz W (2011) $[Fe_4S_4]^-$ and $[Fe_3S_4]^-$ cluster formation in synthetic peptides. Biochimica Et Biophysica Acta-Bioenergetics 1807: 1414-1422.

47. Pauling L, Corey R (1951) The Pleated Sheet, A New Layer Configuration of Polypeptide Chains. Proceedings of the National Academy of Sciences of the United States of America 37: 251-256.

48. Daggett V, Armen RS, Alonso DOV (2004) Anatomy of an amyloidogenic intermediate: Conversion of beta-sheet to alpha-sheet structure in transthyretin at acidic pH. Structure 12: 1847-1863.

49. Sagui C, Babin V, Roland C (2011) The alpha-sheet: A missing-in-action secondary structure? Proteins-Structure Function and Bioinformatics 79: 937-946.

50. Watson JD, Milner-White EJ (2002) A novel main-chain anion-binding site in proteins: The nest. A particular combination of phi,psi values in successive residues gives rise to anion-binding sites that occur commonly and are found often at functionally important regions. Journal of Molecular Biology 315: 171-182.

51. D.A. Case TAD, T.E. Cheatham, III, C.L. Simmerling, J. Wang, R.E. Duke, R. Luo, R.C. Walker, W. Zhang, K.M. Merz, B. Roberts, B. Wang, S. Hayik, A. Roitberg, G. Seabra, I. Kolossvai, K.F. Wong, F. Paesani, J. Vanicek, J. Liu, X. Wu, S.R. Brozell, T. Steinbrecher, H. Gohlke, Q. Cai, X. Ye, J. Wang, M.-J. Hsieh, G. Cui, D.R. Roe, D.H. Mathews, M.G. Seetin, C. Sagui, V. Babin, T. Luchko, S. Gusarov, A. Kovalenko, and P.A. Kollman (2010) AMBER 11. University of California, San Francisco.

52. Hawkins GD, Cramer CJ, Truhlar DG (1995) Pairwise solute descreening of solute charges from a dielectric medium. Chem Phys Lett 246: 122-129

53. Hawkins GD, Cramer CJ, Truhlar DG (1996) Parametrized Models of Aqueous Free Energies of Solvation Based on Pairwise Descreening of Solute Atomic Charges from a Dielectric Medium. J Phys Chem 100 1982419839.

54. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, et al. (1995) A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules Journal of the American Chemical Society 117: 5179-5197.

55. Wang J, Cieplak P, Kollman PA (2000) How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? Journal of Computational Chemistry 21: 1049-1074.

56. Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, et al. (2006) Comparison of multiple Amber force fields and development of improved protein backbone parameters. Proteins 65: 712-725.

57. Feig M, Karanicolas J, Brooks CL (2004) MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology. Journal of Molecular Graphics & Modelling 22: 377-395.

Chapter 4

Discovering the Electronic Circuit Diagram of Life: The Origins and Evolution of Biological Electron Transfer Reactions

Oxidoreductases play a central role in catalyzing enzymatic electron transfer reactions across the tree of life. To first order, the equilibrium thermodynamic properties of these proteins are governed by protein folds associated with specific transition metals and ligands at the active site. A global analysis of holoenzyme structure and function suggests there are fewer than ~500 fundamental oxidoreductases, which can be further clustered into 35 unique groups. These catalysts evolved in prokaryotes early in Earths history and are largely responsible for emergent non-equilibrium biogeochemical cycles on the planets surface. Although the evolutionary history of the amino acid sequences in the oxidoreductases is very difficult to reconstruct due to gene duplication and horizontal gene transfer, the evolution of the folds in the catalytic sites potentially be used to infer the history of these enzymes. Using a novel yet simple analysis of the secondary structures associated with the ligands in oxidoreductases, we developed a structural phylogeny of these enzymes. The results of this composome analysis suggest an early split from a basal set of a small group into two families of oxidoreductases, one dominated by α-helices and the second by β-sheets. The structural evolutionary patterns in both clades trace redox gradient and increased hydrogen bond energy. The overall pattern suggests that the evolution of the oxidoreductases led to decreased entropy over ~2.5 billion years, allowing the enzymes to utilize increasingly oxidized substrates with high specificity.

4.1 Introduction

Biologically-driven, electron transfer reactions are the primary energy transduction processes across the tree of life. These reactions depend upon external energy sources. The two external energy sources on Earth are solar radiation and geothermally derived heat.

These create chemical redox gradients, which are coupled to biological redox reactions and ultimately drive the non-equilibrium thermodynamic reactions that make life possible. Indeed, the origin of life almost certainly began with the evolution of a small set of metabolic processes coupled to redox chemistry (1).

Oxidoreductases (Enzyme Commission 1; EC1) are a class of enzymes that facilitate these proton coupled electron transfer reactions. All of the core metabolic processes mediated by EC1 proteins evolved in prokaryotes and ultimately became coupled on local and planetary scales to facilitate an electron market between the major light elements. This electron market ultimately led to a closed cycle between respiratory reactions and their biological oxidative analogues, and photosynthesis and their biological reductive analogues. These reactions, which are far from thermodynamic equilibrium, allow gas exchanges across the tree of life and transformed Earth's biogeochemical cycles (Fig. 4.1) (1).

Although the exact number of core EC1 proteins, including orthologs, paralogs and analogs is unknown, their functions appear to be encoded by fewer than $\sim$500 unique genes (i.e, genes encode for unique functions including paralogs and analogs) (Fig. 4.1). One reason that such a small set of proteins plays such a large role in Earths elemental cycles is that the different electron transport chains found ubiquitously throughout life share common components. For example, chemoautotrophic, anaerobic and aerobic respiration, as well as anoxygenic and oxygenic photosynthesis all utilize a proton coupled electron transport scheme. The basis of the scheme is separation of protons from electrons across an energized membrane. The electrons are inevitably ferried via carriers, across a set of membrane bound proteins ultimately arriving at a transient sink that allows a negatively charged carrier to be neutralized by a proton. The
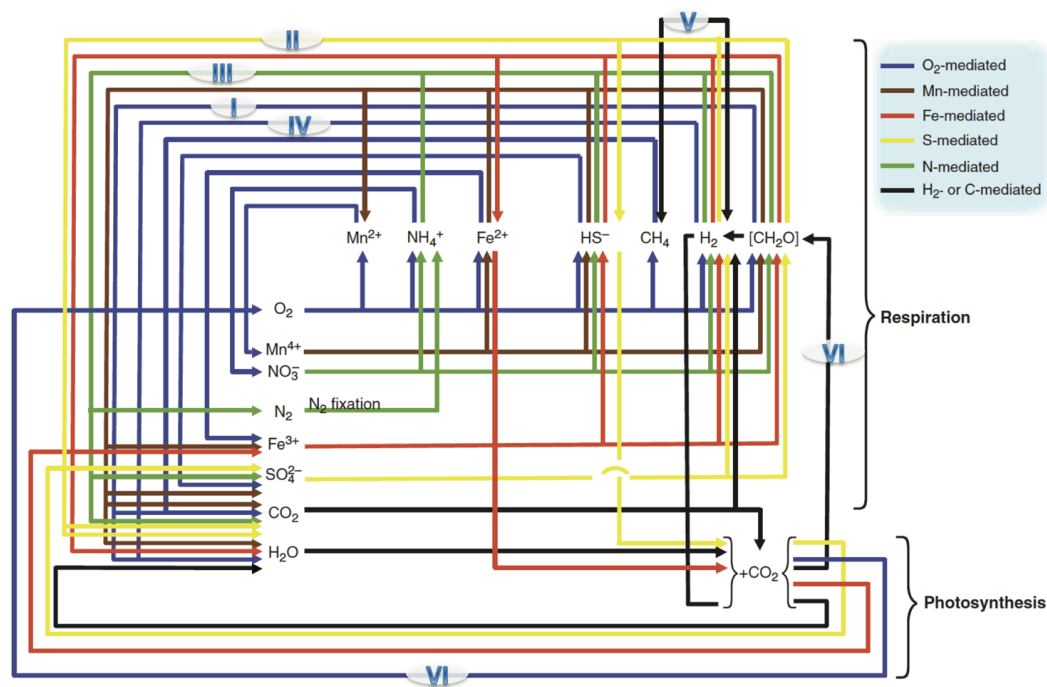
Figure 4.1: Network of life's biologically mediated cycles modified from (1). The interconnected "circuit board" of life's electron transfer reactions (for C, H, O, N, S and Fe) is labeled with pathway groups I to VI.

protons are initially segregated from the electrons by the membrane, thereby forming an asymmetric distribution of charge. The return flow of the protons (i.e., the proton motive force) is coupled to nanomachines, especially the coupling factor, ATP synthase, which conserves the electrochemical energy as chemical bond energy.

Though peripheral components of these proton coupled electron transport reactions have been selected for specific reaction substrates and products, the basic architecture of all the core pathways shares similar protein structures and ligands, including iron-sulfur clusters, pterins, hemes and quinones. These interchangeable structures and ligands have evolved into a metabolic network with overlapping functions across the tree of life (Fig. 4.2). Additionally, all biological energy transduction systems share a small subset of metabolic pathways such as glycolysis, pyruvate oxidation, or the TCA cycle. Thus, metabolism, using a variety of electron donors and acceptors, draws catalysts from a core set of similar components and pathways to enable a flow of electrons and protons (2). This modular approach to metabolism has provided great flexibility on a relatively small number of EC1 genes. Indeed, prokayotes are often able to regulate major components of metabolism in accordance with environmental conditions (3), often at suboptimal efficiency.

