A Bioinformatics Approach to Identifying Radical SAM (*S*-Adenosyl-L-Methionine) Enzymes

Elisa Marie Gagliano

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science in Life Sciences
in
Biochemistry

Anne M. Brown, Committee Chair
Kylie D. Allen
Justin A. Lemkul

May 14, 2020
Blacksburg, Virginia

Keywords: Bioinformatics, Radical Biochemistry, Enzymology

A Bioinformatics Approach to Identifying Radical SAM (*S*-Adenosyl-L-Methionine) Enzymes

Elisa Gagliano

ACADEMIC ABSTRACT

Radical SAM enzymes are ancient, essential enzymes. They perform radical chemical reactions in virtually all living organisms and are involved in producing antibiotics, generating greenhouse gases, human health, and likely many other essential roles that have yet to be established. A wide variety of reactions have been characterized from this group of enzymes, including hydrogen abstractions, the transferring of methylthio groups, complex cyclization and rearrangement reactions, and others. However, many radical SAM enzymes have yet to be identified or characterized. There have been great leaps forward in the amount of enzyme sequences that are available in public databases, but experiments to investigate what chemical reactions the enzymes perform take a great deal of time. In our work, we utilize Hidden Markov Models to identify possible radical SAM enzymes and predict their possible functions through BLAST alignments and homology modelling. We also explore their distribution across the tree of life and determine how it is correlated with organism oxygen tolerances, because the core iron-sulfur cluster is oxygen sensitive. Trends in the abundances of radical SAM enzymes depending on oxygen tolerances were more apparent in prokaryotes than in eukaryotes. Although eukaryotes tend to have fewer radical SAM enzymes than prokaryotes, we were able to analyze uncharacterized radical SAM enzymes from both an aerobic eukaryote (*Entamoeba histolytica*) and a eukaryote capable of oxygenic photosynthesis (*Gossypium barbadense*), and predict the reactions they catalyze. This work sets the stage for the functional characterization of these essential yet elusive enzymes in future laboratory experiments.

A Bioinformatics Approach to Identifying
Radical SAM (*S*-Adenosyl-L-Methionine) Enzymes

Elisa Gagliano

GENERAL AUDIENCE ABSTRACT

Radical SAM enzymes are ancient, essential enzymes that perform chemical reactions in virtually all living organisms. We do know that they are involved in producing antibiotics, human health, and generating greenhouse gases. We also know that there are many radical SAM enzymes whose functions remain a mystery. There have been great leaps forward in the amount of enzyme sequences that are available in public databases, but experiments to investigate what chemical reactions enzymes perform take a great deal of time. The experiments are especially difficult for radical SAM enzymes because the oxygen we breathe can break the enzymes down in a laboratory. In our work, we utilize computational techniques to identify possible radical SAM enzymes and predict what reactions they might catalyze. Because these enzymes are vulnerable to oxygen in laboratory environments, we also explore whether organisms that breathe oxygen have fewer of these enzymes than organisms that perform anaerobic respiration instead. We found that does seem to be the case in microbes like bacteria and archaea, but the results were not as consistent for eukaryotes. We then chose radical SAM enzymes we had identified from both an aerobic eukaryote (*Entamoeba histolytica*) and a eukaryote capable of producing oxygen (*Gossypium barbadense*), and predicted the reactions they catalyze. This work sets the stage for the functional characterization of these essential yet elusive enzymes in future laboratory experiments.

Acknowledgements

This project is a collaboration between the laboratories of Anne Brown and Kylie Allen. I would like to acknowledge the preliminary work that Amanda Sharp and Paulene Sapao contributed to this project through the Bevan and Brown Lab.

My graduate work at has been supported by the Virginia Tech Biochemistry Department and the Virginia Tech George Washington Carver Assistantship.

Some of the illustrations and videos were made in Chimera [1] and PyMOL, and in Adobe After Effects through the Virginia Tech Library Media Design Studios.

I would also like to acknowledge Robert Settlage and Frank Aylward for their perspectives on bioinformatics. In this project I was also able to apply skills I practiced in courses taught by David Haak, Song Li, Roderick Jensen, and Clement Vinauger, as well as skills I learned from David Haak and Aureliano Bombarely as an undergraduate intern in the Virginia Tech Multicultural Academic Opportunities Program. Undergraduate courses at the University of New Mexico, and working with Mariel Campbell, Jennifer Rudgers, and Michelle Facette also provided valuable experience.

I like to acknowledge the online coding community and its helpful publicly available tutorials and discussion boards, and the workshops offered by the Virginia Tech library and Technology-enhanced Learning and Online Strategies.

I would like to acknowledge Robert White for my initial introduction to radical SAM enzymes, and Anne Brown, Kylie Allen, and Justin Lemkul for serving on my committee, and acknowledge all other people who were a part of both planned and unexpected events that led to this document, and support from unexpected places.

Table of Contents

List of Figures

QMEAN Z-scores are very low. The other individual scores are reasonable and do not pass the threshold of -4.

Figure 9. Homology modelling structural overlays. (A) Structural overlay of homology model of A0A5K1U8H1_ENTHI from *Entamoeba histolytica* (teal), and RimO (PDB 4JC0) template with cofactors (magenta). (B) 4JC0 template and iron-sulfur cluster. (C) A0A5K1U8H1_ENTHI homology model overlaid with iron-sulfur cluster from 4JC0 template. Distances between cysteine Sγ atoms and the iron atoms are measured in Å in panels (B) and (C).

Figure 10. Tree of MAFFT alignment between selected standard radical SAM sequences, and *Arabidopsis thaliana* (brown) and *Gossypium barbadense* (blue) sequences that matched radical SAM HMM profiles. The relationship between the circled proteins are explored in Table 7. The A0A5J5PQ sequence from *G. barbadense* did not appear to be a radical SAM enzyme upon further inspection.

Supplemental Figure 1. Phylogenetic tree of the 100 randomly-selected eukaryotic organisms. Counts of radical SAM HMM hits are blue, percentages of radical SAM HMM hits in each proteome is in red.

Supplemental Figure 2. Homology modelling structural overlays. (A) Structural overlay of homology model of A0A5K1U8H1_ENTHI from *Entamoeba histolytica* (teal), and RimO (PDB ID: 4JC0) template with cofactors (magenta). (B) 4JC0 template and iron-sulfur cluster. (C) A0A5K1U8H1_ENTHI homology model overlaid with iron-sulfur cluster from 4JC0 template. Distances between cysteine Sγ atoms of the "nest" and the iron atoms are measured in Å in panels (B) and (C).

List of Tables

Table 4. Motif counts for sets of standard Radical SAMs and ferredoxins. The length cutoff refers to excluding sequences with the CX_4C_2X motif that are shorter than 400 amino acids.

Table 5. Top BLAST hit between each of the four *Entamoeba histolytica* (UP000078387) sequences found from HMM profile searches, and the radical SAM standard sequences.

Table 6. The top blast hits between the four *Entamoeba histolytica* (UP000078387) proteins found from HMM search, and radical SAM enzymes with solved crystal structures.

Table 7. BLAST between *Gossypium barbadense* (UP000327439) uncharacterized protein A0A5J5PDX9 and nearby neighbors circled in Figure 10.

Additional supplemental tables included in '**Supplemental_Tables.xlsx**' attachment. Those tables include:

Supplemental Table 1. The top BLAST hits between each of the 37 *Gossypium barbadense* (UP000327439) proteins found from HMM search, and the radical SAM standard sequences.

Supplemental Table 2. The top BLAST hits between the 37 *Gossypium barbadense* (UP000327439) proteins found from HMM search, and radical SAM enzymes with solved crystal structures.

Supplemental Table 3. Software dependencies

Supplemental Table 4. The UniProtKB accession numbers for proteins included as Radical SAM and Ferredoxin standards

Supplemental Table 5. Sets of proteomes included in the analyses

Supplemental Table 6. Counts of proteome motif matches

Supplemental Table 7. Sequences identified by HMM searches. Hits to Pfam models include E-values.

Supplemental Table 8. Counts of Radical SAM matches per proteome, in the different datasets

List of Videos

Supplemental file "**Radical_SAM_3D.mov**" is also attached separately.
Supplemental file "**Radical_SAM_2D.mov**" is also attached separately.
https://vimeo.com/423775537

1 Introduction

1.1 General Radical SAM (*S*-Adenosyl-L-Methionine) Characteristics

Radical SAM (*S*-Adenosyl-L-Methionine) enzymes comprise a large superfamily of enzymes found across the tree of life. These enzymes use SAM and a four iron-four sulfur (4Fe-4S) cluster to catalyze a multitude of single-electron (radical) transformations on various substrates. The radical SAM superfamily was established through bioinformatic means in 2001 [2]. That study used the sequences of enzymes that had been characterized as far back as 1970, beginning with lysine 2-3-aminomutase (LAM) [3,4]. Discoveries made in the intervening decades revealed that in these enzymes, SAM is used in radical reactions. This was unexpected at the time, as only adenosylcobalamin was known to perform these types of radical reactions [5], and because SAM was widely known to be used by enzymes as only a methylating agent [2]. Radical SAM enzymes were likely the first radical enzymes to have evolved, with adenosylcobalamin, a much more complex cofactor, evolving later [6]. Since the radical SAM superfamily was established in 2001 [7], studies have found that radical SAMs are the most abundant radical bio-catalysts in nature [8], and the list of the reactions they are known to perform grows [6] and exhibits the potential for future growth [9].

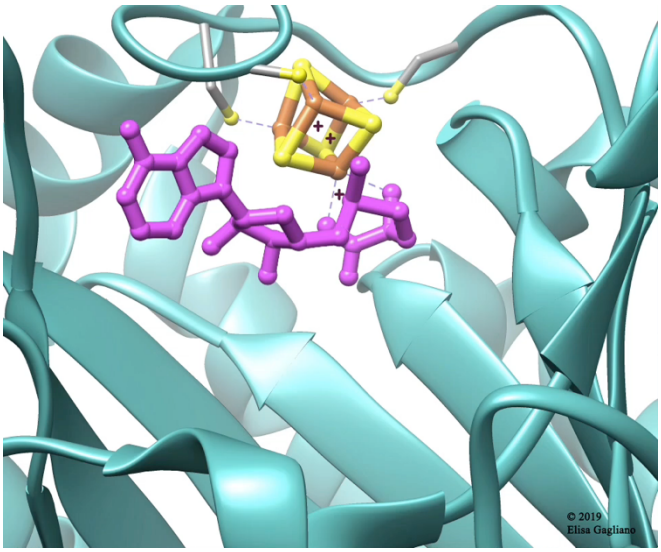


Figure 1. Depiction of GTP 3',8-cyclase MoaA (PDB ID: 1TV8) [13], a radical SAM enzyme. The iron-sulfur cluster is bound to three cysteine residues, and coordinates the SAM molecule. The MoaA enzyme is depicted as a cyan ribbon diagram, the iron-sulfur cluster is shown in yellow and orange balls and sticks. The SAM molecule is shown in magenta balls and sticks.

Although the radical SAM superfamily is large and its enzymes catalyze a wide range of reactions, all radical SAMs have a common first step in their reaction mechanisms. (Figure 1, Figure 2, and Radical_SAM_3D.mov) All radical SAM enzymes coordinate a $[4Fe\text{-}4S]^{+2}$ cluster that accepts an electron, then transfers it to SAM. Subsequent homolytic cleavage of SAM produces a radical intermediate, which is then used to transform the various substrates into various products, usually beginning by abstracting a substrate hydrogen [10,11]. The most common radical intermediate is the 5'-deoxyadenosyl radical, although one exception that cleaves the S-C(γ) bond to generate a 5'-methylthioadenosyl radical has been characterized [6,10]. SAM is usually consumed in the reaction, although some enzymes, such as lysine 2,3-aminomutase and spore photoproduct lyase can recycle the SAM molecule [3,6,12,13]. The first stages of the mechanism are largely conserved, but members of the radical SAM superfamily are capable of harnessing them to transform many different substrates. Some of the substrates include amino acids and other small molecules; in other cases, tRNA and proteins are modified [6]. Other notable characteristics of all radical SAMs is the presence of a partial to full TIM barrel structure, and the presence of a

CX_3CX_2C motif and rarer variations on it (Table 1), which is required to bind the $[4Fe\text{-}4S]^{+2}$ iron-sulfur cluster [6].

Table 1. Radical SAM motifs* found in peer-reviewed literature. Since the establishment of the radical SAM superfamily in 2001, other radical SAM motifs have been discovered. These non-canonical motifs either have more amino acids between the first two cysteines than the canonical motif does, or have diverging patterns.

Motif	Literature Source
CX_3CX_2C (canonical)	(Sofia *et al.*, 2001)[1], (Layer et al., 2004)[11]
CX_4CX_2C	(Challand et al., 2011)[7], (Selvadurai et al., 2014)[12], (Berteau and Benjdia,2017)[13]
CX_5CX_2C	(McGlynn *et al.*, 2010)[14], (Challand et al., 2011)[7], (Berteau and Benjdia, 2017)[13]
CX_7CX_2C	(Parent et al., 2016)[15], (Berteau and Benjdia, 2017)[13]
CX_8CX_2C	(Thweat et al., 2016)[16]
CX_9CX_2C	(Greenwood *et al.*, 2009)[17], (Berteau and Benjdia, 2017)[13]
$CX_{14}CX_2C$	(Dowling et al., 2014)[18], (Berteau and Benjdia, 2017)[13]
$CX_2CX_{21}CX_5C$	(Kamat *et al.*, 2013)[19], (Berteau and Benjdia, 2017)[13]
CX_2CX_4C	(Chatterjee et al., 2008)[20], (Fenwick *et al.*, 2014)[21], (Challand et al., 2011)[7], (Berteau and Benjdia, 2017)[13]

*Information on motifs that bind auxiliary clusters is not included in this table.

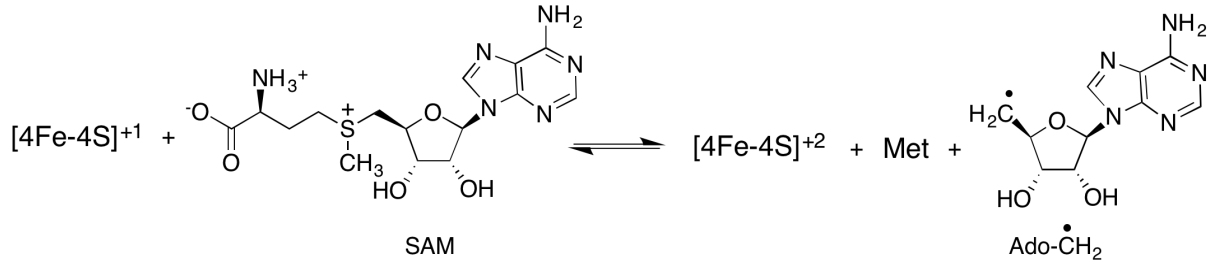


Figure 2. The common first step in a radical SAM enzyme reaction. SAM is cleaved at the 5' carbon by a radical electron donated from the iron-sulfur cluster, forming a 5' deoxyadenosyl radical intermediate. Figure used with permission from Kylie Allen.

Radical SAM enzymes have diversified to catalyze more specific groups of reactions. Radical SAMs are able to break C-H alkyl bonds that are generally considered unreactive [3,6,10]. Glycyl radical enzyme activating enzymes, such as pyruvate formate lyase activating enzyme, simply abstract a hydrogen from its substrate [6,14]. Some radical SAM enzymes, like biotin synthase, insert sulfur into substrate C-H bonds [6,15]. Others perform rearrangement reactions, like lysine 2,3-aminomutase [4,6], or the more complex MoaA-MoaC cyclization reaction [6,16]. Other radical SAM-catalyzed reactions include methylation [17], methylthiolation [18], dehydrogenation [19], forming and breaking carbon bonds [6], synthesizing modified tetrapyrroles [20], and synthesizing complex metal clusters [6,21].

