9-1-2015

# Mining Public Databases for Discovery of Structure and Function within the Hotdog-fold Thioesterase and HAD Phosphatase Enzyme Families

Sarah Toews Keating

Sarah Toews Keating

*Candidate*

Chemistry and Chemical Biology

*Department*

This dissertation is approved, and it is acceptable in quality and form for publication:

*Approved by the Dissertation Committee:*

Debra Dunaway-Mariano , Chairperson

Patrick S. Mariano

Fu-Sen Liang

Karen N. Allen

# MINING PUBLIC DATABASES FOR DISCOVERY OF STRUCTURE AND FUNCTION WITHIN THE HOTDOG-FOLD THIOESTERASE AND HAD PHOSPHATASE ENZYME SUPERFAMILIES

**by**

**SARAH TOEWS KEATING**

B.A., Chemistry, Carleton College, 2009

DISSERTATION

Submitted in Partial Fulfillment of the
Requirements for the Degree of

**Doctor of Philosophy**
**Chemistry**

The University of New Mexico
Albuquerque, New Mexico

**July, 2015**

## ACKNOWLEDGEMENTS

This dissertation is the product of years of hard work shaped by the incredible guidance of my advisor, Dr. Debra Dunaway-Mariano, for whom I would like to express my sincere admiration and gratitude. I am thankful for both her insight and daunting breadth of knowledge but also for her enthusiasm and encouragement over all these years. I am grateful for everything I have learned from her and how she has inspired me, especially as a woman in science.

I also thank the members of my dissertation committee, Dr. Patrick Mariano, Dr. Fu-Sen Liang, and Dr. Karen Allen, for their patience and input.

I owe thanks and friendship to the members of DDM’s group, past and present, from whom I have learned so much. I’d particularly like to thank Luke, John, and Ty, members of my first year cohort and lab-mates with whom I have laughed, learned, grown, struggled, and prevailed. Thanks also to Amy and Kaila for suffering through the tedium of manual bioinformatics with me. You aren’t as fast as Python but you are more fun!

In Karen Allen’s lab, thank you to Shasha, for running the hotdog HTS screens, and Chetanya and Jeremiah, for helping me get my bearings with similarity networks and Cytoscape.

Though not directly part of my doctoral research, teaching has been an integral and fulfilling part of my graduate school experience. For that, I’d like to thank the community of individuals and instructors involved with the General Chemistry sequence here at UNM. I have learned so much about myself, education, learning, and community from that amazing group. I am also grateful

for the support and faith I received from the department and Dr. Dunaway-Mariano as I discovered my love for and cultivated my experience in chemical education.

This journey would not have been possible without the love of my family. Mom (though you were my first female role model, it has taken me years to appreciate that you are, in fact, always right) and Dad (whose bedtime stories about moles, ohms, and volts surely got me started in the sciences): I am ever grateful for your unfailing support, love, encouragement, wisdom, and counsel through all my struggles and successes. Ben: thanks for being an awesome brother but also for all of your encouragement, advice, and help in the many ways I can make Python solve all my problems.

Finally, Bob: my wonderful husband, confidant, cheerleader, support, challenger, editor, beta-tester, solace, safe place, teammate, player two, spider killer, and dog walker but above all else, best friend. You make me a happier, healthier person, a better scientist, and bring so much light and laughter into my life—thank you.

# MINING PUBLIC DATABASES FOR DISCOVERY OF STRUCTURE AND FUNCTION WITHIN THE HOTDOG-FOLD THIOESTERASE AND HAD PHOSPHATASE ENZYME FAMILIES

**SARAH TOEWS KEATING**

**B.A., Chemistry, Carleton College, 2009**

**Ph.D., Chemistry, University of New Mexico, 2015**

## ABSTRACT

For my doctoral work, I have developed strategies to mine public databases for data that can be used to infer structural and functional information for the hotdog-fold and HADSF superfamilies.

For the hotdog-fold superfamily, I used curated and automatically applied annotations of structure, taxonomic lineage, function, and subfamily membership from the UniProtKB, gene context and taxonomic information from the NCBI, and the results of several in-depth explorations of subfamily/function and structural class membership. Based on the distribution of the aforementioned annotations mapped onto a sequence similarity network (SSN), I applied structural assignments to sequences and/or specific function/subfamily assignments to ~143,000 sequences and general subfamily assignments to an additional ~61,000 sequences. I also identified 52 clusters containing nearly 9,000 uncharacterized sequences lacking any annotations whatsoever and several probable instances of cross-domain gene transfer that would be of interest for further study.