Regardless of efficiency, the flux of electrons through the metabolic network is particularly dependent on, and sensitive to the availability of specific transition metals, especially iron (Fig. 4.3). The bioavailability of transition metals is, in turn, highly dependent on the redox state of the environment. A recent whole-genome analysis of phylogenetically diverse microorganisms suggests that the earliest proteins incorporated metals, and that metal usage over time evolved in accordance with environmental availability (4). The metals are invariably coordinated to the protein scaffolds via a small
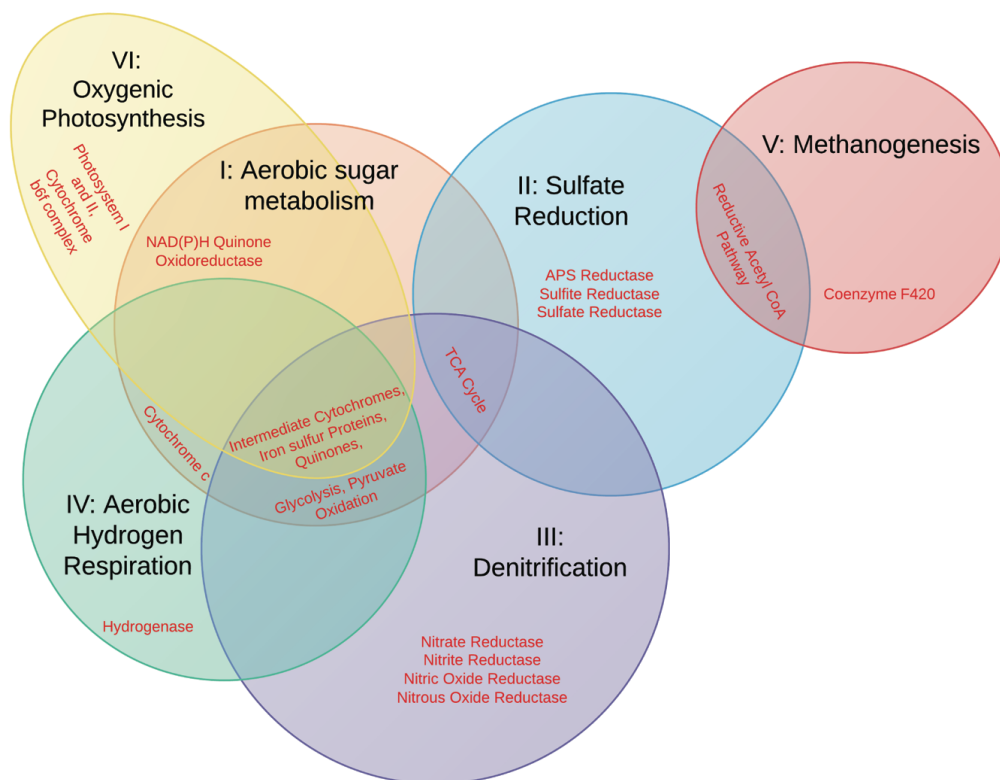
Figure 4.2: Venn diagram of pathway groups I to VI. Pathway groups I to VI are shown to have redox components of electron transport chains as well as auxiliary pathways in common. Size of circles is irrelevant.
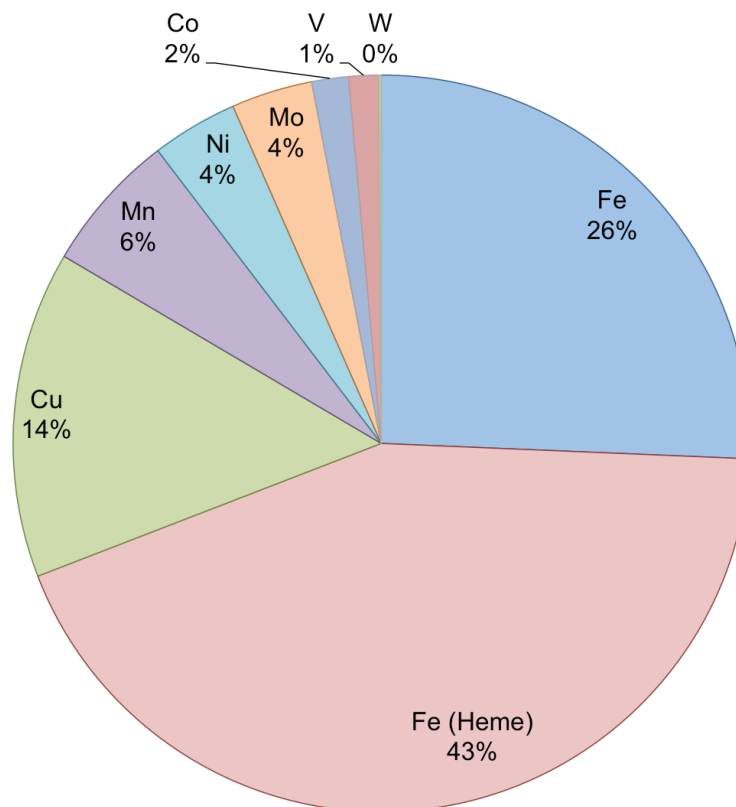
Figure 4.3: The relative contribution of the major transition metals found in PDB structures annotated as oxidoreductases

set of specific protein folds (1). Identifying members and the evolutionary pattern of this set of folds is critical to understanding the evolution of metabolism across the tree of life, as well as the emergence of biogeochemical cycles, far from equilibrium.

In this paper we present an analysis of the evolutionary history of metal usage, the structures of the protein folds, and redox state of the oxidoreductases across the tree of life, which ultimately formed an electronic circuit on a planetary scale. Our results suggest that the that the redox processes connecting metabolism across Earths surface underwent a secular trend in evolutionary transitions that led to successively greater complexity and thermodynamic efficiency in these critical enzymes over the first ~ 2.5 billion years of Earths history.

The vast majority of oxidoreductases catalyze single electron transfer reactions via prosthetic groups containing transition metals that undergo reversible changes in redox state. The transition states are, in turn, controlled by and coordinated to the ligands on the protein scaffold. Transition metals are specifically incorporated into the active sites via coordination bonds, in which the electronic states are three-dimensionally delocalized around the metal. Such a property is unique for d and f orbitals, allowing a high coordination number with multiple coordination geometries. In particular, d-orbitals are highly symmetrical and often capable of populating several orbital states, thereby conferring highly localized degenerate and pseudo-degenerate energy levels. This results in two types of magnetic properties (high/low spin), vibrant colors and multiple coordination numbers (5). Further, multiple oxidation states arising from the operational flexibility in d-orbital transitions allows the reduction-oxidation potential within a specific metal-ligand structure to be tuned," or poised, for specific redox reactions.

Among transition metals, specific elements were naturally selected for their physical-chemical properties, abundance, coordination bond strength, atomic radii, solubility, and polarizability (6). Iron is by far the most common transition metal found on Earth (7). The abundance of this element, especially as a ferrous ion in the Archean and early Proterozoic oceans, is reflected in its wide use as a catalyzing co-factor for oxidoreductases (Fig. 4.3). In Swiss-Prot (8) and PDB (9), iron is identified as a catalytic component in more that 67 % of the EC1 proteins. Iron containing oxidoreductases can employ the metal in a mineral form of a sulfide, or coordinated to imidazole nitrogens in porphyrins, forming hemes. Following iron, oxidoreductases contain metals in the following order of relative abundance: Cu > Mn > Ni> Mo > Co > V > W (Fig. 4.3). It should be noted that under anoxic and/or euxenic conditions, Cu and Mo are highly

insoluble unless the ions are oxidized.

Biopolymer-metal Interactions in the Primordial Soup   Assuming that the generation of energy via dissipation of redox gradients was an early bioinorganic reaction essential for the origin of life, it logically follows that transition metals played a key role. While transition metals can undergo stoichiometric reactions via photochemical processes or in solution phase with other redox couples (10-13), the reactions only become catalytic when surrounded by a biopolymer that provides a specific structural framework facilitating reversible population of the metal ligands with electrons. Lewis basic peptide side chains that include thiolates (cysteine), imidazole nitrogens (e.g. histidines or porphyrins), and carboxylates (aspartic acid, glutamic acid) can continuously provide a source of electrons to d-orbitals in transition metals to form σ-bonds. In turn, electrons in the d-orbitals in the transition metals form π-bonds with the ligands to further enhance the coordination bond energy. Metals bound by multiple side-chains are locked within the peptide/protein matrix. These interactions influence multiple physical properties of the holoprotein, including solvent accessibility, tuned redox potential, optimization of Gibbs free energy, and enhanced substrate specificity. Understanding how the earliest biopolymer-metal interactions evolved is critical to understanding the origins of non-equilibrium bioenergetic reactions, and hence the origins of life. In a primordial soup model of the early Archean (14), metals in minerals may have played a significant role in adsorbing and concentrating organic molecules and catalyzing various chemical reactions implicated in the origin of non-equilibrium redox reactions. Provided with the building blocks of life, metals bound to short peptides could have functioned as protoenzymes, as is proposed by models of early protein evolution (15, 16). An early

protoenzyme would have had to originate and evolve under a strict set of rules:

1. The electronic structure of transition metals must match geometric requirements for metal-ligand coordination number and geometry. As a result, the emergent entactic states constrain the subsequent evolution of the structures and topologies of the coordinately bound polypeptides or other molecules required for catalysis.