Not only do radical SAM enzymes catalyze a diverse range of reactions, but they perform these reactions in a wide range of biological systems. They have been characterized in a wide range of bacteria, archaea, and eukaryotes [6], including humans. To date, eight radical SAM enzymes have been characterized in humans, and have been implicated in diseases such as molybdenum cofactor deficiency, congenital heart disease, and others [16,22]. One of their frequent functions across all life is their role in cofactor biosynthesis [6,23]. In Archaea, it is notable that multiple radical SAMs are required for the process of methanogenesis [24]. Methanogenesis has important implications for the environment [25,26], as well as producing a renewable source or energy [27]. In Bacteria, radical SAMs play notable roles in post-translational modifications and producing antibiotics [6]. The InterPro database system of automatic annotations has predicted radical SAM enzymes in viruses, such as accession Y301_ATV in Acidianus two-tailed virus, and accession MIMI_R756 in acanthamoeba polyphaga mimivirus, although these enzymes have not been studied in-depth at the time of this writing.

There are a multitude of reasons as to why more radical SAM enzymes need to be identified and characterized. As described above, they are involved in many essential roles in a wide range of lifeforms, with implications ranging from human health to the environment. In addition to their importance in biological systems, radical SAMs have the potential to influence industry and environmental sustainability. For example, the process of methanogenesis, which requires multiple radical SAM enzymes, can be harnessed to produce renewable fuels. More generally, it is the ability of radical SAM enzymes to activate otherwise unreactive bonds that is useful to industrial chemical production [28–32]. Collectively, the diversity and utility of radical SAM enzymes and their functions are of interest to a wide variety of fields and warrant further exploration and rapid classification for future work.

1.2 Ancient Times and Evolution in the Context of Radical SAM Enzymes

Proteins with iron-sulfur clusters are generally important parts of living systems and have an interesting evolutionary history. Iron-sulfur minerals may have played a central role in the beginning of life on earth; the iron-sulfur world hypothesis that was developed in the late 20th century introduced this idea [33]. Some studies oppose the idea [34], while others indicate that similar reactions could have been catalyzed by other minerals [35], but many studies are consistent with the idea that iron and sulfur were important catalysts that ushered in life [36–43]. Iron-sulfur clusters may have fostered life by forming spontaneously on organic compounds [44], and eventually evolved into the ancestors of ferredoxins, which then diversified into different families of proteins, including radical SAMs [43,44]. These minerals provided a surface for substances to bind, and a source of reducing power [23]. The iron-sulfur world hypothesis provides a reasonable explanation for why iron-sulfur clusters of various stoichiometries [44] are the most abundant type of cofactor [23].

The gradual increase of atmospheric oxygen posed problems for iron-sulfur proteins. Oxygen and other sources of oxidative stress like hydrogen peroxide tend to react with, and degrade, iron-sulfur clusters [43,45]. A more indirect effect is that an increase in oxygen levels made it more difficult for organisms to utilize iron from their environment [43]. The increase in redox potential caused soluble ferrous iron (Fe^{2+}) to form ferrous (Fe^{3+}) precipitates [46]. Organisms that relied on iron-sulfur cluster proteins thus had to adapt, either by decreasing the number of genes encoding these proteins and evolving alternative pathways, adjusting the structure

of the proteins so that they were less susceptible to oxygen damage, and/or developing new uses for reactive proteins [23,43,46–49].

Since the two billion years since oxygen began increasing in the atmosphere concentrations [46], prokaryotic organisms have adapted to environmental niches and developed different oxygen tolerances, ranging from obligate anaerobic lifestyles [50–52] to obligate aerobic lifestyles that rely on ambient environmental levels of oxygen (around 21%) [53]. Passive diffusion of oxygen across plasma membranes [54] leaves little flexibility for these types of organisms, while others have facultative tolerances to oxygen and anoxic environments [54]. These different categories of organisms could ostensibly have different amounts of radical SAM enzymes based on their oxygen tolerances, or their radical SAM enzymes could have structural differences.

Eukayotes are generally considered to be aerobic organisms. One of the usual characteristics is their reliance on mitochondria that specialize in aerobic respiration [55], which would seem to imply that they would have fewer oxygen-sensitive radical SAM enzymes and iron-sulfur cluster proteins in general. However, a small polyphyletic group of both multicellular and unicellular eukaryotes have mitosomes or hydrogenosomes instead of mitochondria and are considered anaerobic or microaerophilic [56]. Some examples of these organisms are *Arenicola marina, Fusarium oxysporum, Neocallimastic frontalis, Enephalitozoon cuniculi*, *Entamoeba histolytica,* and others [56]. Mitochondria may also play a central role in assembling iron-sulfur clusters in eukaryotes [57], thus potentially modulating the amounts of radical SAM enzymes in these organisms. Because many of the well-characterized radical SAM enzymes are from prokaryotic organisms, and few of those that have been identified in eukaryotes are specific to those anaerobic organisms and their specific lifestyle, it is likely that these organisms would yield an interesting set of novel radical SAM enzymes. It is also interesting to note that some studies have shown that there is evidence that some anaerobic eukaryotes obtained genes from prokaryotes, and that these helped support their anaerobic lifestyle [58,59].

Plants can be considered the opposite side of the spectrum of oxygen levels. They and other eukaryotes that perform photosynthesis produce oxygen by splitting water with photosystem II [60–62]. Plants generate oxygen as they photosynthesize, and that can lead to a build-up of oxygen around leaf microenvironments above ambient levels [63,64]. The increased oxygen levels in plant cells could potentially affect the levels of iron-sulfur cluster proteins like radical SAM enzymes.

Although many studies have explored the roles of radical SAM enzymes in bacteria, relatively few have been explored in eukaryotes. It is possible that radical SAM enzymes with novel functions can be characterized in this domain and that anaerobic eukaryotes might have unique sets of radical SAM enzymes. It is also possible that those with higher oxygen levels might have developed unique pathways to manage any iron-sulfur cluster proteins that are in their proteomes. Exploring the proteomes and radical SAM enzymes of eukaryotes that live at different levels of oxygen concentrations could yield novel insights in the study of radical SAM enzymes.

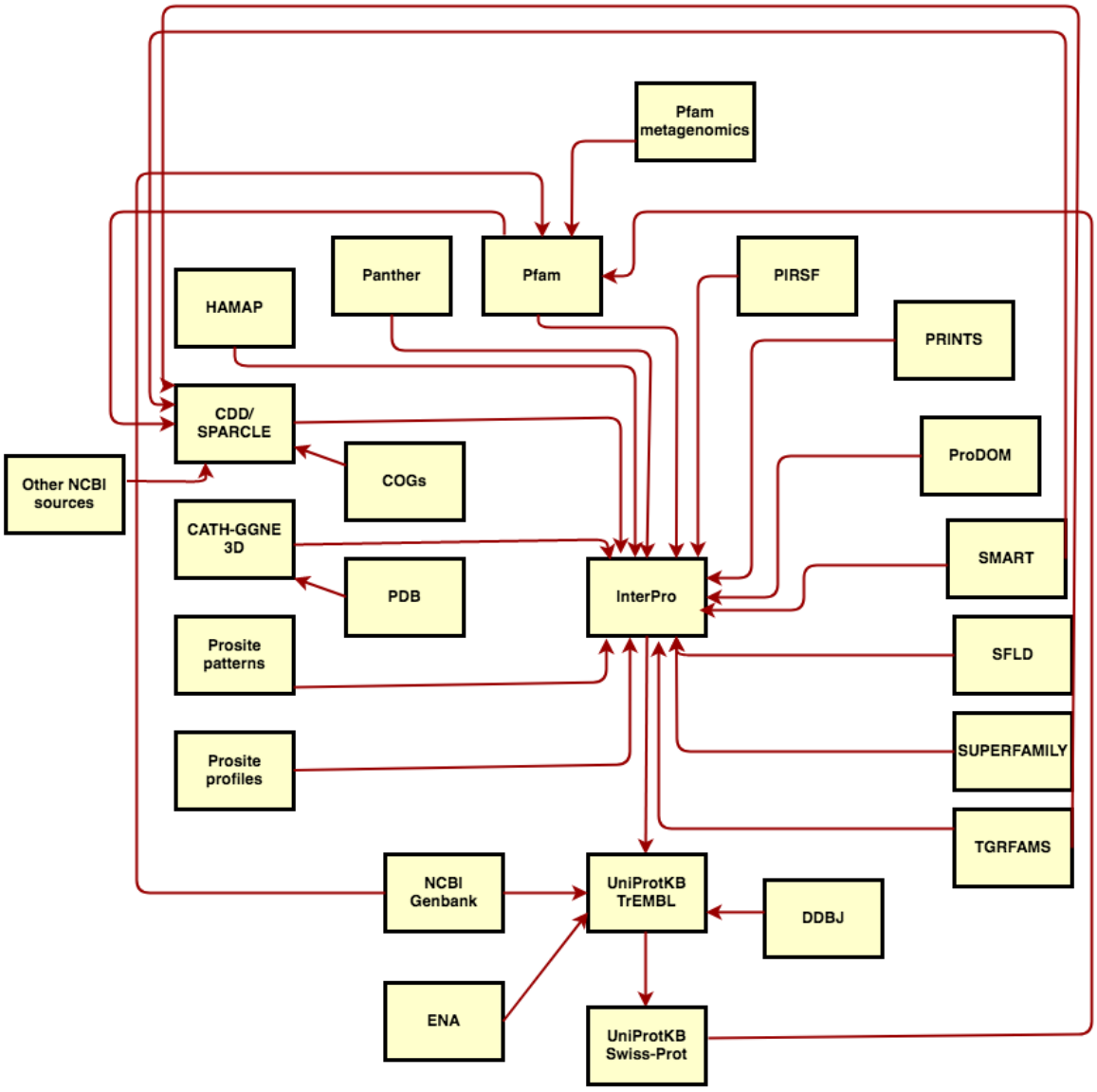


Figure 3. Web diagram representing the sources of sequences and signatures of publicly available biological data. Pfam is one of the sources of protein signatures for InterPro UniProt uses these predictions and its own sets of rules to automatically annotate TrEMBLE sequences. When these sequences are manually curated they are added to the Swiss-Prot database.

1.3 Bioinformatics and Biological Databases

The field of bioinformatics has substantially expanded since the radical SAM superfamily was first established in 2001 [7]. Additionally, public databases and the amount of data deposited, specifically biological sequences, has greatly expanded in the past decade [65,66]. While certain sequence data is housed in singular or specific databases, there is interconnectedness between major databases that curate biological data. Figure 3 represents the web of databases that lead to

UniProtKB, the database that is considered to be the main source for annotated protein sequences. UniProtKB draws from the sequences and annotations of a web of dozens of other databases and sets of rules [66,67]. UniProtKB is divided into Swiss-Prot, the bank of manually-curated sequences, and the more vast collection of sequences that have been only automatically annotated, UniProtKB/TrEMBL [67]. Both access sequences from NCBI Genbank, ENA, and DDBJ [66]. TrEMBLE sequences are automatically annotated – classified – by InterPro, a database derived from a wide range of others [68], as well as a filter of UniProt rules. A range of different techniques are used to curate and predict protein characteristics in these databases. One example that InterPro draws from is CATH-GENE3D, which classifies protein families by the similarity of sequences to those of proteins with 3D structures stored in the PDB [69]. Another example is Pfam, a database that curates protein families based on alignments and Hidden Markov Models from representative sequences [70].

Hidden Markov Models (HMMs) summarize protein sequence alignments and then their profiles can be used to identify potential homologues. HMM methodology is similar to searching for proteins based on the presence of a motif, but it is more flexible and statistically rigorous. HMMs have been applied to the field of bioinformatics since 1989, and had been applied to other fields before then [71–73]. A benefit of using HMMs at the superfamily level is that HMMs are capable of ignoring noise in sequences and determining conserved patterns, such as the conserved cysteine residues in the radical SAM superfamily.

HMMs are based on three concepts: multinomial sequence models, Markov chains, and a third component unique to this type of model. Multinomial sequence models determine the probability of each symbol – type of amino acid, in this case – being present at given position in the sequence, and it depends only on the given position. In Markov chains, the probabilities of a symbol being present in a given position in a sequence also depend on the symbols that precede it. HMMs add another layer of complexity where algorithms detect "hidden states." A simple example of hidden states that HMMs could be used to detect are whether a portion of the sequence is within a membrane or outside of it based on the frequencies of hydrophobic and hydrophilic amino acids in the sequence, although these hidden states in protein sequences can be more abstract [71]. These models and other computational methods provide a solid first step towards characterizing protein sequences of unknown function.

Automatic annotations from databases are very useful, but they do have their limitations. Although these databases build off of each other's information (Figure 3), they do not fully align with each other. Furthermore, some studies have shown that their automatic annotations are not always correct [74–79]. Although the databases are constantly improving, the data utilized and interpreted critically. It is also important to note that automatic annotation improvements have had a difficult time keeping pace with the great increase in available sequencing data [80] and automatic annotations based off of low-quality sequencing data [81]. Such problems are most apparent in unculturable microbes [81,82]. Some studies show that automatic annotations become inaccurate at the family scale [83]. Traditional laboratory tests, which can be expensive and time-consuming [84], are necessary to confirm protein function and characterization, and have an added level of difficulty for anaerobic organisms and oxygen sensitive proteins, like radical SAMs. There are also some proteins which have functions that are completely unknown and have no identifiable signatures. Therefore, identifying approaches to utilize computational methods in a way that optimizes the selection of proteins to assay and characterize.

Computational techniques and automatic annotations are useful in predicting the functions of many proteins on a coarse scale. However, they are based on a core of experimental laboratory

evidence, and such techniques are necessary for confirming fine-scale protein functions, such as at the family level. In this work, we seek to apply HMMs to identify putative members of the radical SAM superfamily for further laboratory characterization. HMMs are useful in identifying proteins that are very distantly related. Additionally, in order to explore and further understand the scope and utility of radical SAM enzymes, we probe the abundance of these enzymes across different phylogenetic groups from all domains of life, and with sub-groups containing varying levels of oxygen tolerances, to better understand the biology of these enzymes.

2 A Bioinformatics Approach to Identifying Radical SAM Enzymes

2.1 Abstract

Radical SAM (*S*-Adenosyl-L-Methionine) enzymes are ancient, essential enzymes. They perform radical chemical reactions in virtually all living organisms and are involved in producing antibiotics, generating greenhouse gases, function in human health, and likely many other essential roles that have yet to be established. A wide variety of reactions have been characterized from this group of enzymes, including hydrogen abstractions, transferring of various functional groups, complex cyclization and rearrangement reactions, among others. However, many radical SAM enzymes may yet be unidentified or uncharacterized. There have been great leaps forward in the amount of enzyme sequences that are available in public databases, but experiments to investigate what chemical reactions the enzymes perform take a great deal of time. In our work, we utilize Hidden Markov Models (HMMs) to identify putative radical SAM enzymes and predict their possible functions through BLAST alignments and homology modelling. We also explore their distribution across the tree of life and determine how it is correlated with organism oxygen tolerances, because the core iron-sulfur cluster is oxygen sensitive. Trends in the abundances of radical SAM enzymes depending on oxygen tolerances were more apparent in prokaryotes than in eukaryotes. Although eukaryotes had fewer radical SAM enzymes than prokaryotes, we were able to propose uncharacterized radical SAM enzymes from both an aerobic eukaryote (*Entamoeba histolytica*) and a eukaryote capable of oxygenic photosynthesis (*Gossypium barbadense*), and predict the reactions they catalyze. This work sets the stage for the functional characterization of these essential, yet elusive, enzymes in future laboratory experiments.

2.2 Introduction

The radical SAM (*S*-adenosyl-L-methionine) superfamily is comprised of a large number of enzymes that catalyze a diverse set of reactions that are essential to life. Those Radical SAM enzymes that have been characterized are known to catalyze reactions between carbon-hydrogen bonds that are difficult to achieve without the use of radical chemistry, including hydrogen abstraction, sulfur insertion, cyclization, and more [6]. It is their ability to complete these reactions that make them promising candidates for the industrial production of chemicals [28–32]. Although their range of chemical reaction involvement is diverse, they all share the earliest steps of these reactions, in which a $[4Fe\text{-}4S]^{+2}$ cluster attracts a single electron, which then directs it to a SAM molecule. This electron transfer causes SAM to break into a 5'-deoxyadenosyl radical, which subsequently reacts with the various substrates that this superfamily transforms [6]. To ligate the iron-sulfur cluster, these enzymes are dependent on a hallmark cysteine motif (CX_3CX_2C) or similar variations of it [6]. In 2001, the presence of these motifs was used to establish the radical SAM superfamily [7].