Within the thioesterase family of the hotdog-fold superfamily, I identified ~450 targets to undergo high-throughput screens in Karen Allen's lab, the SSN-mapped results of which underscore widespread promiscuity across the family. I demonstrated the use of HTS and gene context results to infer functional identities for hotdog-fold superfamily members, though most gene contexts proved to be unilluminating.

In the HADSF, I explored the diversity and function space of Firmicutes members, revealing the wide range of HADSF representatives even within members of the same genus. SSNs mapped according to taxonomic lineage, subfamily membership, and function revealed several instances of probable gene transfer among Firmicutes members, but also across phyla. Related gene context, biological range, and HTS results revealed a member of *Listeria innocua* to be a member of the PTS pathway and provided potentially useful information for other HADSF members.

Two groups of HADSF members were earlier identified as having interesting evolutionary histories. I provide biological range- and gene context-based evidence for the convergent evolution of FMN phosphatase activity in *E. coli* and *Bacteroides thetaiotaomicron* HADSF members, divergent evolution of the same in *E. coli* and *Salmonella enterica* members, and divergent evolution of yidA in *E. coli* and BT3352 in *B. thetaiotaomicron*.

# CONTENTS

# LIST OF ABBREVIATIONS

| | |
|---|---|
| 3HCDH | 3-hydroxyacyl-CoA dehydrogenase |
| 4HBT | 4-hydroxybenzoyl thioesterase |
| aa | amino acids |
| ACOT | Acyl-Coenzyme A thioesterase |
| ACP | Acyl Carrier Protein |
| ADH | Short-chain dehydrogenase |
| ADP | Adenosine diphosphate |
| AMP | Adenosine monophosphate |
| Arg/R | Arginine |
| Asn/N | Asparagine |
| Asp/D | Aspartic acid |
| ATP | Adenosine triphosphate |
| *B. thetaiotaomicron* | *Bacteroides thetaiotaomicron* |
| BACH | brain acyl-Coenzyme A thioesterase |
| ß-PGM | ß-phosphoglucomutase |
| BFIT | Brown fat inducible thioesterase |
| BKAS | beta-keto-acyl synthase |
| BLAST | Basic Local Alignment Search Tool |
| CACH | Cytoplasmic acetyl-CoA hydrolase |
| CBS | Cystathionine beta synthase |
| CFA | Coronafacic acid |
| cNMP | Cyclic nucleotide-monophosphate |

| | |
|---|---|
| CoA | Coenzyme A |
| CoA-SH | Free thiol version of Coenzyme A |
| COBALT | Constraint-based Multiple Protein Alignment Tool (NCBI) |
| Cys/C | Cysteine |
| D | Dimer fold |
| DdhA | Double hotdog version of TA tetramer |
| dgoK | 2-oxo-3-deoxygalactonate kinase |
| dh | Double hotdog fold |
| DHNA-CoA | 1,4-dihydroxy-2-naphthoate-Coenzyme A |
| DNA | Deoxyribonucleic acid |
| DUF | Domain of Unknown Function |
| *E. coli* | *Escherichia coli* |
| EFI | Enzyme Function Initiative |
| E-value | Expectation value for a BLAST result |
| FAD | Flavin adenine dinucleotide |
| FAS | Fatty acid biosynthesis |
| FASTA | text-based format for representing protein sequence |
| FLK | Fluoroacetyl-CoA thioesterase |
| FMN | Flavin mononucleotide |
| Glu/E | Glutamate |
| Gly/G | Glycine |
| GNAT | Gcn5-related N-acetyltransferases |
| H1 | Hexaer with interfacial active site loops |

| | |
|---|---|
| H2 | Hexamer with interfacial N-terminal helices |
| H3 | Hexamer with head-to-tail arrangement |
| HAD | Haloacid Dehalogenase |
| HADSF | Haloacid Dehalogenase Superfamily |
| HBP | D,D-heptose 1,7-bisphosphate |
| HK-MTPenyl-1-P | 2-hydroxy-3-keto-5-methylthiopentenyl-1-phosphate |
| HMM | Hidden Markov model |
| HTS | High-Throughput Screening |
| IPRO | InterPro group number |
| IToL | Interactive Tree of Life |
| KDPG | 2-keto-3-deoxy-6-phosphogalactonate |
| Leu/L | Leucine |
| Lys/K | Lysine |
| Met/M | Methionine |
| MSA | Multiple sequence alignment |
| NAD | Nicotinamide adenine dinucleotide |
| NCBI | National Center for Biotechnology Information |
| PAA | Phenylacetic acid |
| PDB | Protein Data Bank |
| PFA | Polyunsaturated fatty acid |
| PHA | Polyhydroxyalkanoate |
| Phe/F | Phenylalanine |
| PKS | Polyketide biosynthesis |