2. The mildly reducing environment of the ancient oceans required a relatively low midpoint potential of the early redox organocatalysts, especially compared with more recently evolved oxidoreductases. The ligand environment was modulated to tune the midpoint potential to meet the functional requirements.

To meet these constraints, random polypeptides almost certainly evolved in association with the available metals and continuously selected for sequences and folds that satisfied both structural and functional constraints. Such a combinatorial search to local minima would have persisted until the polypeptide-metal complex acquired locally optimized catalytic functions. The specific folds almost certainly co-evolved over geological time with an increasing larger set of coupled biogeochemical cycles. The ancient folds were spread genetically across the nascent tree of life primarily via horizontal gene transfer but ultimately diverged into several motifs. The subsequent structural innovations were accelerated by various modes of evolution such as gene insertion, duplication, and partial loss. Evolved core protein folds became molecular modules from which a variety of biomachines could ultimately be built via a mix and match set of motifs.

Evolution of Sequences and Folds   Because of both their modularity and early spread across the tree of life, it is extremely difficult to determine the evolutionary

heritage of folds in the oxidoreductases via analysis of organisms, sequences or synteny. Inspection reveals major sequence degeneracy in closely related structures. For example, trans-membrane spanning helices of photosystem I and II have highly divergent sequences, yet their structures are almost identical. This basic phenomenon was noted early on by pioneers in the field of bioinformatics. Indeed, in 1966, Eck and Dayhoff (17) noted The processes of natural selection severely inhibit change to a well-adapted system on which several other essential components depend. While their comments were based on the highly conserved structure of ferredoxin, they apply to many ancient proteins, including enzymes that do not catalyze redox reactions. For example, Ribulose-1,5-bisphosphate carboxylase oxygenase (RuBisCO) (EC 4.1.1.39) is a carboxylyase. The enzyme is responsible for the fixation of CO_2 in many photosynthetic and chemoautotrophic organisms. This crucial enzyme cannot easily distinguish between its true substrate CO_2 and O_2. The result is that at present atmospheric levels of O_2, the enzyme is often remarkably inefficient. Moreover, the catalytic turnover of the reaction, even under optimal conditions, is much slower than reactions feeding the substrate (CO_2) or removing the product (3-phospho glycerate). Regardless, the catalytic site of the enzyme is highly conserved and the biological result of this conservation is that organisms often synthesize the enzyme in excess to achieve maximum overall growth efficiency (18). There are many other, similar examples. The fundamental physicochemical properties that govern the major protein fold conformations have remained unchanged. Reinvention of metalloenzyme folds is highly restricted, given that geometric and energetic selection processes limit structural solutions. Let us examine a novel approach to identifying and ordering the structural solutions found in extant oxidoreductases.

4.2 The Composome Approach

We hypothesize that the ensemble of secondary structures in the region surrounding the catalytically active metals has been selected to facilitate catalysis of the holoenzyme. We further hypothesize that these secondary structures are the outcome of selection and provide a window into the processes in which protein folds evolved. We assume that the composition of the secondary structural motifs reflects the evolutionary history of the protoenzyme from which the extant motif is descended, and was inherited with modifications through a myriad of organisms to form the observed protein fold. We further assume that the folds must obey the rules set by the d-block metal coordination chemistry (19). These underlying hypotheses are extensions from our previous work where we proposed that secondary structures around the metal or metal ligand in the active site would be more conserved than elsewhere in the protein (20). We call this quantitative analysis of the secondary structure of the folds in active sites the elucidation of a composome." To our knowledge, this is the first attempt to infer quantitative distances between distinctly different protein folds based secondary structural compositions. The resulting phylogeny represents a linkage of fold relationships in structural space, and, obviously is not a linear history of the evolution of EC1 proteins. The approach uses PDB (Protein Data Bank) files that possess previously determined gold standard domains (21). From this data set we extracted a subset of representatives using the best resolution structure for each organism per gold standard domain. For every domain, the corresponding metal-ligand is treated as the catalytic site. For each catalytic site, we first collected a list of amino acid residues that are within 15 Å from a catalytic metal, based on carbon alpha ($C\alpha$) coordinates. Secondary structures of
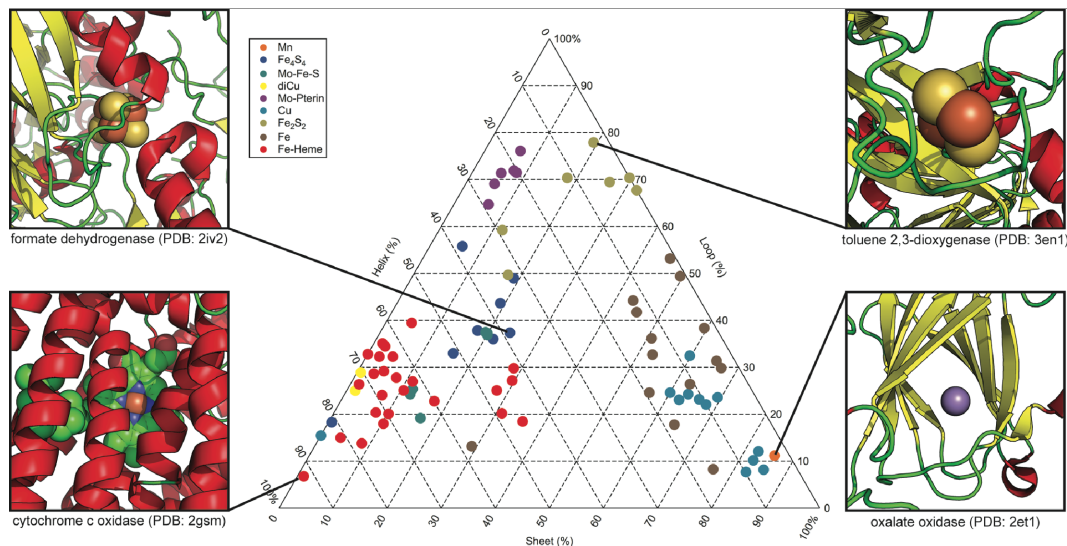
Figure 4.4: Ternary vector space used for the secondary structure composition (composome) and representative structures. Representative structures are shown for each of the major folds in the ternary space.

the residues were assigned using the DSSP database (22). The overall secondary structure composition (i.e., the composome) around each metal bearing catalytic site was determined and further adjusted based on a residue-metal distance by a factor of $1/r^2$. These secondary structural compositions, from each catalytic site, were plotted in a ternary vector space (Fig. 4.4). Structures that contained both identical metals and a Euclidean distance in composome space less than 0.02 Å were collapsed into a single representative structure, resulting in 82 final representative structures. Each data point in this ternary space diagram represents an individual oxidoreductase. The data points largely cluster into two groups with helix rich and sheet rich clades. For each metal or metal-ligand, data points tend to aggregate, suggesting that the physicochemical properties of metals constrain the entactic evolution toward specific compositions of secondary structure around the catalytic site.

The compositional finger printing method described for calculating relations in folds

across all known EC1 structures collapses three-dimensional information into a one-dimensional matrix. To assess the effects of this mathematical simplification and compression, backbone environments for the 82 structurally different catalytic sites were superimposed in a pairwise, all vs. all fashion using an algorithm that compares the composome approach with more complex analyses of the ternary structures (23). The similarity values of highly similar structure pairs (> 40 % structurally similar) correlate with low compositional profile distances (Fig. 4.5). This result strongly suggests that in principle, secondary structures retain sufficient topological information such that a quantitative analysis of the fold can retrieve conformational relationships with sufficient resolution and confidence to derive the structural history of the folds in catalytic sites. Indeed, for dissimilar structural pairs, secondary structure composition maintains a high degree of predictive performance. We find significantly lower secondary structure composition distances for environments of same ligands compared to different ligands (Fig. 4.5), especially for pairs where the calculated Cα backbone structure similarity is less than 40 % and therefore inconclusive. Very high composome distances are generally not observed for same ligand pairs hinting.