The presence and dependence on the iron-sulfur cluster leaves radical SAM enzymes vulnerable to oxygen, which has evolutionary and physiological implications on the abundance of these enzymes. It has been hypothesized that radical SAM enzymes (and their very distant relatives, ferredoxins), were among the first enzymes in the history of life, which began in a period before oxygen was abundant in the atmosphere [36–43,85]. However, once oxygenic photosynthesis evolved, the oxygen concentration of the environment began to rise, and reducing potential plummeted [43,46]. This shift made it more difficult for organisms to use iron from the

environment, but it was also a direct liability for their iron-sulfur proteins [43]. For example, oxygen and reactive oxygen species extract an iron atom from radical SAM enzymes' $[4Fe\text{-}4S]^{+2}$ cluster, rendering the enzyme inactive [8,43]. Sensitive organisms had to adapt to these conditions, by either adjusting the amounts of vulnerable proteins or changing the structure of their proteins so that they were less prone to degradation [23,43,46–49]. It is not currently fully established which evolutionary path(s) radical SAM enzymes took. This situation could prove to have especially interesting implications for aerobic eukaryotes, such as the parasite *Entamoeba histolytica*, which are believed to have at one point performed aerobic respiration via mitochondria, but then lost that capability [56]. Plants and other eukaryotes capable of oxygenic photosynthesis might also have unique sets of radical SAM enzymes. Further examination of the *E. histolytica* and cotton plant *Gossypium barbadense* proteomes for radical SAM motifs could yield novel types of uncharacterized radical SAM enzymes. Doing so may provide insights into the evolution of oxygen tolerance, which has not been quantified this way.

The radical SAM superfamily was established through bioinformatic means, and we seek to expand on previous computational findings. Here, we develop a method to identify candidates for future analysis, based on the presence of the radical SAM motifs and the use of Hidden Markov Models (HMMs), across the tree of life. We analyze the trends in the abundance of these enzymes as a function of oxygen sensitivity of the organisms. Our study examines the ability of the motif search to identify radical SAM enzymes from sets of standard radical SAMs and ferredoxins. To analyze and observe the presence of putative radical SAM enzymes across the tree of life, the proteomes of 400 randomly-selected organisms across the tree of life, 22 model organisms, 134 prokaryotic organisms that are categorized as aerobic or anaerobic, 51 aerobic or anaerobic/microaerophilic eukaryotes, and 60 photosynthetic or non-photosynthetic eukaryotes, were used. We then further explore the putative radical SAM enzymes identified from *Entamoeba histolytica* and *Gosspyium barbadense*. Results indicate some correlations between oxygen levels and the abundance of these enzymes, and some possible functions of some of the enzymes that can be further analyzed.

2.3 Materials and Methods

2.3.1 Methods: Data Curation and Retrieval from UniProt Database

2.3.1.1 Methods: Proteome Selection

To predict the presence and amount of putative radical SAM enzymes across the tree of life, 100 proteomes from the UniProt public database were selected [67], including each of the three domains of life and viruses (Supplemental Table 5**)**. These 400 proteomes were selected from four randomized lists of the proteomes available in the UniProt database. UniProt proteomes vary in degrees of completeness and accuracy, so only proteomes with a "Standard" Complete Proteome Detector (CPD) level were included. BUSCO (Benchmarking Universal Single-Copy Ortholog) [86] scores were not used as a parameter in this part of the analysis, because but these scores were only available for bacteria and eukaryotes. Excluding bacteria and eukaryotes, but not viruses or archaea, could have potentially biased the archaea and virus data to give the impression that it had a relatively large number of radical SAM proteins compared to the other domains.

Twenty-two model organisms that are commonly used in molecular biology studies were selected to serve as points of reference in proteome analysis (Table 2**)**. To compare how oxygen

tolerance is correlated with the amounts of radical SAM enzymes organisms contain, the information for obligately aerobic prokaryotic organisms and obligately anaerobic prokaryotic organisms was selected from the BacDive database [87], and were manually narrowed to sets of 65 aerobic and 69 anaerobic organisms that had proteomes with Standard CPDs in the UniProt database **(**Supplemental Table 5). These processes resulted in lists of organisms and their UniProt proteome accession numbers. The proteomes were downloaded from UniProt in batches using the Bioservices package (version 1.6.0) [88] in Python (version 3.7.3), and manually checked for completeness.

Table 2. Model organisms included in proteome analysis. UniProt lists all of these as "Reference" proteomes. Their CPDs, a measure of the proteomes' completeness, vary in quality.

Organism	Common name	UniProt Proteome ID	UniProt Proteome CPD
Arabidopsis thaliana	mouse-ear cress	UP000006548	Close to Standard
Bos taurus	cattle	UP000009136	Close to Standard
Caenorhabditis elegans	roundworm	UP000001940	Standard
Chlamydomonas reinhardtii	a green algae	UP000006906	Outlier
Danio rerio	zebrafish	UP000000437	Outlier
Dictyostelium discoideum	slime mold	UP000002195	Standard
Drosophila melanogaster	fruit fly	UP000000803	Close to Standard
Escherichia coli O157:H7	*E. coli* strain	UP000000558	Standard
Escherichia coli K12	*E. coli* strain	UP000000625	Standard
Hepatitis C virus	hepatitis C	UP000000518	Outlier
Homo sapiens	human	UP000005640	Close to Standard
Mus musculus	house mouse	UP000000589	Close to Standard
Mycoplasma pneumoniae	mycoplasma	UP000000808	Standard
Oryza sativa	rice	UP000059680	Standard
Plasmodium falciparum	malaria parasite	UP000001450	Standard
Pneumocystis carinii	ascomycete	UP000011958	Standard
Rattus norvegicus	Norway rat	UP000002494	Standard
Saccharomyces cerevisiae	baker's yeast	UP000002311	Standard
Schizosaccharomyces pombe	fission yeast	UP000002485	Standard
Takifugu rubripes	Japanese pufferfish	UP000005226	Standard
Xenopus laevis	African clawed frog	UP000186698	Close to Standard
Zea mays	corn	UP000007305	Close to Standard

Additional sets of proteomes were selected so that differences within Eukarya could be tested, including the difference between photosynthetic and non-photosynthetic eukaryotes. First, a list of all UniProt proteomes from eukaryotes with Standard CPDs was downloaded. Photosynthetic eukaryotes were identified based on their membership in the Viridiplantae, Rhodophyta, and Ochrophyta clades. To ensure that the best possible proteomes would be utilized,

the final set of proteomes of photosynthetic eukaryotes was narrowed down to those with BUSCO completeness scores of 95% or higher (n = 31). The complementary set of proteomes of non-photosynthetic organisms with BUSCO completeness scores of over 95% was randomized, and the final selection (n = 29) was chosen so that the distribution of BUSCO completeness scores was not significantly different from those of the photosynthetic organisms (t-test p-value of 0.9539).

A similar process was used to choose the sets of proteomes from anaerobic and aerobic organisms. Anaerobic/microaerophilic eukaryotes were chosen based on being listed by Altenbach *et al.* [56] as containing mitosomes or hydrogenosomes rather than mitochondria. Most of these proteomes had low BUSCO completeness scores: between 42.2% and 76.5% (n = 21) when outliers – data points outside the whiskers of a boxplot - were excluded. The complementary set of proteomes from aerobic eukaryotes (n = 30) was chosen to have a similar distribution of BUSCO completeness scores, such that the BUSCO completeness score would not be a confounding variable. A t-test between the BUSCO completeness scores of the two final sets yielded an insignificant score (p-value of 0.9948).

2.3.1.2. Methods: Establishment of Sets of Standard Radical SAM and Ferredoxin Sequences

Sets of reference proteins and reference protein signatures were obtained from UniProt. Two sets of standard radical SAM proteins and those of their close relatives, ferredoxins, were downloaded from UniProt [67]. Protein sets were selected by searching UniProt through the Bioservices package (version 1.6.0) [88] in Python (version 3.7.3). Proteins were selected based on a minimum criteria including reviewed status by curators, evidence at the protein level, and either belonged to the ferredoxin family or contained "radical SAM" in one of their fields. These sets were downloaded separately. Sets were then checked manually for false positives, so that UnipProt accessions DRDA_BACT7 and CAF17_YEAST were excluded from the final dataset because they are not radical SAM enzymes. Dph2 proteins were not considered to be a part of the radical SAM superfamily in this study because they do not form a TIM barrel structure [6]. The protein PHNJ_ECOLI was added to include a representative with a $CX_2CX_{21}CX_5C$ motif. This process resulted in sets of 119 radical SAM sequences, and 149 ferredoxin sequences (Supplemental Table 4). These two sets of proteins were then used to assess how well our methods could detect radical SAM enzymes and exclude false positives.

Accessions of radical SAM enzymes with characterized crystal structures was listed in Pfam, and those sequences were downloaded from UniProt. These sequences were later used to assess suitability of homology modelling and other computational biochemistry techniques.

2.3.2. Methods: Identification of Putative Radical SAM Enzymes

2.3.2.1 Methods: Identification and Quantification of Putative Radical SAM Enzymes Using Regular Expression/Motif Search

Proteome datasets were searched for those sequences that contained radical SAM motifs that have been described in literature (Table 3), which yielded mixed results. These motifs are listed in Table 1. These searches for regular expressions in FASTA files were performed using the Biopython package (version 1.74) [89] in Python (version 3.7.3).

Table 3. Radical SAM motifs found in peer-reviewed literature, excluding motifs that bind auxiliary clusters. Since the establishment of the radical SAM superfamily in 2001, other radical SAM motifs have been discovered. These non-canonical motifs either have more amino acids between the first two cysteines than the canonical motif does or have more divergent patterns.

Motif	**Literature Source**
CX_3CX_2C (canonical)	Sofia *et al.*, 2001[1], Layer et al., 2004[11]
CX_4CX_2C	Sofia *et al.*, 2001[1], Selvadurai et al., 2014[12]
CX_5CX_2C	Sofia *et al.*, 2001[1], McGlynn *et al.*, 2010[14]
CX_7CX_2C	Parent et al., 2016[15]
CX_8CX_2C	Thweat et al., 2016[16]
CX_9CX_2C	Greenwood *et al.*, 2009[17]
$CX_{14}CX_2C$	Dowling et al., 2014[18]
$CX_2CX_{21}CX_5C$	Kamat *et al.*, 2013[19]
CX_2CX_4C	Chatterjee et al., 2008[20], Fenwick *et al.*, 2014[21]

Table 4. Motif counts for sets of standard Radical SAMs and ferredoxins. The length cutoff refers to excluding sequences with the CX_4CX_2C motif that are shorter than 400 amino acids.

	Without CX_4CX_2C length cutoff		With CX_4CX_2C length cutoff	
Motif	**Standard Radical SAMs**	**Standard Ferredoxins**	**Standard Radical SAMs**	**Standard Ferredoxins**
CX_3CX_2C	110	2	110	2
CX_4CX_2C	1	***106***	1	***0***
CX_5CX_2C	3	0	3	0
CX_7CX_2C	1	3	1	3
CX_8CX_2C	2	0	2	0
CX_9CX_2C	5	0	5	0
$CX_{14}CX_2C$	1	0	1	0
$CX_2CX_{21}CX_5C$	1	0	1	0
CX_2CX_4C	3	5	3	5
Total Proteins	119	149	119	149
Total Radical SAMs	119*	***112***	119*	*7*

*Radical SAMs can have more than one iron-sulfur-binding cluster and this was taken into account when total radical SAM sequences were calculated.

Initially, the search of the motifs in the standard radical SAM and ferredoxins sequences resulted in a very large number of false positives from ferredoxins with the CX_4CX_2C motif and others (false positive rate of 75.2%), so sequences with CX_4CX_2C motifs that were shorter than 400 amino acids were excluded. This modification decreased the false positive rate to 4.70%. (Table 4). However, because many radical SAM enzymes are shorter than 400 amino acids, this criterion likely increased the rate of false negatives of this method.

The motif search was then performed on full proteomes (Supplemental Table 5). For each batch of proteomes, the script automatically generated a CSV file with rows listing each proteome, the number of each type of motif match, the total number of proteins, the total number of sequences that contained at least one of the radical SAM motifs, and the total numbers of proteins UniProt marked as "Uncharacterized." The CSV files were generated with the Python Pandas package (version 0.24.2) [90]. Further examination of the sequences that matched the radical SAM motifs revealed a large number of sequences that are highly unlikely to be radical SAM enzymes. Many of these sequences belonged to well-characterized proteins that are not radical SAM enzymes, or were sequences sharing little identity with known radical SAM enzymes. These results prompted a new approach.

2.3.2.2 Methods: Identification and Quantification of Putative Radical SAM Enzymes Using Hidden Markov Models (HMMs)

HMMs were first tested on the sets of radical SAM and ferredoxin standards, and yielded positive results. HMMER (version 3.2.1) was used (hmmer.org), in a method based on the protocol by Aylward [91]. The HMM models were built from the seed alignments of Pfam families PF04055, PF06007, and PF01964. An E-value cutoff of 1e-3 was used for the HMM search against the proteome FASTA files, as recommended in the manual (hmmer.org). Proteins in other preliminary analyses with these lower E-values still retained the characteristics of radical SAM enzymes. These HMM searches against the standard radical SAM sequences and standard ferredoxin sequences yielded 0 false positives and 0 false negatives.

HMMs yielded more reliable results than the motif search, so the results produced by HMMs will be the focus for the remaineder of the study. The HMM analyses were performed on proteomes grouped by hypothesis test, since HMM E-values depend on the size of the database that hmmsearch is performed on. An E-value cutoff of 1e-3 was utilized again. R (version 3.3.2) and Python were then used to combine the HMM sequence hits with metadata about each proteome (organism ID, proteome ID, the dataset each organism belonged to, etc.), using the Biopython (version 1.74) [89] and Pandas (version 0.24.2) [90] Python packages. Doing so allowed us to efficiently identify possible radical SAM enzymes from a large number of organisms and quickly summarize the data.

2.3.3 Methods: Statistical Analysis

Statistical analyses were performed in R (version 3.3.2) to test the differences in the amounts of radical SAM enzymes in different types of organisms. Kruskal-Wallace tests were performed to test the differences in counts and percentages of radical SAM proteins between the different domains of life, and were followed up with post-hoc Dunn tests (dunn.test version 1.3.5). The differences between counts and percentages of radical SAM enzymes in aerobic and anaerobic

prokaryotes, aerobic and anaerobic eukaryotes, and photosynthetic and non-photosynthetic eukaryotes were analyzed with two-sided t-tests.

2.3.4 Methods: Analysis of Putative Radical SAMs in Select Eukaryotic Organisms

2.3.4.1 Methods: Putative Radical SAMs in *Entamoeba histolytica*

Putative radical SAM enzymes in *Entamoeba histolytica* (UniProt proteome ID UP000078387) were chosen for further analysis. This organism had a high percentage of radical SAM enzymes based on the motif search, although the results varied significantly in the HMM analysis. Nevertheless, this proteome is valuable because this organism is an anaerobic/microaerophilic eukaryote, and we suspected that the evidence that some prokaryotic genes have been transferred to its genome could prove relevant for the study of radical SAM enzymes [58,59,92]. Further, many of this organism's proteins have not been characterized, suggesting that it is an understudied organism that would benefit from high-throughput computational analysis.

To predict the functions of any putative radical SAM enzymes, we performed BLAST comparisons of the four *Entamoeba histolytica* proteins that had matched radical SAM HMM profiles, against the sets of standard radical SAM enzymes (Table 5). We also performed BLAST comparisons of these four proteins and the sequences of radical SAM enzymes with solved crystal structures (Table 6). One possible *E. histolytica* methylthiotransferase, (Uniprot ID A0A5K1U8H1_ENTHI), was selected to assess its suitability for homology modelling, based on its relatively high degree of sequence similarity (25.48%) to *Thermotoga maratima* RimO, a radical SAM. The RimO structure was downloaded from PDB ID 4JC0 [93].