| | |
|---|---|
| PTS | phophoenolpyruvate:carbohydrate phosphotransferases |
| RefSeq | NCBI Reference Sequence Database |
| ribF | riboflavin kinase/FAD synthetase |
| RNA | Ribonucleic acid |
| RNN | Representative node network |
| *S. enterica* | *Salmonella enterica* |
| SCP | Sterol carrier protein |
| Ser/S | Serine |
| SFLD | Structure Function Linkage Database |
| SI | Sequence identity |
| SNF | Structure, No Function (PDBs without assigned functions) |
| SSN | Sequence similarity network |
| START | StAR-related lipid-transfer |
| STRING | Search Tool for the Retrieval of Interacting Genes/Proteins |
| TA | Tetramer with interfacial helices |
| TB | Tetramer with interfacial beta-sheets |
| Thr/T | Threonine |
| Trdh | Double hotdog version of H2 hexamer |
| tRNA | transfer ribonucleic acid |
| Trp/W | Tryptophan |
| Tyr | Tyrosine |
| UDP | Uridine diphosphate |
| UniProtKB | Universal Protein resource Knowledgebase |

| | |
|---|---|
| Val/V | Valine |

## LIST OF FIGURES

## LIST OF TABLES

# CHAPTER 1

# INTRODUCTION TO EVOLUTION OF STRUCTURE AND FUNCTION WITHIN ENZYME FAMILIES

Fundamentally, evolution occurs on the molecular scale, in enzymes' acquisition of novel functions leading to new phenotypes and evolutionary advantage. Thus, understanding the manner in which such novel function arises is of key importance to our grasp of evolution in general and in specific, for example in the study of the evolution of antibiotic-resistant bacteria and how said resistance develops. In getting to the heart of what drives evolution, we must go beyond the chemistry catalyzed and look at the enzymes that do the work. And because ultimately it is enzyme structure that defines enzyme function, looking at structural evolution is a natural first step towards this goal.

## 1.1: Formation of enzyme families

Gene duplication is one of the major ways evolution occurs on the genetic scale (*1, 2*). A given gene is duplicated, resulting in two proteins that are identical in sequence and structure. Classically, the original protein retains its function whereas the duplicate is freed from selective pressure and is able to acquire new function by a variety of methods (*3-5*). If the duplicate persists over generations and does successfully acquire new function, the two proteins are now classified as paralogs—proteins with divergent function within the same species, related by the gene duplication event (*6*). Orthologs arise as speciation events occur: proteins with the same function in different species which will reflect genetic drift associated with speciation and further evolution (*6*).

Over many, many such generations and speciation events, a population of paralogous and orthologous enzymes, all related to one another evolutionarily, can be said to comprise an enzyme or gene family (*7-10*). Enzyme families thus contain proteins sharing the same inherited backbone fold (*5*), catalytic scaffold (*11*), sequence motifs (*12*), and some degree of sequence homology (*13, 14*). As a result of the shared overall structure and catalytic scaffold, members of the same enzyme family typically catalyze the same reaction chemistry (*5, 15, 16*) and/or use a similar catalytic mechanism (*17, 18*).

Comparative analysis of these structure and function relationships within a family provides a foundation for predicting function of uncharacterized sequences within the family. Indeed, the central dogma of structural biology is that enzyme sequence determines structure, which determines function (*19, 20*). Previous studies demonstrate that, in many cases, two enzymes within the same family can be assigned the same function provided they share sufficient sequence identity, typically ~40-50% (*5, 13, 15, 19, 21*), and assigned similar reaction chemistry at lower sequence identities (*5, 15*). While these thresholds are expected to vary depending on the enzyme family and considerations such as domain inserts and length (*14*), they do provide a strategy for function prediction at high sequence identities. Such analysis also provides insight into the mechanisms by which enzymes evolve.

### 1.2: Enzyme promiscuity and evolvability

Enzyme promiscuity, the ability to carry out alternate chemical reactions or reactions with alternate substrates, is a key feature in enzyme evolution, on both the single enzyme scale and the enzyme family scale (*22-27*). Following a gene duplication event, a promiscuous duplicate is afforded the opportunity to hone whatever low-level promiscuous activity it already possesses without selection pressure to retain its original primary function.