There appears to a limit of the differences between same ligand environment secondary structures. This phenomenon appears to be related to hydrogen bonding patterns that are required by certain electronic structures around metal or metal ligand catalytic sites, leading to similar secondary structure composition signatures. Let us now examine the fold phylogeny derived from the composome analysis.
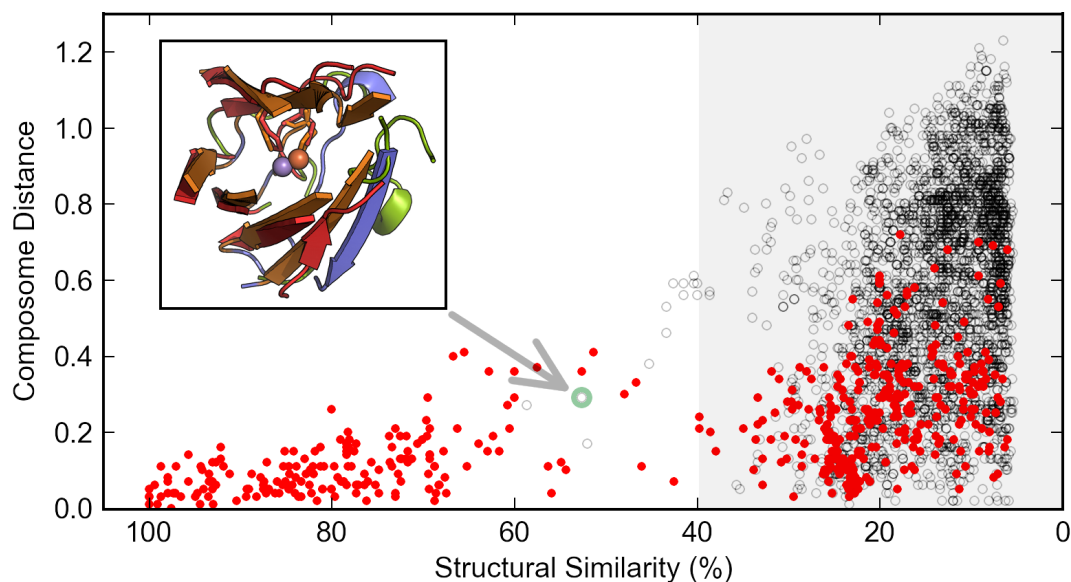
Figure 4.5: Structural similarity and composome distance (described in text) are compared in a dot plot of 3,240 alignments of 82 metal catalytic sites. The 447 alignments between the same metal co-factors are colored red. Alignments that are less than 40% structurally similar (gray shaded area) can be considered inconclusive in terms of structural similarity. The insert shows an alignment between the manganese-binding motif in Oxalate oxidase (2et1, EC 1.2.3.4; orange for matched, blue for unmatched) and the iron-binding motif of clavaminate synthase (1ds1, EC 1.14.11.21; red for matched, green for unmatched). The folds are very similar in secondary structure content and overall fold (52.64% structurally similar, RMSD 1.86 Å). The sequence identity derived from the 50 residue long alignment is 6%, or 3 residues, two of which are the metal coordinating histidines (illustrated in stick representation).

4.3 Fold Phylogeny Based on Composome

A polypeptide chain has an intrinsic property of having dynamic conformational variability, but specific structures with relatively rigid folds are often conferred for specific biological functions. Fold evolution is achieved when a protein structure with a functional promiscuity has a flexible chain, which is evolvable (24). To check the evolvability of each protein fold, we calculated the average hydrogen bond energy per residue around each metal site. We postulated that the degree of hydrogen bond is inversely proportional to the evolvability of the protein fold. The analysis (Fig. 4.6) suggests that loop rich folds have low hydrogen bond energy per residue, making the fold more evolvable; loops and coils tend to be more flexible. In contrast, α-helices or β-sheets form more extended hydrogen-bonding networks, making the fold far more rigid. While evolution can obviously still occur in these folds the resulting structures are highly conserved. Mixed usage of both helices and sheets require adjoining loop regions. The result is a variable degree of hydrogen bonding across peptide backbone. Protein folds with both helix and sheet are expected to have high evolvability. In our analysis, we observe two evolvability hot spots (Fig. 4.6). One such area is occupied by a ferredoxin fold, which might be one of the earliest structures (17), whereas the other hotspot does not exist among extant folds.

Based on the evolvability, we developed a Bayesian prior that fold evolution occurred from more evolvable to less evolvable folds. Loop and coil rich protein structures, lacking secondary structure would be located close to the root of the phylogenetic tree of electron transfer folds. Using the secondary structure compositional vectors for each protein fold, a matrix of Euclidian distances was generated and a phylogenetic tree of protein folds (Fig. 4.7) was calculated using the Fitch-Margoliash algorithm and global
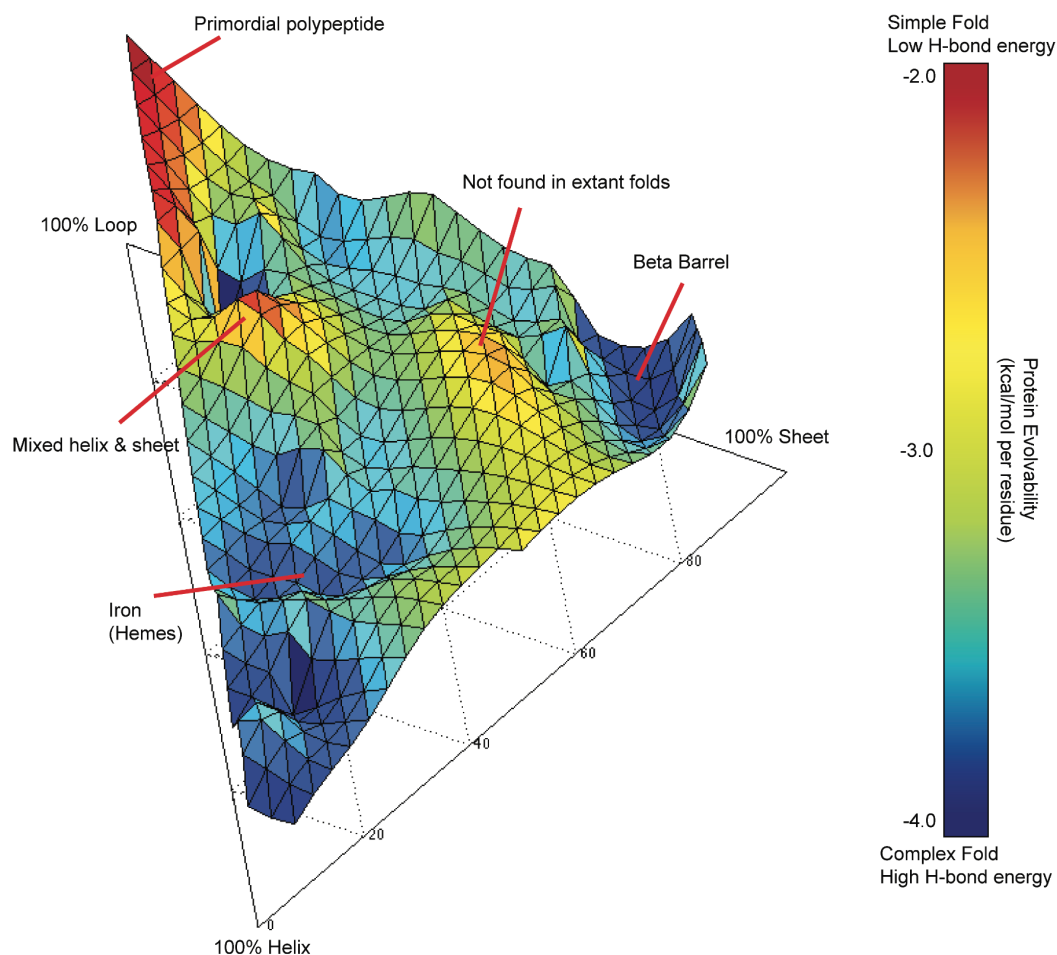
Figure 4.6: Average hydrogen bond energy per amino acid residue associated with the metal/metal-ligand center across the composome surface. Hydrogen bond energy is a proxy for fold evolvability. The evolvability contour plot is generated based on average hydrogen bond energy per residue from 82 data points and overlayed on the ternary plot (Fig. 4.4). As protein folds acquire more helix or sheet secondary structures, they become less flexible and their folds are increasingly fixed. The evolution of fold occurred starting from simple disordered state (red) to complex organized state (blue).

tree optimization as provided by the PHYLIP package (25). Rieske fold was chosen as a root for building a monophyletic tree. It is impossible to prove that the actual evolution of protein fold occurred in a monophyletic fashion. The tree arranges protein folds in the order of complexity based on a Bayesian prior (hydrogen bond energy).