The A0A5K1U8H1_ENTHI sequence was submitted to I-TASSER [94] for homology modelling. Energy minimization was then performed in Schrödinger [95]. Model validation on the energy minimized model was performed in Swiss-Model [96]. A structural overlay was produced from the energy minimized model and the structure that I-TASSER had used as its primary reference structure (PDB 4JC0). To determine the quality of the model in terms of the formation of a likely $[4Fe\text{-}4S]^{+2}$ binding site, the distances between the cysteine residues around the corresponding location in the model and the primary iron-sulfur cluster in RimO were measured (Figure 9), as were the distances between each cysteine residue, using PyMOL (version 1.7.4.5) (Supplemental Figure 2) [97].

2.3.4.2 Methods: Putative Radical SAMs in *Gossypium barbadense*

Preliminary results indicated that plants had higher amounts of radical SAM enzymes in eukaryotes, and that photosynthetic eukaryotes seemed to have high levels compared to non-photosynthetic eukaryotes. The 37 sequences that fit radical SAM HMM profiles from *G. barbadense* (UP000327439) were further characterized. As with *E. histolytica*, we performed BLAST comparisons between these sequences with HMM matches against standard sets of radical SAM enzymes (Supplemental Table 1), and against a set of radical SAM enzymes with solved crystal structures (Supplemental Table 2).

Analysis was also performed to compare the *G. barbadense* hits against the 16 *Arabidopsis thaliana* (UP000006548) hits. Because *A. thaliana* is a relatively well-characterized model organism, this plant was thought to be a good baseline to compare the *G. barbadense* hits; in fact,

some of the sequences were already included in the set of standard radical SAM sequences. A tree was constructed in MAFFT [98] and visualized in iTOL [99]. The tree was built from the standard radical SAM sequences, the 37 *G. barbadense* (UP000327439) hits, and the 16 *A. thaliana* (UP000006548) relatively well-characterized hits. MAFFT settings were as follows: Output format: Pearson/fasta (default), Matrix: BLOSUM62 (default), Gap open penalty: 1.53 (default), Gap extension penalty: 0.123 (default), Order: aligned (default), Tree rebuilding number: 100 (raised from default 2 to be maximally rigorous), Guide tree output: ON (default), Maxiterate: 100 (raised from default 2 to be maximally rigorous), Perform FFTS: none (default). This phylogenetic protein tree showed a potentially interesting relationship between three *Gossypium barbadense* proteins (A0A5J5PDX9, A0A5J5PNT2, and A0A5J5U143) and one of the *Arabidopsis thaliana* proteins (Q8H0V1), so these were compared in a separate BLAST analysis (Table 7).

2.4 Results

2.4.1 Results of Large-Scale Proteome HMM Searches

The results of the search for sequences that matched HMM profiles of radical SAM enzymes are summarized in Figures 4-7 and Supplemental Figure 1, and the full results are contained in Supplemental Table 7 and Supplemental Table 8. Most of the matches are from the Pfam PF04055 family, which contains radical SAM enzymes with canonical motifs and variations. Most living organisms have at least a few sequences that match radical SAM profiles, although some exceptions exist, including *Tremblaya princeps*, *Mycoplasma pneumoniae*, and *Lactobacillus paucivorans*. None of the viruses analyzed contained any sequences that matched radical SAM HMM profiles. In organisms that had sequences with HMM matches, the counts ranged from 1 to 131, and their percentages in the proteome ranged from 0.01% to 3.95%. Many methanogens appear to be at the high end of the count and percentage spectrum in archaea, although this outcome was not analyzed further. It was observed that the majority of eukaryotic proteomes used in the analyses belonged to funguses.

We found statistically significant differences between some groups of organisms (Figure 4). The Kruskal Wallis test between the counts in the three domains of life yielded a p-value of $p < 2.2e\text{-}16$. This means that the Kruskal Wallis test determined that the three distributions were not identical. The follow-up Dunn test found a $p = 0.1366$ between archaea and bacteria, $p < 1e\text{-}4$ between archaea and eukaryotes, as well as between bacteria and eukaryotes. A way to interpret this is that the comparisons that yielded p-values less than an alpha of 0.05 have distributions that are statistically significantly different from each other. The p-values are based on the means, standard deviations, and (large) sample sizes of each of the groups compared. The Kruskal Wallace test between the percentages of putative radical SAM enzymes in proteomes of the three groups of organisms yielded a p-value of $p < 2.2e\text{-}16$. The follow-up Dunn test determine $p=0.0001$ between archaea and bacteria, and $p < 1e\text{-}4$ between archaea and eukaryotes, as well as between bacteria and eukaryotes.

The number of radical SAM motifs appear to be different between aerobic and anaerobic prokaryotes (Figure 5). The t-tests yielded a yielded $p = 8.165e\text{-}12$ for counts, and $p < 2.2e\text{-}16$ for percentages. Comparisons between eukaryotes that live in different oxygen levels had more mixed results. In aerobic vs. anaerobic eukaryotes (Figure 6), the t-test between the counts yielded $p=0.1937$. The percentages yield a $p=0.0936$, but it is decreased to $3.686e\text{-}05$ if the information from *Nephila clavipes*, an extreme outlier with 0.33% radical SAM enzymes, is excluded.

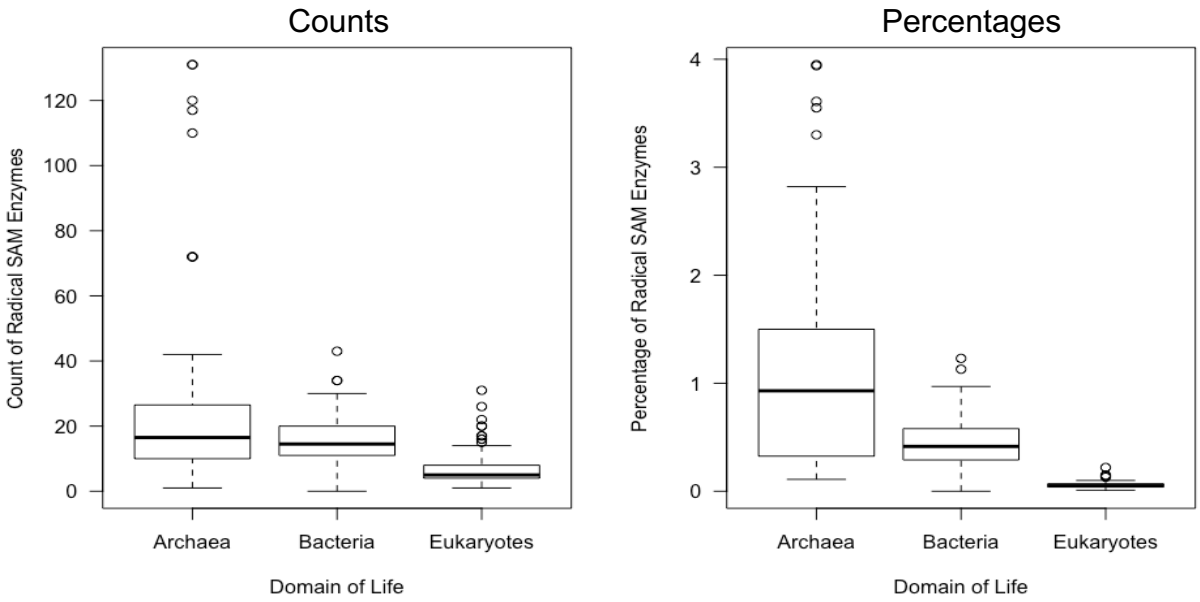


Figure 4. Boxplots of the percentages of putative radical SAM enzymes found in the proteomes of the 400 organisms randomly chosen from the three domains of life. None of the sequences in any of the 100 virus proteomes matched a radical SAM HMM signature. The archaeal outliers are strains of *Methosarcian mazei* and *Saccarolobus solfataricus*. Some of the bacterial outliers include *Acidobacteria bacterium* and *Clostridiodes difficile*. Some of the eukaryotic outliers are plants, mammals, and algae. The exhaustive list of these results is in Supplemental Table 8.

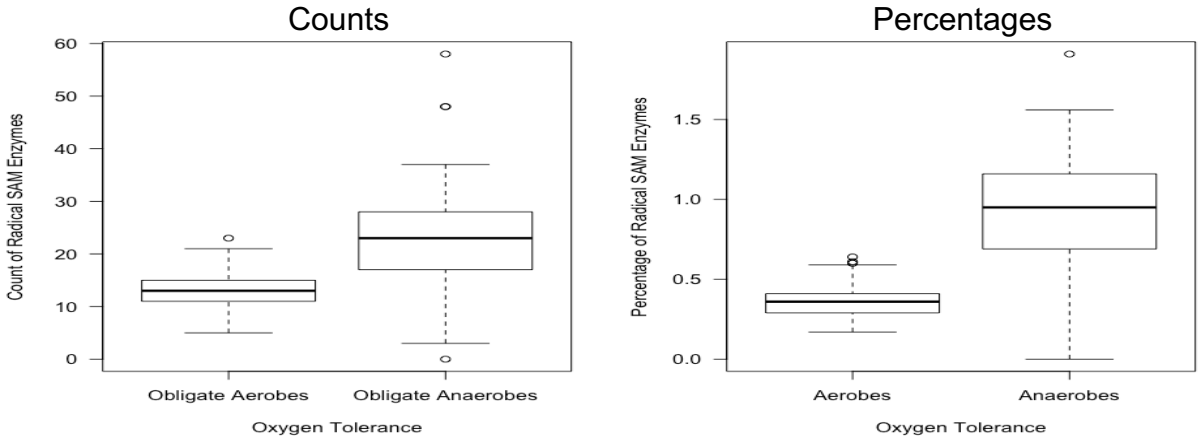


Figure 5. Boxplots comparing the amounts of putative radical SAM enzymes found in the proteomes of aerobic prokaryotes (n=65) and anaerobic prokaryotes (n=69). There is a significant difference between the two groups of organisms, in both the count of radical SAMs present, and the percentage of radical SAMs present. Some of the aerobic outliers include *Aeribacillus composti* and *Acetobacter senegalensis*, and anaerobic outliers include *Lachnotalea glycerini*, and *Thermatoga petrophila*. The exhaustive list of these results is in Supplemental Table 8.

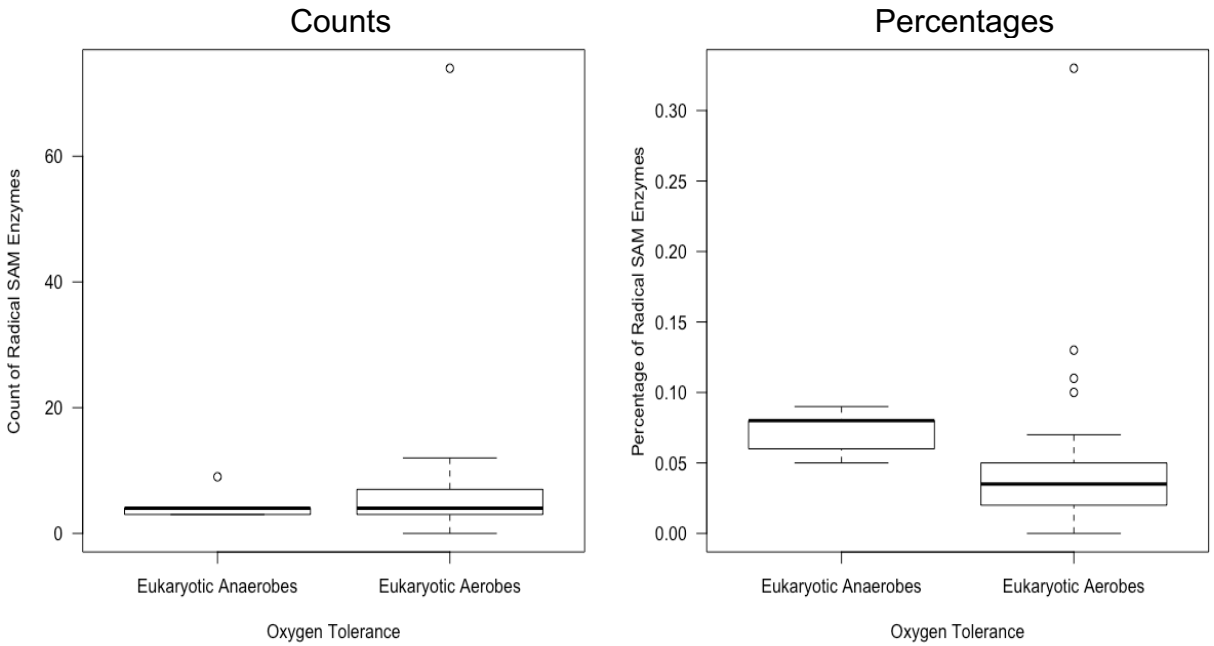


Figure 6. Boxplots comparing counts and percentages of putative radical SAM enzymes from aerobic eukaryotes (n=30) and anaerobic/microaerophilic eukaryotes (n=21). The anaerobic outlier is a strain of *Piromyces*. The most dramatic aerobic outlier is *Nephila clavipes*, and others include *Syncephalis pseudolumigal* and *Sugiyamaella lignohabitans*. The exhaustive list of these results is in Supplemental Table 8.

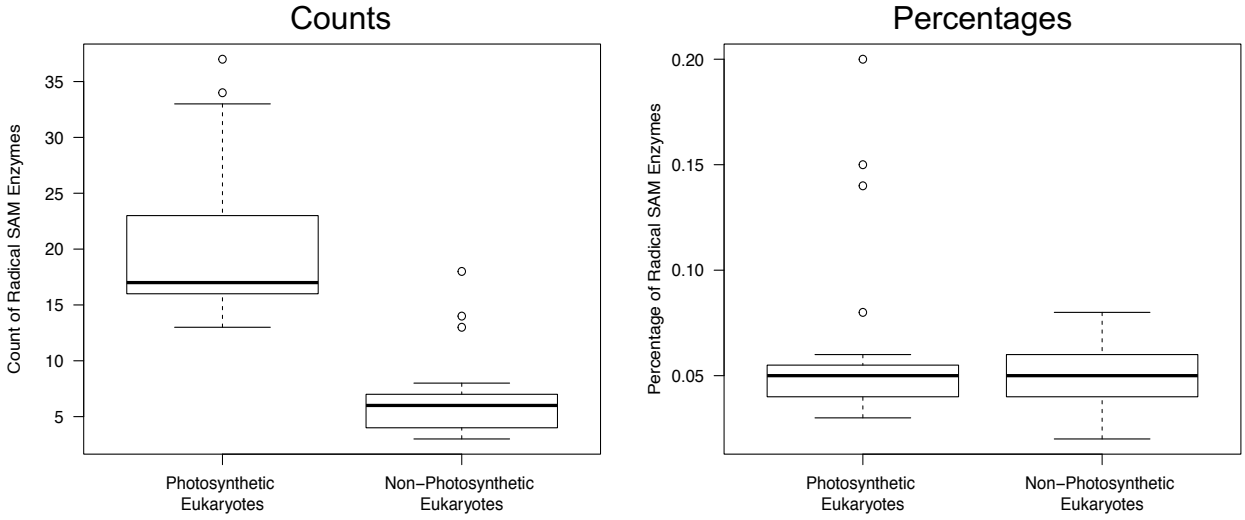


Figure 7. Boxplots comparing counts and percentages of putative radical SAM enzymes in photosynthetic eukaryotes (n=31) and non-photosynthetic eukaryotes (n=29). Outliers in photosynthetic eukaryotes include species of *Gossypium* and algae. Outliers in non-photosynthetic eukaryotes include *Saimiri boliviensis* and *Byssochlamys spectabilis*. The exhaustive list of these results is in Supplemental Table 8.