Substrate promiscuity enhances the 'evolvability' of a protein family. As enzymes evolve, two key concepts are in conflict: robustness, or the degree to which the enzyme's native activity is unaffected by mutations, and plasticity, or the ability to gain novel function with a minimal number of mutations (*28-30*). Enzymes must be robust if they are to withstand the many deleterious mutations which naturally occur over generations and which don't confer any selective advantage; thus, mutations must not affect the physiological role of the enzyme. Contrarily, they must gain new function through minimal mutation if they are to evolve and gain new function on any useful timescale while avoiding deletion. The requirement for robustness is slightly lessened under gene duplication circumstances, as the original protein still fills the original physiological role. But duplication subjects the duplicate enzyme to Ohno's dilemma, that the duplicate must undergo the rare mutations necessary to acquire new function, and that it must do so quickly enough that it does not undergo deleterious mutations that would remove it from the population (*3*).

It is thus advantageous for highly evolvable (robust and plastic) enzymes to be promiscuous, as well—in such a case, they retain normal function but have some small function towards other substrates, which in turn makes them more evolvable (*31*). Unless they are also robust in structure, these enzymes risk undergoing mutations that render them unfoldable and, hence, useless in terms of evolutionary advantage and propagation. This conflict is navigated by enzymes which manage to have plasticity and promiscuity that do not significantly alter the native enzyme activity (*30*). This is a particularly useful template for evolution in that these promiscuous activities can lay dormant until there is a gene duplication event, after which they can quickly optimize the promiscuous activity without concern for decreasing physiological activity. Indeed, directed evolution experiments demonstrate that promiscuous enzymes are quickly capable of specializing to new functions with few mutations (*32*).

However, it is important to note the range of enzyme promiscuity possibilities. In some cases promiscuity takes the form of substrate ambiguity, in which the enzyme has a physiological substrate but is capable of catalyzing very similar off-targets as well (*24*). Function assignment is particularly difficult with very promiscuous enzymes. Because promiscuous enzymes are able to catalyze a wide range of reactions, the physiological substrate is not always immediately apparent from activity assays. As such, other methods are necessary for function assignment, including contextual clues from neighboring genes and regulatory proteins, as described below.

## 1.3: Beyond sequence similarity: using other clues to infer enzyme function

### *1.3.1: The need for and the problem with automated tools*

Since the first sequenced genome, the scientific community has increasingly been in the position of having more data than we know what to do with. Nearly two decades ago, when only seven genomes had been completely sequenced, entire genomes were grouped into clustered orthologous groups which could all be expected to have the same function; assigning function to one member allowed the same functional assignment for all members (*10*). This idea of sequence-based clustering has been increasingly useful as the numbers of available protein sequences have exploded. Programs like CD-HIT (*33*) automatically cluster sequences together based on a certain sequence threshold, allowing for a less curated but nonetheless similar function assignment method. However, gene and protein sequencing are now sufficiently inexpensive that the limit to our scientific knowledge of enzymes is not how many genomes or sequences are available, but how many sequences we can reasonably and correctly inspect and functionally annotate. In lieu of running extensive assays or even high-throughput screens to experimentally assign function to individual proteins, bioinformatics approaches offer *in silico* alternatives capable of working on much larger scales. They may also provide guidance for further confirmation of function by narrowing the probable substrate library.

Databases such NCBI and the UniProtKB automate sequence annotation (UniProt 2015, October 2002), but such annotations are not without significant errors. Indeed, Schnoes *et al* demonstrated that in four major databases including the NCBI and UniProtKB, up to 40% of sequences were misannotated (*34, 77, 78*). UniProt combats annotation errors by hosting a database of manually annotated and reviewed sequences, Swiss-Prot, in addition to the automatically annotated database, TrEMBL; however, requiring manual annotation severely limits the number of annotated sequences. As of May 2015, Swiss-Prot contains 548,208 sequences compared to TrEMBL's 46,714,516 sequences. The TrEMBL database is too large to work with manually whereas the Swiss-Prot database does not contain enough manually curated information to cover an entire superfamily. What is required is annotation tools that are more accurate than current automatic methods but faster than current manual methods. Additional tools can be used to guide this approach.

### *1.3.2: Gene context and biological range*

A gene context or gene neighborhood describes the region surrounding a gene encoding a protein of interest. Proteins with related function or stepwise function within a pathway are not uncommonly encoded in geographically compact operons; thus, identifying gene context may indicate the biological role or pathway of a protein of interest (*35*). Gene context has been used to assign enzyme function (*36*), then verified with in-vitro screenings (*35*), or it can be used in conjunction with other clues, such as high-throughput screening results, to provide suggestions as to a range of possible functions. A mini-review by Gerlt *et al* provides examples of both approaches in the enolase superfamily (*37*).