Iron-Sulfur Proteins   Iron-sulfur containing ferredoxin folds are located near the center of the ternary plot, indicating relatively equal amounts of helix, sheet and loop composition. The location of iron-sulfur proteins on a ternary plot coincide with high a evolvability hotspot, suggesting ferredoxin folds might be the common ancestor of many extant protein folds for a number of reasons. First, iron sulfur minerals were thought to be relatively abundant in the early Archean ocean and it has been speculated that iron-sulfur clusters played an important role in the evolution of bioenergetic redox transduction systems (12). These hypotheses postulate that the earliest biologically relevant redox reactions occurred on iron-sulfur mineral deposits associated with hydrothermal vents. Second, ferredoxin is found across the tree of life, and is often a domain of larger associated in domains in large proteins, including many of which are those encoded by the core redox genes of life. Third, sequence and structural symmetry of ferredoxins suggests that they may have evolved from a gene duplication event of a 28-30 amino acid sequence, each capable of binding one iron-sulfur cluster. Sequence analyses by Eck and Dayhoff revealed even shorter repeats of four amino acids, suggesting a prebiotic protoferredoxin that was potentially composed of a primeval subset of the twenty amino acids (17). Fourth, all ferredoxins have a simple, conserved fold that binds two Fe_4S_4 clusters and is composed of fifty to sixty amino acids. Lastly, the earliest microorganisms, such as methanogens, may have utilized hydrogen gas as
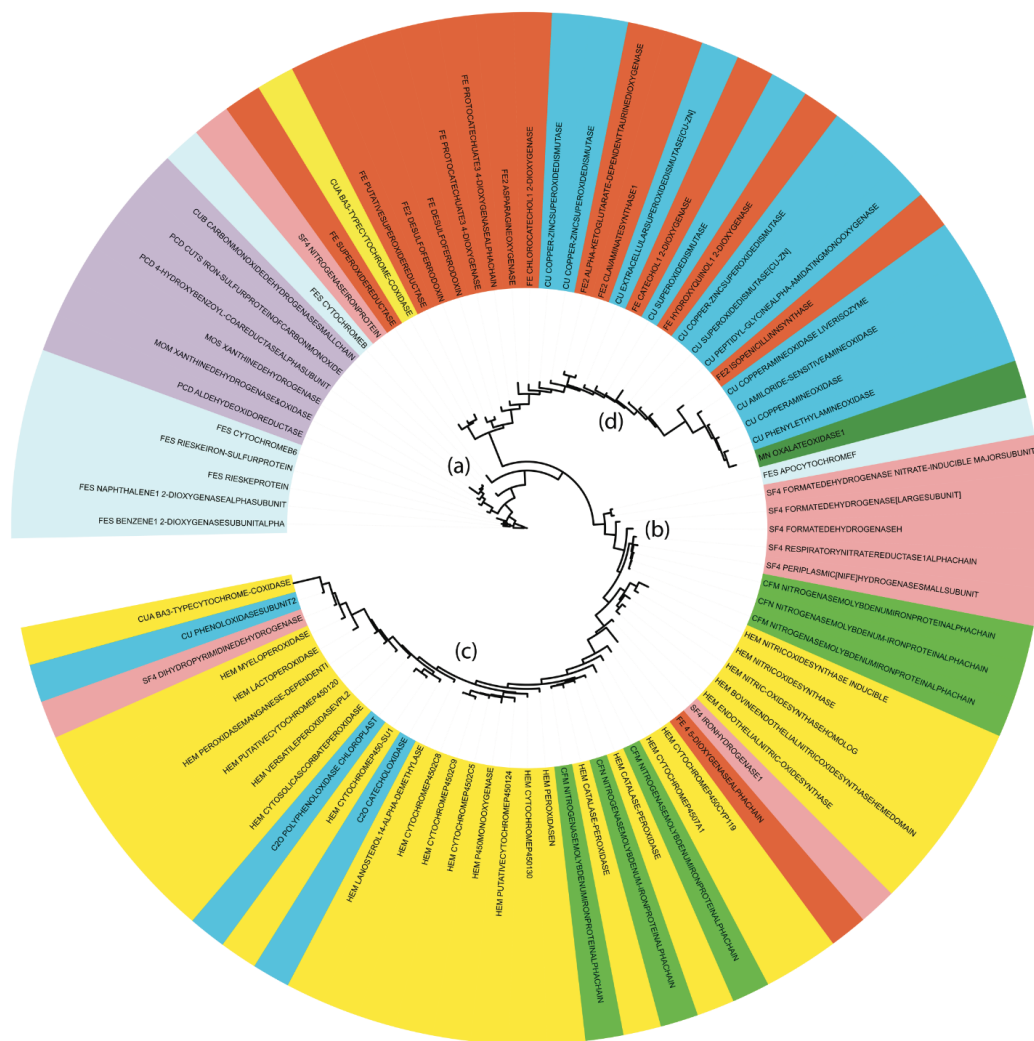
Figure 4.7: A tree of protein folds based on evolvability and a Euclidean distance matrix based on the composome analysis. In a primordial soup, polypeptide chains gain catalytic ability through random mutations. The earliest protein folds would resemble extant Rieske or molybdopterin folds, which have high evolvability (a). Conformational searches yields folds with both α-helix and β-sheet structures, such as ferredoxins (b). Further conformational searches result in the divergence of protein structure, giving rise to α-helix rich (c) and β-sheet folds (d).

electron donor using hydrogenase, which has a ferredoxin fold at its catalytic site (26).

Molybdopterin & Rieske Proteins   The lack of a helix/sheet is characteristic of Rieske iron-sulfur containing proteins and molybdopterin proteins. The former almost certainly is older. Regardless, it is unclear how these proteins carry out specific biological functions without a fixed helix or sheet secondary structure, but our analysis suggests these folds have the highest evolvability potential (Fig. 4.6). Unlike nitrogenases, molybdopterin proteins, such as DMSO reductase or nitrate reductase have a molybdenum atom chelated by pterins, which is surrounded by a protein environment that lack either α-helix or β-sheet. In the absence of Mo, studies have shown tungsten can replace molybdenum, due to its similar chemical properties, and it is possible that tungsten containing proteins gave rise to molybdenum containing protein folds in a prebiotic world.

Molybdenum-Iron-Sulfur Proteins   Reduction of N_2 to NH_3 is catalyzed by nitrogenases, the modern forms of which contain molybdenum iron-sulfur clusters. Molybdenum containing nitrogenase folds are closely located near ferredoxin folds in a ternary composome space, indicating their similarity in the secondary structure composition with mixed α-helix and β-sheet. Composome analysis suggests the Mo-Fe-S cluster containing folds may have evolved from Fe-S folds. Both iron-sulfur folds and molybdenum-iron-sulfur folds have α-helix and β-sheet secondary structure elements with a significant loop content, adding flexibility to the core structure. As a result, these folds may have diversified by exploring different conformations and compositions, mainly diverging into two distinct clades: an α-helix rich clade and a β-sheet rich clade.

Rubredoxin & Mn & Cu Proteins The Fe atom is found alone or as a ligand complex with a porphyrin ring in most extant folds. Rubredoxin is a single iron-containing fold with high β-sheet content. Manganese and copper folds also appear at the far edges of the composome-space in the β-sheet clade. These folds form extensive hydrogen bond networks across backbone amide and carbonyl, making the fold less evolvable. In return, these folds gained high catalytic specificity and accuracy with rigid structure, but lost their evolvability.

Heme Proteins Heme proteins are a hallmark of electron transfer proteins. Unlike rubredoxins, heme folds have an iron atom residing in a porphyrin ring surrounded by α-helices. The iron atom has octahedral coordination geometry, where axial positions are ligated to histidine side-chains. A surrounding porphyrin ring allows iron atoms to be incorporated into a protein scaffold with just one amino acid residue binding the iron directly. The usage of porphyrin rings may have relieved the stringent geometric requirement of the iron coordination (i.e. Rubredoxin), allowing an explosive diversification of heme folds. Heme is one of the most abundant protein cofactors in oxidoreductases (Fig. 4.3) and widely covered in composome ternary space (Fig. 4.4).

4.4   Conclusion

Life is dependent on catalytic function of a relatively small set of core proteins. Within this core set, oxidoreductases play an outsized role but their evolutionary history is poorly understood. Our brief analysis conforms to the hypothesis that the evolution of the oxidoreductases is, like for all proteins, a tradeoff between enzyme specificity and evolvability. Assuming the evolutionary trajectory followed from disordered state to

ordered state, proteins appear to have evolved from loop rich into either α-helix rich or β-sheet rich, becoming more specific but less evolvable. The early split between the two major clades of oxidoreductases continued to follow a decrease in the free energy within the active site a long term displacement of internal entropy that coincides with an increase in the redox potential of the active sites. The selection pressure that led to the increased hydrogen bonding energy concordant with increased redox potential in oxidoreductases may be an outcome of evolution, or a result of a feedback between the result of the ignition of a non-equilibrium thermodynamic system that is self-sustaining with a positive feedback. Regardless, the redox space explored by biology over the past two billion years appears to be extremely limited. We conclude that life has evolved a very small set of key core catalysts, the genes of which are transmitted across vast expanse of geological time by microbes, allowing a fundamental set of electron-transfer reactions to permit energy extraction from an open system. How these catalysts came to be coupled in specific sequences and with other chemical reactions to obviate non-equilibrium systems remains a major challenge to understanding the origins of life on Earth.