2.4.2 Results: Sequence and Structural Analysis of Putative *Entamoeba histolytica* Radical SAM Enzymes

The HMM search marked four sequences from the *E. histolytica* proteome (UniProt proteome ID UP000078387). Each of these sequences has a high degree of similarity to at least one of the sequences in the set of standard radical SAM enzymes (Table 5). The four *E. histolytica* sequences are similar to a methylthiotransferase, a glycyl-radical activating enzyme, an elongator complex protein 3, and an anaerobic sulfatase-maturating enzyme, respectively. Alignments between these four sequences and those of radical SAM enzymes with solved crystal structures exhibit a low degree of sequence identity (Table 6). Despite the low sequence identity, we attempted to build a structural model of the *E.. histolytica* sequence A0A5K1U8H1_ENTHI.

Table 5. Top BLAST hits between each of the four *Entamoeba histolytica* (UP000078387) sequences found from HMM profile searches, and the radical SAM standard sequences

Entamoeba histolytica **Query Sequences**	**Radical SAM Sequences**	**Query Cover**	**E-value**	**Percent Identity**
A0A5K1U8H1_ENTHI	*Homo sapiens* CDKAL_HUMAN Threonylcarbamoyladenosine tRNA methylthiotransferase	97%	6e-131	46.21%
A0A5K1UL08_ENTHI	*Escherichia coli* YJJW_ECOLI Putative glycyl-radical activating enzyme	52%	1e-09	23.08%
A0A5K1UWY4_ENTHI	*Arabidopsis thaliana* ELP3_ARATH Elongator complex protein 3	93%	0.0	58.21%
A0A5K1VD25_ENTHI	*Clostridium perfringens* ANSME_CLOP1 Anaerobic sulfatase-maturating enzyme	90%	2e-12	23.76%

Table 6. The top blast hits between the four *Entamoeba histolytica* (UP000078387) proteins found from HMM search, and radical SAM enzymes with solved crystal structures

Entamoeba histolytica **Query Sequences**	**Radical SAM Sequences**	**Query Cover**	**E-value**	**Percent Identity**
A0A5K1U8H1_ENTHI	*Thermatoga maritima* RIMO_THEMA ribosomal protein S12 methylthiotransferase	79%	2e-31	25.48%
A0A5K1U8H1_ENTHI	*Thermotoga maritima* HYDE_THEMA [FeFe] hydrogenase maturase subunit	57%	5e-04	24.63%
A0A5K1UL08_ENTHI	*Escherichia coli* PFLA_ECOLI Pyruvate formate-lyase 1-activating enzyme	61%	2e-08	18.42%
A0A5K1UL08_ENTHI	*Bacteroides vulgatus* A6L094_BACV8 Pyruvate-formate lyase-activating enzyme	48%	2e-05	21.55%
A0A5K1UL08_ENTHI	*Staphylococcus aureus* MOAA_STAAN GTP 3',8-cyclase	45%	2e-05	23.67%
A0A5K1UWY4_ENTHI	*Streptoalloteichus tenebrarius* Q2MFI7_STRSD Putative apramycin biosynthetic oxidoreductase	34%	0.067	22.30%
A0A5K1VD25_ENTHI	*Clostridium perfringens* ANSME_CLOP1 Anaerobic sulfatase-maturating enzyme	90%	3e-13	23.76%
A0A5K1VD25_ENTHI	Hungateiclostridium thermocellum A3DDW1_HUNT2 Radical SAM domain protein	69%	2e-11	25.35%

The output of the validation of the *E. histolytica* A0A5K1U8H1_ENTHI structural model is summarized in Figure 8. The quality scores along different portions of the protein structure are highlighted in Figure 8B, in which residues with low scores are red, and residues with high scores are blue. The QMEAN Z-scores (Figure 8C) for the Cβ, all-atom, and solvation geometrical properties are close to zero, indicating reasonable agreement between this model and those of structures of a similar size [100]. The QMEAN Z-scores for torsion and the model overall are much lower than the acceptable threshold of -4 (Figure 8C). This indicates that the model has overall low-quality [100].

The structural overlays are shown in Figure 9 and Supplemental Figure 2. Supplemental Figure 2 shows the measured distances between the sulfur atoms of the cysteine residues that comprise the "nest" that coordinates the iron-sulfur cluster [42].

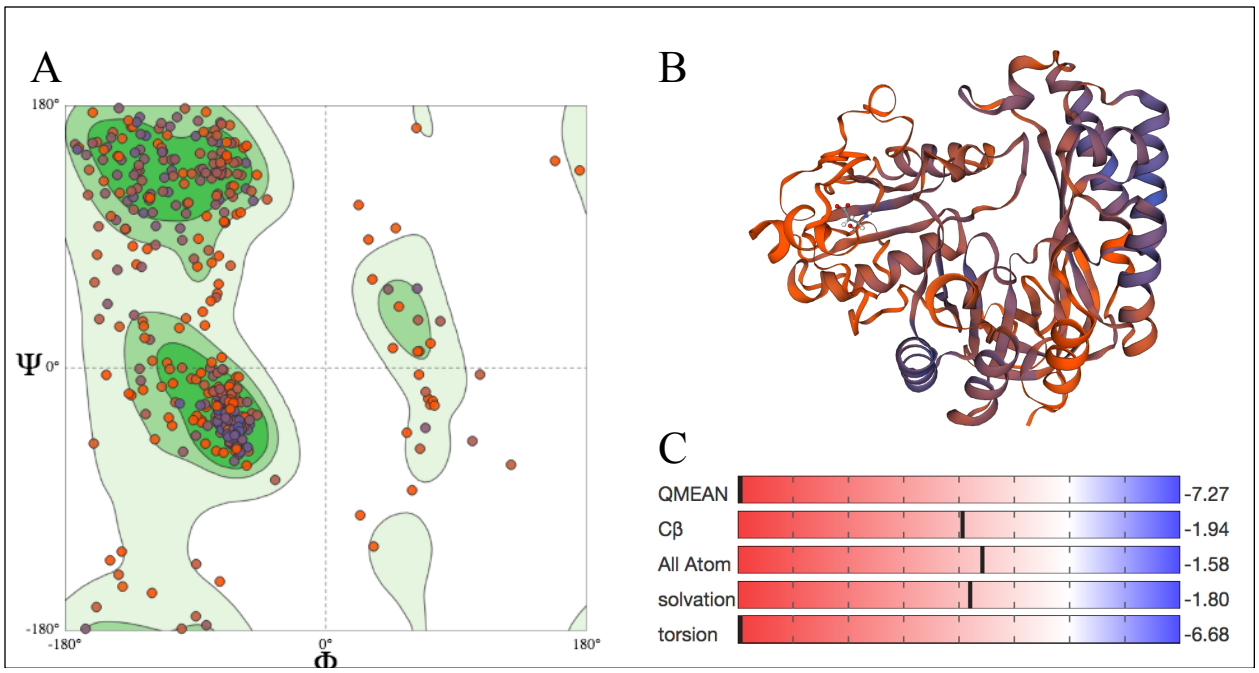


Figure 8. Validation of energy-minimized A0A5K1U8H1_ENTHI homology model. (A) Ramachandran Plot. (B) Cartoon depiction of model, colored by QMEAN model quality. Blue is high quality and red is low quality. (C) QMEAN quality estimates. The torsion and overall QMEAN Z-scores are very low. The other individual scores are reasonable and do not pass the threshold of -4.

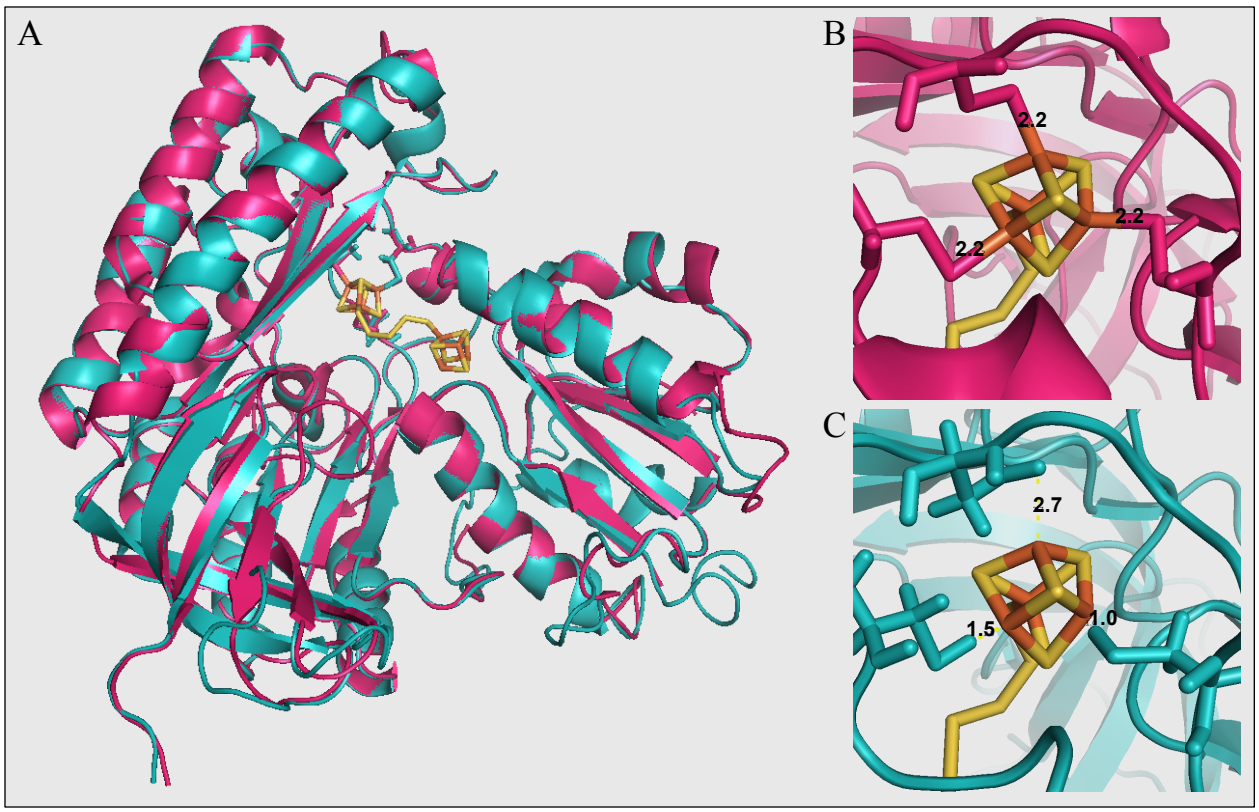


Figure 9. Homology modelling structural overlays. (A) Structural overlay of homology model of A0A5K1U8H1_ENTHI from *Entamoeba histolytica* (teal), and RimO (PDB 4JC0) template with cofactors (magenta). (B) 4JC0 template and iron-sulfur cluster. (C) A0A5K1U8H1_ENTHI homology model overlaid with iron-sulfur cluster from 4JC0 template. Distances between cysteine Sγ atoms and the iron atoms are measured in Å in panels (B) and (C).

2.4.3 Results: Sequence Analysis of Putative *Gossypium barbadense* Radical SAM Enzymes

The radical SAM HMM profiles matched 37 sequences from the *Gossypium barbadense* (UP000327439) proteome. BLAST alignments between these sequences and the set of standard radical SAM enzymes show that most of them are homologous sequences (Supplemental Table 1). However, one sequence (A0A5J5PQX6_GOSBA) was found to have low similarity to any of the sequences. This sequence also had a relatively unfavorable E-value (0.00089) in the HMM profile search, so it is likely not a radical SAM enzyme. BLAST alignments between the *G. barbadense* sequences and radical SAM enzymes with solved crystal structures have some sequences with a relatively high degree of similarity (Supplemental Table 2).

All of the 36 *Gossypium barbadense* sequences that are likely to be radical SAM enzymes are similar to the 16 *Arabidopsis thaliana* sequences to some degree, as seen in the phylogenetic tree (Figure 10). The relationship between the group of sequences with the longest branch lengths is summarized in Table 7. The query cover and E-values between the searches are good. The degrees of percent identity between the proteins are high but show some degree of divergence.

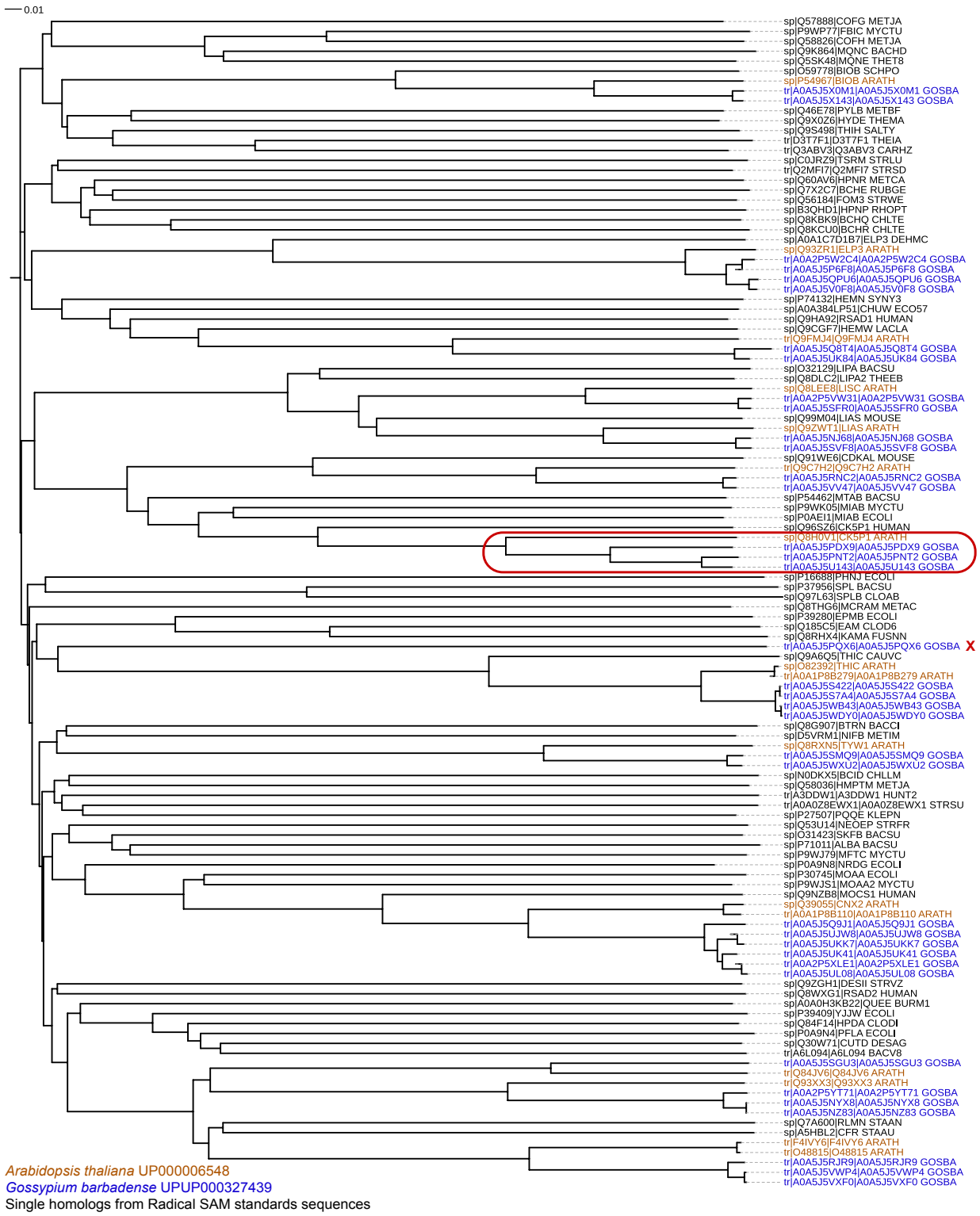


Figure 10. Tree of MAFFT alignment between selected standard radical SAM sequences, and *Arabidopsis thaliana* (brown) and *Gossypium barbadense* (blue) sequences that matched radical SAM HMM profiles. The relationship between the circled proteins are explored in Table 7. The A0A5J5PQ sequence from *G. barbadense* did not appear to be a radical SAM enzyme upon further inspection.

Table 7. BLAST between *Gossypium barbadense* (UP000327439) uncharacterized protein A0A5J5PDX9 and nearby neighbors circled in Figure 10.