4.5 Methods

Composome Code The composome algorithm takes a list of pdb and DSSP files. For each pdb file, the program builds a 15 angstrom shell around the metal center and calculates a composition score (helix, sheet, loop) based on DSSP assignments. Manually curated PDB files are provided with a metal center coordinates. A center of mass is chosen for cofactors with multiple metal atoms (e.g. Fe_4S_4).

```perl
#!/usr/bin/perl -w

use Math::Trig;

use warnings;


sub trim($);

sub getrefcoord;

sub getShell;


$shellsize = 15;


$extension = ".pdb";

$extension2 = ".dssp";

$pdbfile = $ARGV[0].$extension;

$dsspfile = $ARGV[0].$extension2;


$ligandid = trim($ARGV[1]);

$ligandchain = $ARGV[2];
```

```perl
$ligandnum = $ARGV[3]+0;


print "open... $pdbfile\n";


open(PDB, "<", "$pdbfile") or die("cannot open the $pdbfile\n");

open(DSSP, "<", "$dsspfile") or die("cannot open the file\n");

open(OUT, ">>", "result_size$shell.txt")

or die("cannot write the output file\n");


@LIG=();

@hetarray=();

@atomarray=();

@dssparray = ();


while ($dssp = <DSSP>)

{

push @dssparray, $dssp;

if ($dssp =~ m/^COMPND/)

{

$name = substr($dssp, 21, 70);

}

}


while ($data = <PDB>)
```

```perl
{

if ($data =~ m/^HETATM/)

{

push @hetarray, $data;

}


if ($data =~ m/^ATOM/)

{

push @atomarray, $data;

}

}


foreach $line (@hetarray)

{

$ligandofinterest=trim(substr($line, 17, 3));

$ligand_chain =substr($line, 21,1);

$ligandnumber=substr($line, 22, 4)+0;

$metaltype=substr($line, 76, 2);


if ($metaltype eq "FE" or

$metaltype eq "MN" or

$metaltype eq "ZN" or

$metaltype eq "CU" or

$metaltype eq "MO")
```

```
{

if ("$ligandofinterest" eq "$ligandid" and

"$ligandnumber" == "$ligandnum" and

"$ligandchain" eq "$ligand_chain")

{

print "Metal of interest\n$line\n";

push @LIG, $line;

}

}

}


@refcoord = getrefcoord(@LIG);

print "Reference Coords (X,Y,Z):

 $refcoord[0]\t$refcoord[1]\t$refcoord[2]\n ";


#print OUT "$pdbfile\n";


$count = 0;

$Gscore = 0;$Hscore = 0;$Iscore = 0;

$Bscore = 0;$Escore = 0;$Sscore = 0;

$Tscore = 0;$totalscore = 0;$hbenergy = 0;

$hbtotal = 0;
```

```perl
foreach $line (@atomarray)

{

$carbonalpha=substr($line, 13, 2);


if ($carbonalpha =~ m/^CA/)

{

$shelly = getShell($line, $refcoord[0], $refcoord[1], $refcoord[2]);

if ($shellsize > $shelly)

{

$score = 1/($shelly*$shelly);

$chain = substr($line, 21,1);

$num = substr($line, 22, 4)+0;


foreach $dsspline (@dssparray)

{

$breakindssp = substr ($dsspline, 13, 1);

if ($breakindssp !~ m/!/ or $breakindssp !~ m/#/)

{

$getchain = substr($dsspline, 11, 1);

$getnum = substr($dsspline, 5, 5);

if ($getchain eq $chain and $getnum == $num)

{

$count++;

$sec = substr($dsspline, 16, 1);
```

```perl
if ($sec eq "G"){$Gscore = $Gscore+$score;}

elsif($sec eq "H"){$Hscore = $Hscore+$score;}

elsif($sec eq "I"){$Iscore = $Iscore+$score;}

elsif($sec eq "B"){$Bscore = $Bscore+$score;}

elsif($sec eq "E"){$Escore = $Escore+$score;}

elsif($sec eq "S"){$Sscore = $Sscore+$score;}

elsif($sec eq "T"){$Tscore = $Tscore+$score;}


$hbenergy1 = substr($dsspline, 46, 4);

$hbenergy2 = substr($dsspline, 57, 4);

$hbenergy3 = substr($dsspline, 68, 4);

$hbenergy4 = substr($dsspline, 79, 4);

$hbtotal = ($hbenergy1 + $hbenergy2 + $hbenergy3 + $hbenergy4);

$hbenergy = $hbenergy + $hbtotal;




$helixscore = $Gscore + $Hscore + $Iscore;

$sheetscore = $Bscore + $Escore;

$loopscore = $Sscore + $Tscore;

$totalscore = $helixscore + $sheetscore + $loopscore;


#print OUT "$dsspline\n";
```

```
        }

        }

        }

        }

        }


    $helix = sprintf("\%4.3f", $helixscore/$totalscore);

    $sheet = sprintf("\%4.3f", $sheetscore/$totalscore);

    $loop = sprintf("\%4.3f", $loopscore/$totalscore);

    $ligandid = sprintf("\%3s", $ligandid);

    $hbenergy = sprintf("\%4.3f", $hbenergy/$count);

    $name = trim($name);



    print OUT "$pdbfile $ligandid\t$helix\t$sheet\t$loop\t$hbenergy\t$name\n";



    close(PDB);

    close(DSSP);

    close(OUT);



    sub trim($)

    {
```

```perl
my $string = shift;

$string =~ s/^\s+//;

$string =~ s/\s+$//;

return $string;

}




sub getrefcoord

{

$sum_x = $sum_y = $sum_z = 0;

    my @r = ();

    my $size = scalar(@_);

    foreach my $a (@_)

    {

        $x = substr($a, 30, 8) + 0;

$y = substr($a, 38, 8) + 0;

        $z = substr($a, 46, 8) + 0;


        $sum_x = $sum_x + $x;

        $sum_y = $sum_y + $y;

        $sum_z = $sum_z + $z;

    }


    $refx= sprintf("\%4.3f",$sum_x/$size);
```

```perl
    $refy= sprintf("\%4.3f",$sum_y/$size);

    $refz= sprintf("\%4.3f",$sum_z/$size);


    push @r, $refx;

    push @r, $refy;

    push @r, $refz;


    return @r;

}



sub getShell

{

my $Cx = substr($_[0], 30, 8) + 0;

my $Cy = substr($_[0], 38, 8) + 0;

my $Cz = substr($_[0], 46, 8) + 0;


my $Dx = $_[1] + 0;

my $Dy = $_[2] + 0;

my $Dz = $_[3] + 0;


my $d = sqrt(($Cx-$Dx)*($Cx-$Dx)+($Cy-$Dy)

*($Cy-$Dy)+($Cz-$Dz)*($Cz-$Dz));


return $d;
```

}

PHYLIP output   Fitch-Margoliash algorithm with global rearrangement for EC1

proteins based on EC1 composome distance matrix. Output file is in a Newick tree

format.

((2oqe:0.00021,1w6g:0.00311):0.01349,((((((1odm:0.02883,

(1yi9:0.03054,(1yi9:0.01480,((1q0e:0.00437,(1f1g:0.00346,

((((1oal:0.00229,2aps:0.00602):0.00054,1z9n:0.00488):0.00298,

(1dmh:0.01757,((1tmx:0.04986,((1xso:0.00848,(1srd:0.02529,

(((2bur:0.01731,(3pcg:0.02019,((((1ehk:0.01264,2amu:0.02080):0.00369,

(2hvb:0.00064,1dqi:0.00181):0.02660):0.06543,((((1vlb:0.00773,

(2e1q:0.00966,(1v97:0.01066,(1rm6:0.04594,(1ffv:0.00321,

1n62:0.00539):0.00542):0.01665):0.01907):0.04793):0.02081,

(1vf5:0.04516,(3en1:0.09268,((1rfs:0.02324,((1rie:0.00128,

1ezv:0.00296):0.00809,2nvg:0.01147):0.00564):0.01596,2hmj:0.03169)

:0.01549):0.03659):0.05950):0.05165,

1zrt:0.01378):0.04558,(1cp2:0.04174,((1q90:0.00770,1kqf:0.00945):0.03582,

(1h0h:0.01977,((((1q16:0.01227,(1h1l:0.00268,(1m1n:0.00134,

1mio:0.00405):0.00111):0.00542):0.00677,1yq9:0.01082):0.00814,

(((1mjt:0.02250,((1d0c:0.00781,3e7g:0.00492):0.00942,1m9m:0.01671)

:0.01947):0.00909,

(2ort:0.04414,1m7v:0.01415):0.03277):0.01749,((((((1lfk:0.00283,

((3czh:0.00761,2f9q:0.00229):0.00214,((2fdv:0.00323,2pg5:0.00243):0.00132,

(3e6i:0.00465,(2q9f:0.00463,((3b4x:0.00191,1io7:0.00967):0.01511,

3eqm:0.00456):0.00954):0.00845):0.00328):0.00391):0.00845):0.01272,

1r9o:0.00559):0.00244,2nnj:0.00243):0.00651,(2wm5:0.00997,

((((1z8o:0.00324,1nr6:0.00100):0.00203,2hi4:0.00285):0.00046,

1po5:0.00347):0.00330,(1ued:0.01996,(((((((1itk:0.01196,

1h1l:0.00764):0.00298,1m1n:0.00665):0.00745,1ub2:0.00640):0.01252,

((((2zwu:0.00281,2wx2:0.00433):0.00492,((1cpt:0.00529,(3ejb:0.00840,

(((1gwu:0.00551,(((1b80:0.00059,1qpa:0.00241):0.00423,(1llp:0.00173,

((((3fmu:0.00088,1lyk:0.00370):0.00224,1sch:0.00180):0.00264,

1yyd:0.02146):0.00576,3e2o:0.00318):0.00325):0.00208):0.00437,

2ve3:0.01680):0.00581):0.01369,1apx:0.00817):0.00362,((1n40:0.00499,

(1q5d:0.00116,1q4g:0.00129):0.00305):0.00194,((2jjn:0.00580,

(2zbx:0.00681,((1gte:0.01594,(3bxi:0.02174,(1d2v:0.01452,

((3hhs:0.00109,((3hhs:0.00078,3hhs:0.00629):0.00204,(2gsm:0.00014,

(3hb3:0.00021,1ehk:0.00943):0.00065):0.09694):0.01984):0.01894,

(3hhs:0.00262,1gte:0.00277):0.00317):0.00764):0.03289):0.03722):0.01121,

2p3x:0.00428):0.01038):0.00314):0.00853,(3pgh:0.00570,2cca:0.00178):