Description	Query Cover	E-value	% Identity
Gossypium barbadense uncharacterized protein A0A5J5PNT2	99%	0	85.32%
Gossypium barbadense uncharacterized protein A0A5J5U143	99%	0	79.71%
Arabidopsis thaliana CDK5RAP1-like protein Q8H0V1	91%	0	75.26%

2.5 Discussion

2.5.1 Hidden Markov Model (HMM) Method Discussion

Utilization of HMMs provided much more reliable results than simple motif searches to identify putative radical SAM enzymes. The high number of motif matches in the eukaryotes was rational given that they have thousands more sequences in which the pattern can match by random chance. We found that the HMMs provided a more statistically robust method of detecting radical SAM enzymes that still incorporated information about the conserved cysteine motifs. This search alone does not reveal what the functions of the proteins are, but narrows down the candidates for laboratory experiments.

2.5.2 Large-Scale Proteome Searches and Comparisons Between Phylogenetic Groups Discussion

Differences were observed between phylogenetic groups. The most apparent difference, besides no viral sequences matching any of the HMMs, is the small number and percentage of putative radical SAM enzymes in eukaryotic organisms compared to either of the prokaryotic lineages (Figure 4). Eukaryotes had counts of 1-31 (mean = 7.24) and percentages of 0.01-0.22% (mean 0.058%), while prokaryotes had counts of 0-131 (mean = 19.82) and percentages of 0-3.95% (mean = 0.79%) (Figure 4, Supplemental Table 8). The stark difference might be correlated with the fact that most eukaryotes rely on high concentrations of oxygen and aerobic respiration.

Results should be critically interpreted because many of the eukaryotes that were included in the analysis turned out to be funguses. Although this was not tested, it appears in Supplemental Figure 1 that fungi may have fewer radical SAM enzymes, and at a lower percentage, than other eukaryotes. This means that it is possible that in the analyses between eukaryotic groups, the anaerobic eukaryotes vs. aerobic eukaryotes (Figure 7), and photosynthetic vs. non-photosynthetic eukaryotes (Figure 8), might be better interpreted as comparisons against fungi than against all eukaryotes with the given traits. The p-value between percentages of anaerobic vs. aerobic eukaryotes changed drastically when the outlier *Nephila clavipes*, a spider, was removed. It is important to consider that this was one of the rare proteomes that did not belong to a fungus that was included in that dataset. Future studies that examine these comparisons should utilize stratified random sampling or cluster random sampling, or compare the means between all groups, with fungi comprising one of the individual subgroups. In fungi, it might also be important to consider that they are haploid for most of their lives [54], while many plants are diploid or polyploid [101,102], although it is unknown in what manner UniProt takes ploidy into account in proteome

annotations. In future analyses, the BUSCO score could also be treated as a covariate, such that more organisms could be included in the analysis.

We also observed differences between organisms of different oxygen tolerances. This was especially apparent in the comparison between anaerobic and aerobic prokaryotes, which has been discussed in literature but perhaps not quantified before as we did here [11]. The results are consistent with the idea that there are well-characterized radical SAM proteins that are known to be directly involved in anaerobic metabolism [11,103,104].

2.5.3 *Entamoeba histolytica* and Other Anaerobic Eukaryotes Discussion

Correlation was observed between an anaerobic lifestyle and the quantity of radical SAM enzymes in organisms proteomes (Figure 6). With the current comparison it cannot be ruled out that the outcome reflected fungi more than then general aerobic eukaryote population, although the correlation that we found is in the expected direction. It also cannot be ruled out that the counts of radical SAM enzymes between the two groups is not significantly different because all eukaryotes could have a conserved set of radical SAM enzymes. It would be expected for anaerobic eukaryotes to have more radical SAM enzymes than aerobic eukaryotes, because of the different exposures to oxygen which degrades iron-sulfur clusters, and because of evidence that some prokaryotic genes have been transferred to the genomes of anaerobic eukaryotes [58,59,92]. It also follows the logic that radical SAM enzymes are sometimes involved in pathways under low-oxygen conditions but other types of enzymes are utilized under aerobic conditions [52,105].

Data gathered from the large-scale phylogenetic analysis aided in choosing candidates for finer-scale analysis of the functions of the HMM-matching sequences in *E. histolytica* (UP000078387). We were able to use this method to identify potential candidates to further characterize from the *E. histolytica* proteome. *E. histolytica* is especially interesting for studying radical SAM enzymes because iron-sulfur proteins tend to be different in these and other protozoan parasites compared to their hosts [106]. Phylogenetic analysis performed on similar differences in iron-sulfur proteins (or other markers) dependent on environment or lifecycle have yielded interesting results in past studies [47–49,107]. All four of the sequences identified by HMMs share a percentage of identity with more well-characterized radical SAM enzymes (Table 5), which increases the likelihood of future characterization.

The potential of studying these four enzymes through homology modeling is somewhat limited by the available radical SAM enzymes with solved crystal structures. None of the four sequences matched any of the radical SAM enzymes with solved crystal structures especially well (Table 6). The low percentage of identity between *Entamoeba histolytica*'s A0A5K1U8H1_ENTHI sequence and 4JC0 was reflected in the homology model.

If homology modeling, molecular dynamics simulations, docking, or other computational biochemistry techniques are to be used on any of these four *Entamoeba histolytica* enzymes, it appears necessary for the crystal structures of more homologous proteins to be solved. Because it is not possible to include the iron-sulfur cluster or other cofactors in homology models, it would also be necessary for the cluster to be super-positioned into the starting coordinates for molecular dynamics simulations. It is possible that at least a small amount of the low QMEAN scores can be attributed to the lack of the iron-sulfur cluster in the model, though most of the degradation in quality can be attributed to the fact that A0A5K1U8H1_ENTHI had a higher sequence identity to a different methylthiotransferase than the RimO methylthiotransferase that formed the basis for most of the model.

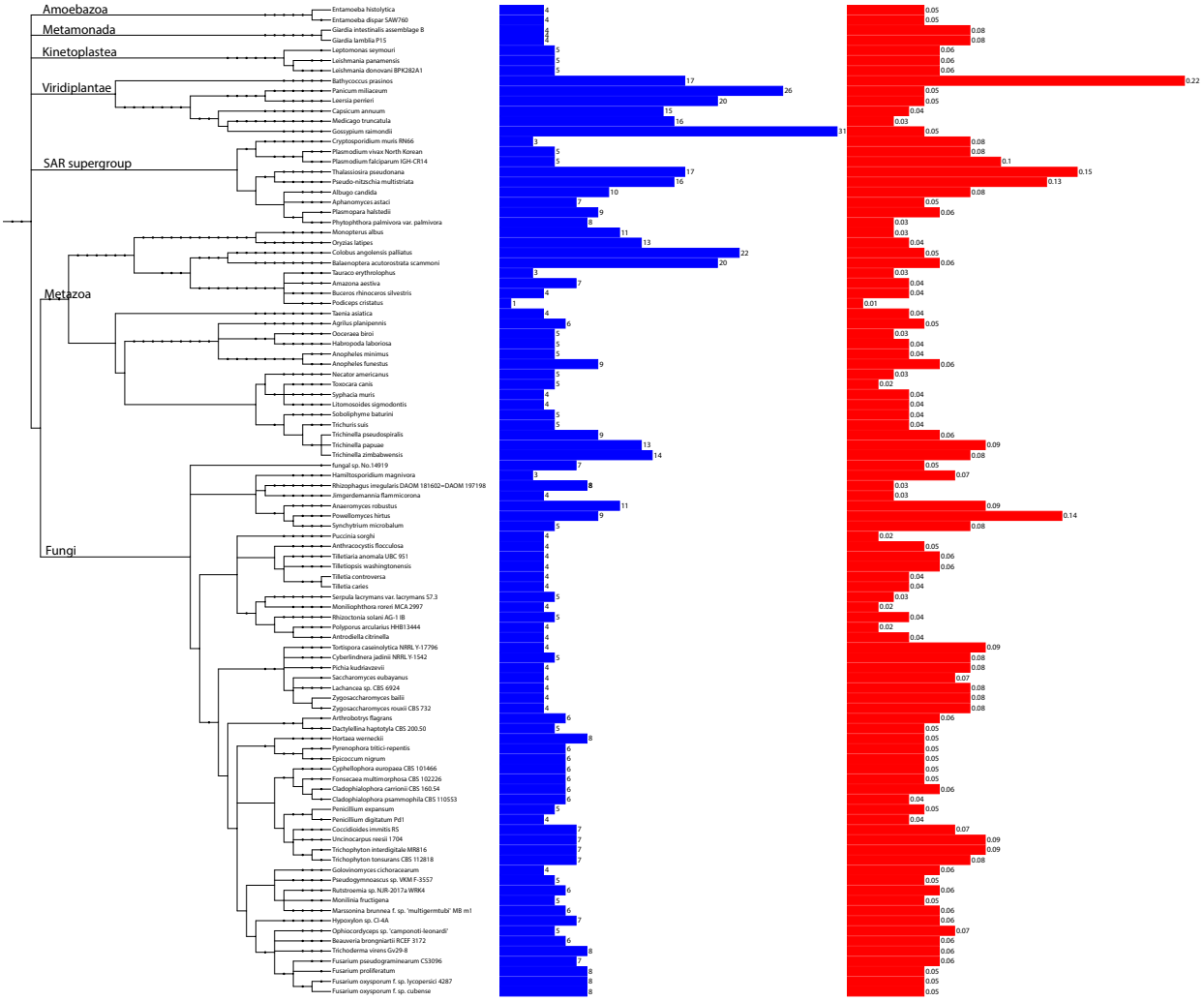
2.5.4 *Gossypium barbadense* and Other Photosynthetic Eukaryotes Discussion

Considering the other trends observed between oxygen levels and the prevalence of radical SAM enzymes, it was surprising to see a higher number of putative radical SAM enzymes in eukaryotes capable of oxygenic photosynthesis compared to other eukaryotes (Figure 7). Although at this time we are unable to rule out that the polyploidy of plants might be a confounding variable [101,102], the result might be consistent with the complex physiology of plants. For example, although it is well-established that oxygen levels in photosynthetic plant cells can build to the point that it inhibits the carbon-fixation enzyme [63,64], plants have developed Krantz anatomy and CAM photosynthesis to isolate RUBISCO from oxygen [54,64]. However, even oxygenic cyanobacteria are able to separate oxygen production from oxygen-sensitive processes [108]. Additional analysis would need to be performed to thoroughly test whether or not there is a trend in the abundance of radical SAM enzymes in organisms capable of producing oxygen, compared to organisms that do not.

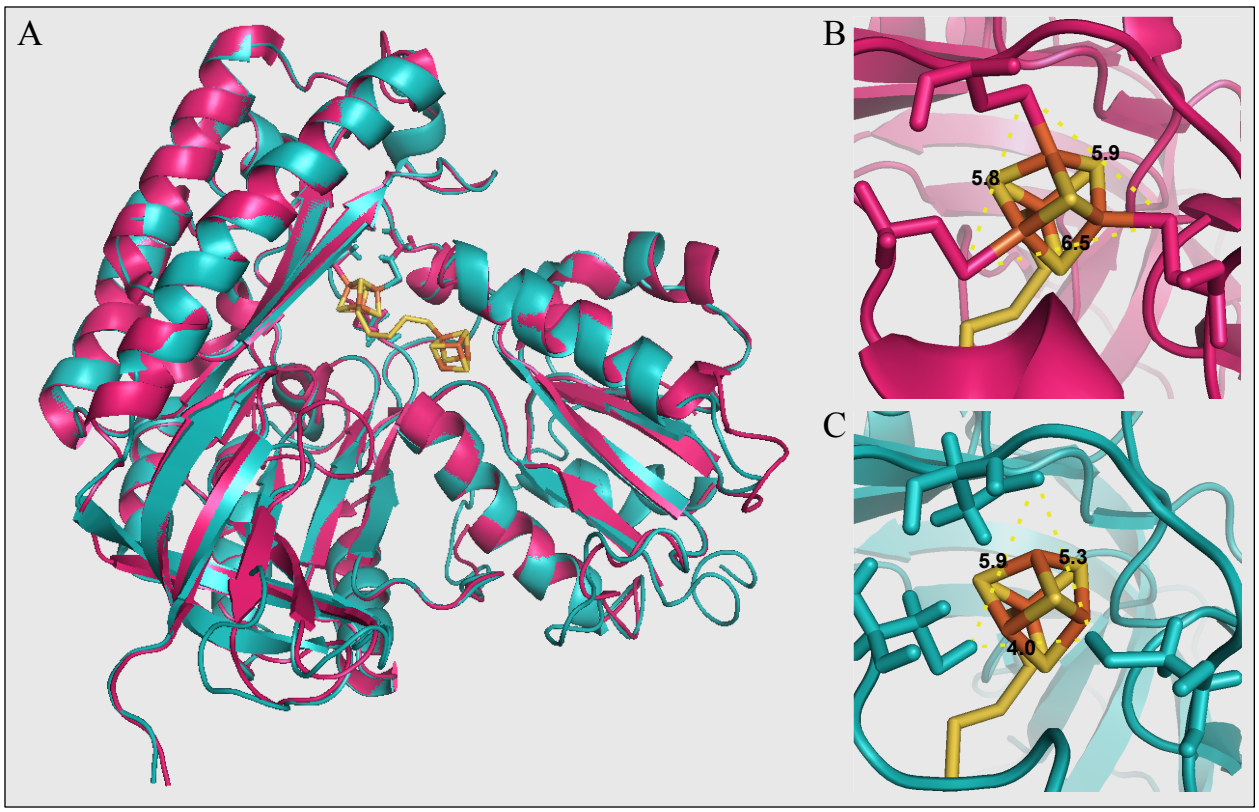
All of the 36 *G. barbadense* sequences that are likely to be radical SAM enzymes (Supplemental Table 1) share a good amount of sequence identity with radical SAM enzymes that have been well-characterized. It is interesting to note that many of them share sequence identity with radical SAM enzymes that are encoded in mitochondria or chloroplasts (Supplemental Table 1). There are also sequences that pass the recommended minimum 30% sequence identity with radical SAM enzymes with solved crystal structures, for homology modeling (Supplemental Table 2). Once these *Gossypium* sequences are examined further and found to share other signatures with the crystal structures they align with, homology modeling could be performed.

The phylogenetic comparison between the *G. barbadense, A. thaliana*, and other standard radical SAM sequences showed an interesting trend. The tree (Figure 10) demonstrates that all of the 36 *G. barbadense* sequences that are likely to be radical SAM enzymes share a relatively high amount of sequence similarity to the set of 16 *A. thaliana* sequences, rather than other standard radical SAM sequences. They are likely homologues. However, in most cases, there is more than one *Gossypium* sequence for every *Arabidopsis* sequence. It may be relevant that *G. barbadense* is a tetraploid species [109], whereas laboratory Arabidopsis thaliana is diploid [110]. It will be important to assess the status of the homology of each of the *Gossypium* sequences and determine whether or not they are artifacts of either the sequencing process or automatic annotations before further analysis is conducted. If they instead represent true *G. barbadense* biology, it could be important to identify if they are the products of polyploidy or gene duplications, and determine what type of homology the sequences share. The results of the four sequences further analyzed in a BLAST alignment (Table 7) likely tell an interesting evolutionary story, because the sequence of A0A5J5PDX9 is so different from the neighboring *Gossypium* sequences. The sequence itself is similar to a CK5P1_RAT mitochondrial tRNA methylthiotransferase (Table 5). Further analysis will need to be performed to determine whether the sequences are paralogous or simply highly divergent.

2.6 Supplemental Figures and Tables



Supplemental Figure 1. Phylogenetic tree of the 100 randomly-selected eukaryotic organisms. Counts of radical SAM HMM hits are blue, percentages of radical SAM HMM hits in each proteome is in red.



Supplemental Figure 2. Homology modelling structural overlays. (A) Structural overlay of homology model of A0A5K1U8H1_ENTHI from *Entamoeba histolytica* (teal), and RimO (PDB ID: 4JC0) template with cofactors (magenta). (B) 4JC0 template and iron-sulfur cluster. (C) A0A5K1U8H1_ENTHI homology model overlaid with iron-sulfur cluster from 4JC0 template. Distances between cysteine Sγ atoms of the "nest" and the iron atoms are measured in Å in panels (B) and (C).