0.00554):0.01148):0.00770):0.00299):0.00262):0.00786,

1bt3:0.01825):0.01160):0.00268,1pa2:0.00891):0.00251,1mio:0.04830)

:0.00195):0.00370,

(1mwv:0.00628,1qgj:0.00232):0.01214):0.00782,2uuq:0.00598):0.01184,

3dax:0.04193):0.01070):0.00707):0.01061):0.01142):0.09310,

(3c8y:0.00506,1hfe:0.00538):0.00576):0.00801,1bou:0.07164):0.06553)

:0.01706):0.01978,

2iv2:0.02483):0.03947):0.03968):0.01910):0.07368):0.12442):0.00604,

```
(2ji2:0.02027,1dfx:0.01154):0.03288):0.05378):0.01671):0.03810,

2og7:0.04773):0.01191,1s9a:0.02106):0.01580):0.02504):0.00984,

(1os7:0.01031,1ds1:0.01164):0.03683):0.01007):0.00146,2jlp:0.00945)

:0.00885):0.00279):0.00627,

1to4:0.00322):0.01735):0.01775):0.00652,1mfm:0.01003):0.00246)

:0.03998):0.05928):0.05591,

1oac:0.01157):0.01048,3hi7:0.00506):0.00327,1ksi:0.00179):0.00323,

1tu5:0.02118):0.00313,2c10:0.00378):0.00482,2et1:0.03147);
```

4.6 Supplementary Data

PDB	Catalytic site (x,y,z)			Composome			DSSP
PDB code	CmX	CmY	CmZ	Helix	Sheet	Loop	HBonds
1apx	17.69	0.13	1.53	0.69	0.07	0.24	-3.16
1bou	50.67	53.68	8.06	0.58	0.29	0.13	-3.01
1bt3	7.37	21.38	9.97	0.7	0.01	0.29	-3.55
1cp2	-0.96	-0.36	5.63	0.38	0.06	0.56	-2.73
1d0c	7.7	7.92	32.44	0.42	0.28	0.3	-3.38
1d2v	20.02	-12.45	-6.45	0.81	0.04	0.15	-3.67
1dfx	39.69	42.65	47.81	0.13	0.45	0.42	-3.07
1dmh	1.75	80.6	74.42	0.11	0.63	0.26	-2.69
1ds1	-1.82	4.15	4.14	0.03	0.65	0.31	-3.52
1ehk	78.24	17.44	6.04	0.92	0.01	0.07	-4.01
1f1g	-9.73	33.21	63.61	0.11	0.66	0.23	-3.28
1ffv	16.75	93.7	93.05	0.21	0.07	0.72	-2.52
1gte	60.58	66.52	82.51	0.81	0.01	0.18	-3.24
1h0h	15.13	-6.05	50.33	0.37	0.19	0.44	-2.8
1h1l	38.79	13.04	39.71	0.43	0.2	0.37	-3.14
1h1l	31.7	-4.26	35.09	0.63	0.12	0.25	-3.04
1io7	39.18	11.78	45.21	0.56	0.05	0.4	-3.22
1itk	6.56	-9.87	21.33	0.62	0.11	0.27	-3.46
1kqf	8.85	59.51	110.71	0.32	0.19	0.49	-2.74
1m1n	61.1	6.54	-6.37	0.43	0.2	0.37	-2.99
1m1n	61.05	-10.54	2.88	0.64	0.12	0.24	-2.95
1m7v	-0.35	29.05	8.88	0.49	0.31	0.2	-3.06
1m9m	15.82	14.28	57.29	0.43	0.29	0.27	-3.24
1mio	-2.97	-5.67	79.28	0.43	0.2	0.38	-2.83
1mio	13.56	-12.9	72.82	0.64	0.17	0.19	-2.89
1mjt	24.67	78.87	34.88	0.47	0.28	0.25	-3.11
1n62	98.08	188.92	171.74	0.2	0.08	0.71	-2.79
1nr6	-11.52	-20.02	-31.39	0.65	0.02	0.32	-3.42
1oac	67.82	68.62	-4.2	0.05	0.82	0.12	-3.4
1oal	27.32	33.74	14.15	0.12	0.64	0.24	-3.43
1odm	7.71	40.7	0.76	0.16	0.76	0.08	-3.93
1os7	29.98	79.06	23.71	0.03	0.67	0.3	-3.04
1q0e	12.34	0.05	74.47	0.1	0.68	0.22	-3.22
1q16	-138.83	-61.39	-22.52	0.42	0.22	0.36	-2.54
1q90	25.73	154.55	116.84	0.33	0.17	0.5	-2.9
1qgj	20.95	29.39	42.09	0.64	0.08	0.28	-3.17
1r9o	5.89	24.57	-1.06	0.63	0.02	0.35	-3.46
1rfs	4.65	19.02	32.75	0	0.3	0.7	-2.62
1rie	8.31	0.58	5.66	0	0.32	0.68	-2.74
1rm6	13.47	55.98	157.41	0.17	0.07	0.76	-2.4
1s9a	23.01	59.14	301.27	0.19	0.56	0.25	-2.94

1srd	0.79	4.13	-4.94	0.08	0.6	0.33	-2.6
1tmx	46.88	-0.16	11.04	0.18	0.64	0.18	-2.67
1tu5	27.84	-10.48	39.68	0.1	0.83	0.08	-3.44
1ub2	45.03	7.74	-81.38	0.64	0.1	0.25	-3.44
1ued	0.02	3.39	41.97	0.67	0	0.33	-3.67
1v97	149.19	45.87	32.32	0.23	0.05	0.71	-2.72
1vf5	57.67	-21.36	122.94	0.12	0.18	0.7	-1.99
1vlb	43.82	-22.64	26.95	0.29	0.06	0.65	-2.51
1w6g	-7.34	-2.17	27.75	0.06	0.86	0.08	-3.25
1xso	22.29	33.39	4	0.15	0.6	0.25	-2.99
1yi9	20.68	25.36	56.34	0.07	0.69	0.24	-3.34
1yq9	23.72	81.31	47.83	0.44	0.18	0.38	-3.09
1yyd	57.22	21.38	37.7	0.72	0.1	0.18	-3.26
1zrt	0.01	-4.14	-8.32	0.29	0.12	0.59	-2.97
2amu	20.06	23.14	21.73	0.01	0.46	0.53	-2.75
2bur	7.19	-13.33	42	0.15	0.53	0.33	-2.68
2e1q	33.35	19.56	151.86	0.26	0.05	0.69	-2.83
2et1	19.55	12.13	53.49	0.03	0.86	0.11	-3.83
2hmj	25.41	94.79	86.57	0.04	0.26	0.69	-2.89
2hvb	-27.06	7.04	13.94	0.01	0.49	0.49	-2.52
2iv2	74.37	26.5	34.32	0.39	0.24	0.37	-2.78
2ji2	16.95	45.85	40.61	0.12	0.43	0.44	-2.94
2jlp	17.46	1.73	21.74	0.14	0.62	0.23	-3.34
2nnj	50.84	22.12	-28.77	0.63	0.02	0.34	-3.08
2og7	-1.05	58.57	-11.82	0.02	0.59	0.38	-3.65
2ort	68.04	-9.38	55.71	0.46	0.36	0.18	-3.28
2p3x	14.35	-14.23	15.46	0.73	0.02	0.25	-3.32
2uuq	60.21	20.21	8.4	0.66	0.05	0.29	-3.41
2ve3	41.68	-3.21	6.6	0.72	0.08	0.2	-3.21
2wm5	22.15	-3.84	-1.89	0.63	0.05	0.32	-3.72
2wx2	5.93	20.86	13.38	0.68	0.03	0.29	-3.53
2zbx	9.77	8.67	12.42	0.72	0.02	0.26	-3.54
3bxi	10.3	2.43	27.04	0.78	0.08	0.14	-3.21
3c8y	19.41	17.61	19.64	0.51	0.16	0.33	-3.39
3dax	16.12	19.26	-40.76	0.6	0.17	0.23	-3.65
3en1	15.15	143.59	180.46	0.03	0.19	0.78	-2.5
3fmu	9.95	17.09	-0.2	0.7	0.1	0.2	-3.43
3hhs	-5.35	51.92	3.11	0.84	0	0.15	-3.65
3hi7	-21.26	-14.78	62.14	0.07	0.83	0.1	-3.48
3pcg	13.22	6.78	42.28	0.13	0.51	0.36	-2.72

Figure 4.8: Supplementary data: Dataset used for the composome analysis

4.7 Acknowledgements

4.8 References

1. Falkowski PG, Fenchel T, Delong EF. The microbial engines that drive Earth's biogeochemical cycles. Science. 2008 May 23;320 (5879):1034-9.