3 Conclusions

3.1 Summary

The main purpose of this work was to identify candidate radical SAM enzymes for biochemical characterization, and Hidden Markov Models allowed us to select the sequences. By establishing a computational workflow, our investigation narrowed the full sets of proteomes down to those sequences that contain known radical SAM signatures. We were also able to analyze trends across coarse phylogenetic groups and organisms with different oxygen tolerances. We were able to further identify putative radical SAM enzymes from the proteomes of microaerophilic parasite *Entamoeba histolytica* and cotton species *Gossypium barbadense* by comparing these sequences to those of well-characterized radical SAM enzymes. The *Entamoeba histolytica* sequence A0A5K1U8H1_ENTHI did not share enough sequence identity with any of the radical SAM enzymes with solved crystal structures to yield a good homology model, but several *Gosspyium barbadense* sequences share a reasonable amount of sequence identity to be used in future computational biochemistry studies. Possible biochemical reactions of the *Entamoeba* and *Gossypium* sequences have been identified, and can be further tested and characterized in laboratory experiments.

3.2 Conclusions and Future Directions

We have curated a set of proteins that are likely to function as radical SAM enzymes and have yet to be fully characterized. Many of the sequences contained in the UniProt database are based on automatic annotations of metagenomics data, so the next step in any analysis should be establishing the true sequence of the DNA and/or protein of interest. One of the next steps is to investigate the *Entamoeba histolytica* sequences that were selected with Hidden Markov Models, which do not have high similarity to radical SAM enzymes with solved crystal structures but are sufficiently similar to radical SAMs that have been experimentally characterized in other ways.

The *Gossypium barbadense* sequences selected by Hidden Markov Models can also be further characterized. Their BLAST similarities to well-characterized radical SAM sequences are encouraging, and some of the sequences share a degree of similarity with radical SAMs that have solved crystal structures that would be suitable for computational biochemistry techniques such as homology modeling, molecular dynamics simulations, and docking.

To study radical SAM enzymes in plants at a larger scale, larger phylogenetic trees or Sequence Similarity Networks could be composed from the HMM-matching sequences of more plant species, and then examined to see if there are any outliers that might be involved in pathways particular to plant physiology. The relationship between sequences like *Gosspyium barbadense*'s A0A5J5PDX9 and its homologues should also be further investigated to see whether or not they might be involved in novel pathways.

As for evolutionary trends in the varying amounts of radical SAM enzymes due to oxygen tolerances of the organisms analyzed, some expected trends were confirmed, while other results were more surprising. We saw a trend of fewer radical SAM enzymes in prokaryotes that were anaerobic compared to those that were aerobic. The trends were not consistent for oxygen levels in eukaryotic organisms; this either reveals a biological quirk in eukaryotes, or is an artifact of either the sampling methods used, eukaryotic ploidy, or the UniProt method of annotating sequences.

4 References

1. Pettersen, E.F.; Goddard, T.D.; Huang, C.C.; Couch, G.S.; Greenblatt, D.M.; Meng, E.C.; Ferrin, T.E. UCSF Chimera?A visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004**, *25*, 1605–1612, doi:10.1002/jcc.20084.
2. Wang, S.C.; Frey, P.A. S-adenosylmethionine as an oxidant: the radical SAM superfamily. *Trends Biochem. Sci.* **2007**, *32*, 101–110, doi:10.1016/j.tibs.2007.01.002.
3. Berteau, O.; Benjdia, A. DNA Repair by the Radical SAM Enzyme Spore Photoproduct Lyase: From Biochemistry to Structural Investigations. *Photochem. Photobiol.* **2017**, *93*, 67–77, doi:10.1111/php.12702.
4. Chirpich, T.P.; Zappia, V.; Costilow, R.N.; Barker, H.A. Lysine 2,3-Aminomutase: PURIFICATION AND PROPERTIES OF A PYRIDOXAL PHOSPHATE AND S-ADENOSYLMETHIONINE-ACTIVATED ENZYME. *J. Biol. Chem.* **1970**, *245*, 1778–1789.
5. Martin, W.F. Carbon–Metal Bonds: Rare and Primordial in Metabolism. *Trends Biochem. Sci.* **2019**, *44*, 807–818, doi:10.1016/j.tibs.2019.04.010.
6. Broderick, J.B.; Duffus, B.R.; Duschene, K.S.; Shepard, E.M. Radical *S* -Adenosylmethionine Enzymes. *Chem. Rev.* **2014**, *114*, 4229–4317, doi:10.1021/cr4004709.
7. Sofia, H.J. Radical SAM, a novel protein superfamily linking unresolved steps in familiar biosynthetic pathways with radical mechanisms: functional characterization using new analysis and information visualization methods. *Nucleic Acids Res.* **2001**, *29*, 1097–1106, doi:10.1093/nar/29.5.1097.
8. Imlay, J.A.; Sethu, R.; Rohaun, S.K. Evolutionary adaptations that enable enzymes to tolerate oxidative stress. *Free Radic. Biol. Med.* **2019**, *140*, 4–13, doi:10.1016/j.freeradbiomed.2019.01.048.
9. Holliday, G.L.; Akiva, E.; Meng, E.C.; Brown, S.D.; Calhoun, S.; Pieper, U.; Sali, A.; Booker, S.J.; Babbitt, P.C. Atlas of the Radical SAM Superfamily: Divergent Evolution of Function Using a “Plug and Play” Domain. In *Methods in Enzymology*; Elsevier, 2018; Vol. 606, pp. 1–71 ISBN 978-0-12-812794-0.
10. Challand, M.R.; Driesener, R.C.; Roach, P.L. Radical S-adenosylmethionine enzymes: Mechanism, control and function. *Nat. Prod. Rep.* **2011**, *28*, 1696, doi:10.1039/c1np00036e.
11. Benjdia, A.; Balty, C.; Berteau, O. Radical SAM Enzymes in the Biosynthesis of Ribosomally Synthesized and Post-translationally Modified Peptides (RiPPs). *Front. Chem.* **2017**, *5*, 87, doi:10.3389/fchem.2017.00087.
12. Chandor, A.; Berteau, O.; Douki, T.; Gasparutto, D.; Sanakis, Y.; Ollagnier-de-Choudens, S.; Atta, M.; Fontecave, M. Dinucleotide Spore Photoproduct, a Minimal Substrate of the DNA Repair Spore Photoproduct Lyase Enzyme from *Bacillus subtilis*. *J. Biol. Chem.* **2006**, *281*, 26922–26931, doi:10.1074/jbc.M602297200.
13. Layer, G.; Heinz, D.W.; Jahn, D.; Schubert, W.-D. Structure and function of radical SAM enzymes. *Curr. Opin. Chem. Biol.* **2004**, *8*, 468–476, doi:10.1016/j.cbpa.2004.08.001.
14. Utter, M. F.; Lipmann, F.; Werkman, C. H. J. Biol. Chem. 1945, 158, 521.
15. Ifuku, O.; Kishimoto, J.; Haze, S.; Yanagi, M.; Fukushima, S. Biosci., Biotechnol., Biochem. 1992, 56, 1780.
16. Hanzelmann, P.; Schindelin, H. Crystal structure of the S-adenosylmethionine-dependent enzyme MoaA and its implications for molybdenum cofactor deficiency in humans. *Proc. Natl. Acad. Sci.* **2004**, *101*, 12870–12875, doi:10.1073/pnas.0404624101.

17. Giessing, A.M.B.; Jensen, S.S.; Rasmussen, A.; Hansen, L.H.; Gondela, A.; Long, K.; Vester, B.; Kirpekar, F. Identification of 8-methyladenosine as the modification catalyzed by the radical SAM methyltransferase Cfr that confers antibiotic resistance in bacteria. *RNA* **2009**, *15*, 327–336, doi:10.1261/rna.1371409.
18. Lee, K.-H.; Saleh, L.; Anton, B.P.; Madinger, C.L.; Benner, J.S.; Iwig, D.F.; Roberts, R.J.; Krebs, C.; Booker, S.J. Characterization of RimO, a New Member of the Methylthiotransferase Subclass of the Radical SAM Superfamily. *Biochemistry* **2009**, *48*, 10162–10174, doi:10.1021/bi900939w.
19. Yokoyama, K.; Numakura, M.; Kudo, F.; Ohmori, D.; Eguchi, T. Characterization and Mechanistic Study of a Radical SAM Dehydrogenase in the Biosynthesis of Butirosin. *J. Am. Chem. Soc.* **2007**, *129*, 15147–15155, doi:10.1021/ja072481t.
20. Heinemann, I.U.; Jahn, M.; Jahn, D. The biochemistry of heme biosynthesis. *Arch. Biochem. Biophys.* **2008**, *474*, 238–251, doi:10.1016/j.abb.2008.02.015.
21. Allen, R.M.; Chatterjee, R.; Ludden, P.W.; Shah, V.K. Incorporation of Iron and Sulfur from NifB Cofactor into the Iron-Molybdenum Cofactor of Dinitrogenase. *J. Biol. Chem.* **1995**, *270*, 26890–26896, doi:10.1074/jbc.270.45.26890.
22. Landgraf, B.J.; McCarthy, E.L.; Booker, S.J. Radical *S* -Adenosylmethionine Enzymes in Human Health and Disease. *Annu. Rev. Biochem.* **2016**, *85*, 485–514, doi:10.1146/annurev-biochem-060713-035504.
23. Holliday, G.L.; Thornton, J.M.; Marquet, A.; Smith, A.G.; Rébeillé, F.; Mendel, R.; Schubert, H.L.; Lawrence, A.D.; Warren, M.J. Evolution of enzymes and pathways for the biosynthesis of cofactors. *Nat. Prod. Rep.* **2007**, *24*, 972, doi:10.1039/b703107f.
24. Liu, Y.; Beer, L.L.; Whitman, W.B. Methanogens: a window into ancient sulfur metabolism. *Trends Microbiol.* **2012**, *20*, 251–258, doi:10.1016/j.tim.2012.02.002.
25. Wuebbles, D. Atmospheric methane and global change. *Earth-Sci. Rev.* **2002**, *57*, 177–210, doi:10.1016/S0012-8252(01)00062-9.
26. Bridgham, S.D.; Cadillo-Quiroz, H.; Keller, J.K.; Zhuang, Q. Methane emissions from wetlands: biogeochemical, microbial, and modeling perspectives from local to global scales. *Glob. Change Biol.* **2013**, *19*, 1325–1346, doi:10.1111/gcb.12131.
27. Preiner, M.; Xavier, J.; Sousa, F.; Zimorski, V.; Neubeck, A.; Lang, S.; Greenwell, H.; Kleinermanns, K.; Tüysüz, H.; McCollom, T.; et al. Serpentinization: Connecting Geochemistry, Ancient Metabolism and Industrial Hydrogenation. *Life* **2018**, *8*, 41, doi:10.3390/life8040041.
28. Jäger, C.M.; Croft, A.K. Anaerobic Radical Enzymes for Biotechnology. *ChemBioEng Rev.* **2018**, *5*, 143–162, doi:10.1002/cben.201800003.
29. Chen, K.; Zhang, P.; Li, H. Direct Amination of Unreactive C-H Bonds Catalyzed by N-hydroxyphthalimide. *Postdoc J.* **2013**, doi:10.14304/SURYA.JPR.V1N8.6.
30. Maugh, T.H. Activating Unreactive C--H Bonds. *Science* **1983**, *220*, 1261–1263, doi:10.1126/science.220.4603.1261.
31. Babu, K.R.; Zhu, N.; Bao, H. Iron-Catalyzed C–H Alkylation of Heterocyclic C–H Bonds. *Org. Lett.* **2017**, *19*, 46–49, doi:10.1021/acs.orglett.6b03287.
32. Labinger, J.A.; Bercaw, J.E. Understanding and exploiting C–H bond activation. *Nature* **2002**, *417*, 507–514, doi:10.1038/417507a.
33. Wächtershäuser, G. Groundworks for an evolutionary biochemistry: The iron-sulphur world. *Prog. Biophys. Mol. Biol.* **1992**, *58*, 85–201, doi:10.1016/0079-6107(92)90022-X.
34. Ross, D.S. A Quantitative Evaluation of the Iron-Sulfur World and Its Relevance to Life's

Origins. *Astrobiology* **2008**, *8*, 267–272, doi:10.1089/ast.2007.0199.
35. Jelen, B.I.; Giovannelli, D.; Falkowski, P.G. The Role of Microbial Electron Transfer in the Coevolution of the Biosphere and Geosphere. *Annu. Rev. Microbiol.* **2016**, *70*, 45–62, doi:10.1146/annurev-micro-102215-095521.
36. Sahai, N.; Kaddour, H.; Dalai, P. The Transition from Geochemistry to Biogeochemistry. *Elements* **2016**, *12*, 389–394, doi:10.2113/gselements.12.6.389.
37. Stirling, A.; Rozgonyi, T.; Krack, M.; Bernasconi, M. Prebiotic NH_3 Formation: Insights from Simulations. *Inorg. Chem.* **2016**, *55*, 1934–1939, doi:10.1021/acs.inorgchem.5b02911.
38. Thiel, J.; Byrne, J.M.; Kappler, A.; Schink, B.; Pester, M. Pyrite formation from FeS and H_2S is mediated through microbial redox activity. *Proc. Natl. Acad. Sci.* **2019**, *116*, 6897–6902, doi:10.1073/pnas.1814412116.
39. Vallee, B.L.; Williams, R.J. Metalloenzymes: the entatic nature of their active sites. *Proc. Natl. Acad. Sci.* **1968**, *59*, 498–505, doi:10.1073/pnas.59.2.498.
40. Williams, R.J.P. Energised (entatic) States of Groups and of Secondary Structures in Proteins and Metalloproteins. *Eur. J. Biochem.* **1995**, *234*, 363–381, doi:10.1111/j.1432-1033.1995.363_b.x.
41. Suzuki, T.; Yano, T.; Hara, M.; Ebisuzaki, T. Cysteine and cystine adsorption on FeS2(100). *Surf. Sci.* **2018**, *674*, 6–12, doi:10.1016/j.susc.2018.03.011.
42. Hanscam, R.; Shepard, E.M.; Broderick, J.B.; Copié, V.; Szilagyi, R.K. Secondary structure analysis of peptides with relevance to iron–sulfur cluster nesting. *J. Comput. Chem.* **2019**, *40*, 515–526, doi:10.1002/jcc.25741.
43. Imlay, J.A. Iron-sulphur clusters and the problem with oxygen. *Mol. Microbiol.* **2006**, *59*, 1073–1082, doi:10.1111/j.1365-2958.2006.05028.x.
44. Meyer, J. Iron–sulfur protein folds, iron–sulfur chemistry, and evolution. *JBIC J. Biol. Inorg. Chem.* **2008**, *13*, 157–170, doi:10.1007/s00775-007-0318-7.
45. Bruska, M.K.; Stiebritz, M.T.; Reiher, M. Analysis of differences in oxygen sensitivity of Fe–S clusters. *Dalton Trans.* **2013**, *42*, 8729, doi:10.1039/c3dt50763g.
46. Zerkle, A.L. Biogeochemical signatures through time as inferred from whole microbial genomes. *Am. J. Sci.* **2005**, *305*, 467–502, doi:10.2475/ajs.305.6-8.467.
47. Kendall, J.J.; Barrero-Tobon, A.M.; Hendrixson, D.R.; Kelly, D.J. Hemerythrins in the microaerophilic bacterium *C ampylobacter jejuni* help protect key iron-sulphur cluster enzymes from oxidative damage: Microaerophily in *C. jejuni*. *Environ. Microbiol.* **2014**, *16*, 1105–1121, doi:10.1111/1462-2920.12341.
48. Vieira-Silva, S.; Rocha, E.P.C. An Assessment of the Impacts of Molecular Oxygen on the Evolution of Proteomes. *Mol. Biol. Evol.* **2008**, *25*, 1931–1942, doi:10.1093/molbev/msn142.
49. Pierella Karlusich, J.J.; Ceccoli, R.D.; Graña, M.; Romero, H.; Carrillo, N. Environmental Selection Pressures Related to Iron Utilization Are Involved in the Loss of the Flavodoxin Gene from the Plant Genome. *Genome Biol. Evol.* **2015**, *7*, 750–767, doi:10.1093/gbe/evv031.
50. Loesche, W.J. Oxygen Sensitivity of Various Anaerobic Bacteria. *Appl. Microbiol.* **1969**, *18*, 723.
51. Tally, F.P.; Stewart, P.R.; Sutter, V.L.; Rosenblatt, J.E. Oxygen tolerance of fresh clinical anaerobic bacteria. *J. Clin. Microbiol.* **1975**, *1*, 161.
52. Imlay, J.A. How oxygen damages microbes: Oxygen tolerance and obligate anaerobiosis. In *Advances in Microbial Physiology*; Elsevier, 2002; Vol. 46, pp. 111–153 ISBN 978-0-12-

027746-9.
53. Samuelson, L.J.; Teskey, R.O. Net photosynthesis and leaf conductance of loblolly pine seedlings in 2 and 21% oxygen as influenced by irradiance, temperature and provenance. *Tree Physiol.* **1991**, *8*, 205–211, doi:10.1093/treephys/8.2.205.
54. Freeman, S. *Biological science*; 4th ed.; Benjamin Cummings: Boston, 2011; ISBN 978-0-321-59820-2.
55. Martin, W.F.; Garg, S.; Zimorski, V. Endosymbiotic theories for eukaryote origin. *Philos. Trans. R. Soc. B Biol. Sci.* **2015**, *370*, 20140330, doi:10.1098/rstb.2014.0330.
56. *Anoxia: evidence for eukaryote survival and paleontological strategies*; Altenbach, A.V., Bernhard, J.M., Seckbach, J., Eds.; Cellular origin, life in extreme habitats and astrobiology; Springer: Dordrecht, 2012; ISBN 978-94-007-1895-1.
57. Lill, R. Function and biogenesis of iron–sulphur proteins. *Nature* **2009**, *460*, 831–838, doi:10.1038/nature08301.
58. Field, J.; Rosenthal, B.; Samuelson, J. Early lateral transfer of genes encoding malic enzyme, acetyl-CoA synthetase and alcohol dehydrogenases from anaerobic prokaryotes to Entamoeba histolytica. *Mol. Microbiol.* **2000**, *38*, 446–455, doi:10.1046/j.1365-2958.2000.02143.x.
59. Nixon, J.E.J.; Wang, A.; Field, J.; Morrison, H.G.; McArthur, A.G.; Sogin, M.L.; Loftus, B.J.; Samuelson, J. Evidence for Lateral Transfer of Genes Encoding Ferredoxins, Nitroreductases, NADH Oxidase, and Alcohol Dehydrogenase 3 from Anaerobic Prokaryotes to Giardialamblia and Entamoebahistolytica. *Eukaryot. Cell* **2002**, *1*, 181–190, doi:10.1128/EC.1.2.181-190.2002.
60. Keeling, P.J. Diversity and evolutionary history of plastids and their hosts. *Am. J. Bot.* **2004**, *91*, 1481–1493, doi:10.3732/ajb.91.10.1481.
61. Sagan, L. On the origin of mitosing cells. *J. Theor. Biol.* **1967**, *14*, 225-IN6, doi:10.1016/0022-5193(67)90079-3.
62. Järvi, S.; Suorsa, M.; Aro, E.-M. Photosystem II repair in plant chloroplasts — Regulation, assisting proteins and shared components with photosystem II biogenesis. *Biochim. Biophys. Acta BBA - Bioenerg.* **2015**, *1847*, 900–909, doi:10.1016/j.bbabio.2015.01.006.
63. Erb, T.J.; Zarzycki, J. A short history of RubisCO: the rise and fall (?) of Nature's predominant CO2 fixing enzyme. *Curr. Opin. Biotechnol.* **2018**, *49*, 100–107, doi:10.1016/j.copbio.2017.07.017.
64. Raven, P.H.; Evert, R.F.; Eichhorn, S.E. *Biology of plants*; Eighth edition.; W.H. Freeman and Company Publishers: New York, 2013; ISBN 978-1-4292-1961-7.
65. Poux, S.; Arighi, C.N.; Magrane, M.; Bateman, A.; Wei, C.-H.; Lu, Z.; Boutet, E.; Bye-A-Jee, H.; Famiglietti, M.L.; Roechert, B.; et al. On expert curation and scalability: UniProtKB/Swiss-Prot as a case study. *Bioinformatics* **2017**, *33*, 3454–3460, doi:10.1093/bioinformatics/btx439.
66. Apweiler, R. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.* **2004**, *32*, 115D – 119, doi:10.1093/nar/gkh131.
67. The UniProt Consortium UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **2019**, *47*, D506–D515, doi:10.1093/nar/gky1049.
68. Mitchell, A.L.; Attwood, T.K.; Babbitt, P.C.; Blum, M.; Bork, P.; Bridge, A.; Brown, S.D.; Chang, H.-Y.; El-Gebali, S.; Fraser, M.I.; et al. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res.* **2019**, *47*, D351–D360, doi:10.1093/nar/gky1100.

69. Sillitoe, I.; Lewis, T.E.; Cuff, A.; Das, S.; Ashford, P.; Dawson, N.L.; Furnham, N.; Laskowski, R.A.; Lee, D.; Lees, J.G.; et al. CATH: comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Res.* **2015**, *43*, D376–D381, doi:10.1093/nar/gku947.
70. El-Gebali, S.; Mistry, J.; Bateman, A.; Eddy, S.R.; Luciani, A.; Potter, S.C.; Qureshi, M.; Richardson, L.J.; Salazar, G.A.; Smart, A.; et al. The Pfam protein families database in 2019. *Nucleic Acids Res.* **2019**, *47*, D427–D432, doi:10.1093/nar/gky995.
71. Cristianini, N.; Hahn, M.W. *Introduction to computational genomics: a case studies approach*; Cambridge University Press: Cambridge, UK; New York, 2007; ISBN 978-0-511-64888-5.
72. Felsenstein, J.; Churchill, G.A. A Hidden Markov Model approach to variation among sites in rate of evolution. *Mol. Biol. Evol.* **1996**, *13*, 93–104, doi:10.1093/oxfordjournals.molbev.a025575.
73. Eddy, S.R.; Mitchison, G.; Durbin, R. Maximum Discrimination Hidden Markov Models of Sequence Consensus. *J. Comput. Biol.* **1995**, *2*, 9–23, doi:10.1089/cmb.1995.2.9.
74. Schnoes, A.M.; Brown, S.D.; Dodevski, I.; Babbitt, P.C. Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies. *PLoS Comput. Biol.* **2009**, *5*, e1000605, doi:10.1371/journal.pcbi.1000605.
75. Gilks, W.R.; Audit, B.; De Angelis, D.; Tsoka, S.; Ouzounis, C.A. Modeling the percolation of annotation errors in a database of protein sequences. *Bioinformatics* **2002**, *18*, 1641–1649, doi:10.1093/bioinformatics/18.12.1641.
76. Gilks, W.R.; Audit, B.; de Angelis, D.; Tsoka, S.; Ouzounis, C.A. Percolation of annotation errors through hierarchically structured protein sequence databases. *Math. Biosci.* **2005**, *193*, 223–234, doi:10.1016/j.mbs.2004.08.001.
77. Jones, C.E.; Brown, A.L.; Baumann, U. Estimating the annotation error rate of curated GO database sequence annotations. *BMC Bioinformatics* **2007**, *8*, 170, doi:10.1186/1471-2105-8-170.
78. Kyrpides, N.C.; Ouzounis, C.A. Whole-genome sequence annotation: "Going wrong with confidence." *Mol. Microbiol.* **1999**, *32*, 886–887, doi:10.1046/j.1365-2958.1999.01380.x.
79. Pallen, M.; Wren, B.; Parkhill, J. "Going wrong with confidence": misleading sequence analyses of CiaB and ClpX. *Mol. Microbiol.* **1999**, *34*, 195–195, doi:10.1046/j.1365-2958.1999.01561.x.
80. Radivojac, P.; Clark, W.T.; Oron, T.R.; Schnoes, A.M.; Wittkop, T.; Sokolov, A.; Graim, K.; Funk, C.; Verspoor, K.; Ben-Hur, A.; et al. A large-scale evaluation of computational protein function prediction. *Nat. Methods* **2013**, *10*, 221–227, doi:10.1038/nmeth.2340.
81. Makarova, K.S.; Wolf, Y.I.; Koonin, E.V. Towards functional characterization of archaeal genomic dark matter. *Biochem. Soc. Trans.* **2019**, *47*, 389–398, doi:10.1042/BST20180560.
82. Hanson, A.D.; Pribat, A.; Waller, J.C.; Crécy-Lagard, V. de 'Unknown' proteins and 'orphan' enzymes: the missing half of the engineering parts list – and how to find it. *Biochem. J.* **2010**, *425*, 1–11, doi:10.1042/BJ20091328.
83. Ellens, K.W.; Christian, N.; Singh, C.; Satagopam, V.P.; May, P.; Linster, C.L. Confronting the catalytic dark matter encoded by sequenced genomes. *Nucleic Acids Res.* **2017**, *45*, 11495–11514, doi:10.1093/nar/gkx937.
84. Niehaus, T.D.; Thamm, A.M.; de Crécy-Lagard, V.; Hanson, A.D. Proteins of unknown biochemical function - A persistent problem and a roadmap to help overcome it. *Plant Physiol.* **2015**, pp.00959.2015, doi:10.1104/pp.15.00959.

85. Michalkova, A.; Kholod, Y.; Kosenkov, D.; Gorb, L.; Leszczynski, J. Viability of pyrite pulled metabolism in the 'iron-sulfur world' theory: Quantum chemical assessment. *Geochim. Cosmochim. Acta* **2011**, *75*, 1933–1941, doi:10.1016/j.gca.2011.01.015.
86. Simão, F.A.; Waterhouse, R.M.; Ioannidis, P.; Kriventseva, E.V.; Zdobnov, E.M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **2015**, *31*, 3210–3212, doi:10.1093/bioinformatics/btv351.
87. Reimer, L.C.; Vetcininova, A.; Carbasse, J.S.; Söhngen, C.; Gleim, D.; Ebeling, C.; Overmann, J. Bac *Dive* in 2019: bacterial phenotypic data for High-throughput biodiversity analysis. *Nucleic Acids Res.* **2019**, *47*, D631–D636, doi:10.1093/nar/gky879.
88. Cokelaer, T.; Pultz, D.; Harder, L.M.; Serra-Musach, J.; Saez-Rodriguez, J. BioServices: a common Python package to access biological Web Services programmatically. *Bioinformatics* **2013**, *29*, 3241–3242, doi:10.1093/bioinformatics/btt547.
89. Cock, P.J.A.; Antao, T.; Chang, J.T.; Chapman, B.A.; Cox, C.J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **2009**, *25*, 1422–1423, doi:10.1093/bioinformatics/btp163.
90. SciPy 1.0 Contributors; Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **2020**, *17*, 261–272, doi:10.1038/s41592-019-0686-2.
91. Aylward, F. Introduction to protein annotation with Hidden Markov Models v1 (protocols.io.pijdkcn).
92. Loftus, B.; Anderson, I.; Davies, R.; Alsmark, U.C.M.; Samuelson, J.; Amedeo, P.; Roncaglia, P.; Berriman, M.; Hirt, R.P.; Mann, B.J.; et al. The genome of the protist parasite Entamoeba histolytica. *Nature* **2005**, *433*, 865–868, doi:10.1038/nature03291.
93. Forouhar, F.; Arragain, S.; Atta, M.; Gambarelli, S.; Mouesca, J.-M.; Hussain, M.; Xiao, R.; Kieffer-Jaquinod, S.; Seetharaman, J.; Acton, T.B.; et al. Two Fe-S clusters catalyze sulfur insertion by radical-SAM methylthiotransferases. *Nat. Chem. Biol.* **2013**, *9*, 333–338, doi:10.1038/nchembio.1229.
94. Yang, J.; Zhang, Y. I-TASSER server: new development for protein structure and function predictions. *Nucleic Acids Res.* **2015**, *43*, W174–W181, doi:10.1093/nar/gkv342.
95. Maestro, S., LLC. (2019) Schrodinger Release 2019-2, New York, NY.
96. Guex, N.; Peitsch, M.C.; Schwede, T. Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: A historical perspective. *ELECTROPHORESIS* **2009**, *30*, S162–S173, doi:10.1002/elps.200900140.
97. The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.
98. Madeira, F.; Park, Y. mi; Lee, J.; Buso, N.; Gur, T.; Madhusoodanan, N.; Basutkar, P.; Tivey, A.R.N.; Potter, S.C.; Finn, R.D.; et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.* **2019**, *47*, W636–W641, doi:10.1093/nar/gkz268.
99. Letunic, I.; Bork, P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res.* **2011**, *39*, W475–W478, doi:10.1093/nar/gkr201.
100. Benkert, P.; Biasini, M.; Schwede, T. Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics* **2011**, *27*, 343–350, doi:10.1093/bioinformatics/btq662.
101. Cheng, F.; Wu, J.; Cai, X.; Liang, J.; Freeling, M.; Wang, X. Gene retention, fractionation

and subgenome differences in polyploid plants. *Nat. Plants* **2018**, *4*, 258–268, doi:10.1038/s41477-018-0136-7.
102. Clark, J.W.; Donoghue, P.C.J. Whole-Genome Duplication and Plant Macroevolution. *Trends Plant Sci.* **2018**, *23*, 933–945, doi:10.1016/j.tplants.2018.07.006.
103. Knappe, J.; Schmitt, T. A novel reaction of S-adenosyl-L-methionine correlated with the activation of pyruvate formate-lyase. *Biochem. Biophys. Res. Commun.* **1976**, *71*, 1110–1117, doi:10.1016/0006-291X(76)90768-3.
104. Layer, G.; Verfürth, K.; Mahlitz, E.; Jahn, D. Oxygen-independent Coproporphyrinogen-III Oxidase HemN from *Escherichia coli*. *J. Biol. Chem.* **2002**, *277*, 34136–34142, doi:10.1074/jbc.M205247200.
105. Li, B.; Bridwell-Rabb, J. Aerobic Enzymes and Their Radical SAM Enzyme Counterparts in Tetrapyrrole Pathways. *Biochemistry* **2019**, *58*, 85–93, doi:10.1021/acs.biochem.8b00906.
106. Ali, V.; Nozaki, T. Iron–Sulphur Clusters, Their Biosynthesis, and Biological Functions in Protozoan Parasites. In *Advances in Parasitology*; Elsevier, 2013; Vol. 83, pp. 1–92 ISBN 978-0-12-407705-8.
107. Sharma, A.; Sharma, D.; Verma, S.K. Proteome wide identification of iron binding proteins of Xanthomonas translucens pv. undulosa: focus on secretory virulent proteins. *BioMetals* **2017**, *30*, 127–141, doi:10.1007/s10534-017-9991-3.
108. Berman-Frank, I. Segregation of Nitrogen Fixation and Oxygenic Photosynthesis in the Marine Cyanobacterium Trichodesmium. *Science* **2001**, *294*, 1534–1537, doi:10.1126/science.1064082.
109. Paterson, A.H.; Wendel, J.F.; Gundlach, H.; Guo, H.; Jenkins, J.; Jin, D.; Llewellyn, D.; Showmaker, K.C.; Shu, S.; Udall, J.; et al. Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. *Nature* **2012**, *492*, 423–427, doi:10.1038/nature11798.
110. Ravi, M.; Marimuthu, M.P.A.; Tan, E.H.; Maheshwari, S.; Henry, I.M.; Marin-Rodriguez, B.; Urtecho, G.; Tan, J.; Thornhill, K.; Zhu, F.; et al. A haploid genetics toolbox for Arabidopsis thaliana. *Nat. Commun.* **2014**, *5*, 5334, doi:10.1038/ncomms6334.