2. Braakman R, Smith E. The Emergence and Early Evolution of Biological Carbon-Fixation. Plos Comput Biol. 2012 Apr;8 (4).

3. Unden G, Bongaerts J. Alternative respiratory pathways of Escherichia coli: energetics and transcriptional regulation in response to electron acceptors. Biochim Biophys Acta. 1997 Jul 4;1320 (3):217-34.

4. Dupont CL, Yang S, Palenik B, Bourne PE. Modern proteomes contain putative imprints of ancient shifts in trace metal geochemistry. P Natl Acad Sci USA. 2006 Nov 21;103 (47):17822-7.

5. Holm RH, Kennepohl P, Solomon EI. Structural and functional aspects of metal sites in biology. Chem Rev. 1996 Nov;96 (7):2239-314.

6. Metz S, Thiel W. Theoretical studies on the reactivity of molybdenum enzymes. Coordination Chemistry Reviews. 2011;255 (9):1085-103.

7. Dey A, Jenney Jr FE, Adams MW, Babini E, Takahashi Y, Fukuyama K, et al. Solvent tuning of electrochemical potentials in the active sites of HiPIP versus ferredoxin. Science. 2007;318 (5855):1464-8.

8. Williams RJP. Overview of Biological Electron-Transfer. Adv Chem Ser. 1990 (226):3-23.

9. Schlesinger KJ, Johnson JA, Rockosi CM, Lee YS, Morrison HL, Schonrich R, et al. The Metallicity Distribution Functions of Segue G and K Dwarfs: Constraints for Disk Chemical Evolution and Formation. Astrophys J. 2012 Dec 20;761(2).

10. Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Res. 2012 Jan;40(Database issue):D71-5.

11. Bernstein FC, Koetzle TF, Williams GJ, Meyer EF, Jr., Brice MD, Rodgers JR, et al. The Protein Data Bank. A computer-based archival file for macromolecular structures. Eur J Biochem. 1977 Nov 1;80 (2):319-24.

12. Mauzerall DC. The photochemical origins of life and photoreaction of ferrous ion in the archaean oceans. Origins of Life and Evolution of Biospheres. 1990;20 (3):293-302.

13. Braterman PS, Cairns-Smith AG, Sloper RW. Photo-oxidation of hydrated Fe_2^+ significance for banded iron formations. 1983.

14. Wachtershauser G. Pyrite Formation, the 1st Energy-Source for Life - a Hypothesis. Syst Appl Microbiol. 1988 Aug;10 (3):207-10.

15. Martin W, Russell MJ. On the origins of cells: a hypothesis for the evolutionary transitions from abiotic geochemistry to chemoautotrophic prokaryotes, and from prokaryotes to nucleated cells. Philosophical Transactions of the Royal Society of London Series B: Biological Sciences. 2003;358 (1429):59-85.

16. Cody GD. Transition metal sulfides and the origins of metabolism. Annu Rev Earth Planet Sci. 2004; 32:569-99.

17. Oparin AI. Origin of life: Dover Publications; 2003.

18. Hazen RM, Sverjensky DA. Mineral Surfaces, Geochemical Complexities, and the Origins of Life. Csh Perspect Biol. 2010 May;2 (5).

19. Wachtershauser G. Before Enzymes and Templates - Theory of Surface Metabolism. Microbiol Rev. 1988 Dec;52 (4):452-84.

20. Holland HD. The chemical evolution of the atmosphere and oceans. Princeton, N.J.: Princeton University Press; 1984.

21. Sadekar S, Raymond J, Blankenship RE. Conservation of distantly related membrane proteins: photosynthetic reaction centers share a common structural core. Molecular biology and evolution. 2006;23 (11):2001-7.

22. Eck RV, Dayhoff MO. Evolution of the structure of ferredoxin based on living relics of primitive amino acid sequences. Science. 1966;152:363-6.

23. Lorimer GH. The carboxylation and oxygenation of ribulose 1, 5-bisphosphate: the primary events in photosynthesis and photorespiration. Annual Review of Plant Physiology. 1981;32 (1):349-82.

24. Kim JD, Rodriguez-Granillo A, Case DA, Nanda V, Falkowski PG. Energetic selection of topology in ferredoxins. PLoS Comput Biol. 2012;8 (4):e1002463.

25. Harel A, Falkowski P, Bromberg Y. TrAnsFuSE refines the search for protein function: oxidoreductases. Integrative biology : quantitative biosciences from nano to macro. 2012 Jul;4 (7):765-77.

26. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983 Dec;22 (12):2577-637.

27. Sippl MJ, Wiederstein M. Detection of spatial correlations in protein structures and molecular complexes. Structure. 2012 Apr 4;20 (4):718-28.

28. Tokuriki N, Tawfik DS. Protein dynamism and evolvability. Science. 2009;324 (5924):203-7.

29. Plaxco KW, Simons KT, Baker D. Contact order, transition state placement and the refolding rates of single domain proteins. Journal of molecular biology. 1998;277 (4):985-94.

30. Felsenstein J. Phylogenies and the Comparative Method. Am Nat. 1985;125 (1):1-15.

31. Schoepp-Cothenet B, van Lis R, Philippot P, Magalon A, Russell MJ, Nitschke W. The ineluctable requirement for the trans-iron elements molybdenum and/or tungsten in the origin of life. Scientific Reports. 2012; 2.

Curriculum Vitae

J. Dongun Kim

Degrees Conferred

2013   PhD in Biological Chemistry, Rutgers University, The State University of New Jersey, USA

2009   BA in Chemistry (minor Mathematics), Rutgers University, The State University of New Jersey, USA

Professional Experience

2010-2013   Graduate Research Assistant, Rutgers University

2009-2010   NSF IGERT (Nanotechnology for Clean Energy) Fellow

Publications

8. Kim, J.D.; Senn, S.; Harel, A.; Jelen, B.I.; Falkowski, P.G. "Discovering the electronic circuit diagram of life: Structural relationships among transition metal binding sites in oxidoreductases", ***Philosophical Transactions of The Royal Society B***, 368, 1622, 20120257 (2013)

7. Mass, T.; Drake, J.L.; Haramaty, L.; **Kim, J.D.**; Zelzion, E.; Bhattacharya, D.; Falkowski, P.G. "Cloning and characterization of four novel coral acid-rich proteins that precipitate carbonates in vitro", ***Current Biology***, 23, 12, 1126-1131 (2013)

6. Kim, J.D.; Yee, N.; Nanda, V.; Falkowski, P.G. "Anoxic photochemical oxidation of siderite generates molecular hydrogen and iron oxides", ***Proceedings of the National Academy of Sciences***, 110, 25, 10073-10077 (2013) ***Editor's Choice**, ***Science***, 340, 6140, 1501(2013)

5. Kim, J.D.; Rodriguez-Granillo, A.; Case, D.A.; Nanda, V.; Falkowski, P.G. "Energetic selection of topology in ferredoxins", ***PLoS Computational Biology*** 8(4) (2012)

4. Kim, J.D. "Hot skies over a cold landscape", ***Science***, 333, 41 (2011)

3. Gorbunov, M.Y.; Kuzminov F.I.; Fadeev V.V.; **Kim, J.D.**; Falkowski P.G. "A kinetic model of non-photochemical quenching in cyanobacteria", ***Biochimica Biophysica Acta Bioenergetics***, 1807, 12, 15911599 (2011)

2. Solanki, A.; **Kim, J.D.**; Lee, K.B. "Nanotechnology for Regenerative Medicine: nanomaterials for stem cell imaging ", ***Nanomedicine***, 3, 567-578 (2008)

1. Lee, K.B.; Solanki, A.; **Kim, J.D.**; Jung J.,"Nanomedicine: dynamic integration of nanotechnology with biomedical science", Zhang, M., Editors, World Scientific (2008)

Presentations

7. *Evolution of biological redox reactions*, **Institute for Basic Science, Center for RNA Research**, Seoul, South Korea, Feb 2013 (Talk)

6. *Evolution of proteins studied by secondary structure composition*, **3DSIG: Structural Bioinformatics and Computational Biophysics**, Long Beach, CA, USA, July 2012 (Poster)

5. *Chemists perspective on coevolution of life and geochemical cycles on our planet*, **Korea Polar Research Institute**, Incheon, South Korea, May 2012 (Invited Talk)

4. *Energetic selection of topology and chirality in feredoxins*, 5th Annual Advances in Biomolecular Engineering Symposium, **New York Academy of Science**, New York, NY, USA, Apr 2012 (Poster)

3. *Evolution of protein folds based on secondary structure composition*, **Gordon Research Conference, Protons & Membrane Reactions**, Ventura, CA, USA, Feb 2012 (Poster)

2. *The origin of topological chirality in iron-sulfur proteins*, **Carnegie Institution of Washington Geophysical Laboratory**, Washington DC, USA, Apr 2011 (Informal Talk)

1. *The origin of topological chirality in iron-sulfur proteins*, **Gordon Research Conference, Metals in Biology**, Ventura, CA, USA, Jan 2011 (Poster)