The biological range of a protein and its orthologs is also a useful tool. It may be used to track the acquisition and loss of function as well as the rise of orthologs by speciation. Together with ortholog studies, it can be used to track the evolution of and manner of acquisition of novel function according to structural and catalytic site changes (*38*). Comparing the numbers of protein family members across the biological range (gains and losses) give a baseline for understanding the evolutionary landscape for the family (*8*). Abnormal biological ranges can also indicate horizontal gene transfer, as in the case of horizontal gene transfer of similar *mariner* elements that were found in both insects and flatworms, but not close relatives of the flatworms, indicating a transfer event (*39*).

### *1.3.3: The Enzyme Function Initiative (EFI)*

Several research groups across the country are working together under the Enzyme Function Initiative (EFI) to develop strategies for determining enzyme function based on structure, sequence, reaction results, and especially the interplay among all three (*40*). The Dunaway-Mariano lab, in collaboration with the Allen lab, has focused on high-throughput screens and crystallization of HADSF and hotdog-fold members. The goal is to explore the sequence and structure landscape for previously uncharacterized structures which may indicate novel function.

## 1.4: The hotdog-fold superfamily

### *1.4.1: Structural diversity within the hotdog-fold family*

The hotdog-fold superfamily is a functionally diverse family of evolutionarily related enzymes which share a common α + β-fold. Janet Smith and her coworkers dubbed the superfamily the “hotdog-fold” based on its founding member, the *E. coli* β-hydroxydecanoyl-holo acyl carrier protein (ACP) dehydrase/isomerase (*41*). The general tertiary structure (Figure 1.1) of the family takes the form of a 5-turn α-helix (the hotdog), nested in a curved, 7-stranded anti-parallel β-sheet (the bun). The essential functional unit is a dimer, with the subunit interface joining the two β-sheets to form a continuous 14-stranded sheet; the two active sites are located at opposite ends of the interfaced sheets (*41*-*44*).

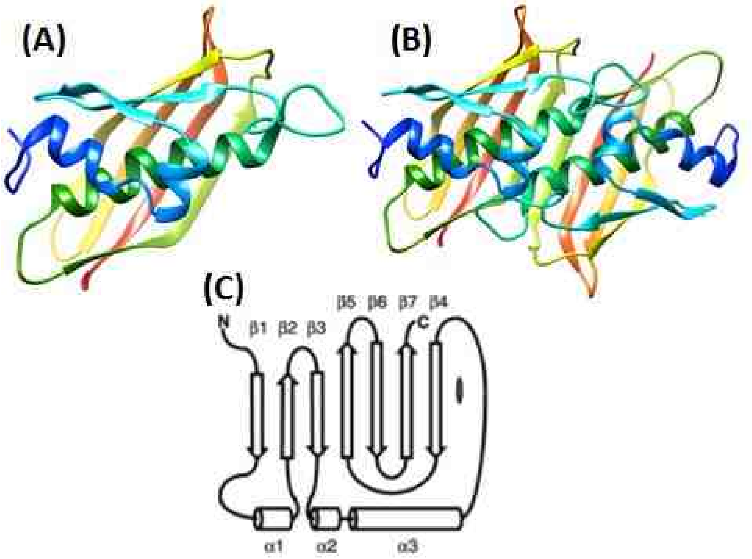


**Figure 1.1:** (A) *E. coli* ydiI monomer (*43*) and (B) dimer, both visualized in Chimera (*45*) (PDB ID: 4K49). (C) Typical monomeric structure of a hotdog domain; image from (*41*).

While the minimum functional unit is a dimer, hotdog-fold members may take on a number of different quaternary structures. Pidugu *et al* identified seven such variances, shown in Figure 1.2 (*46*). The identified structures are: dimer (D), double hotdog (dh), hexamer (trimer of dimers) with active site loops at their interfaces (H1), hexamer (trimer of dimers) with N-terminal helices at their interfaces (H2), hexamer (trimer of dimers) with head-to-tail arrangement (H3), tetramer (dimer of dimers) with helix interactions at their interface (TA), tetramer (dimer of dimers) with β-sheet interactions at their interface (TB). Double hotdog tertiary structures take on similar quaternary structures to dimer formulations in the following ways: the TA tetramer made of dimers is similar in shape to a dimer of double hotdogs whose helices interact at the interface (DdhA), the TB tetramer made of dimers is similar in shape to a dimer of double hotdogs with β-sheet interactions at their interface (DdhB), and the H2 hexamer made of dimers is similar in shape to a trimer of double hotdogs whose helices interact at the interfaces (Trdh).

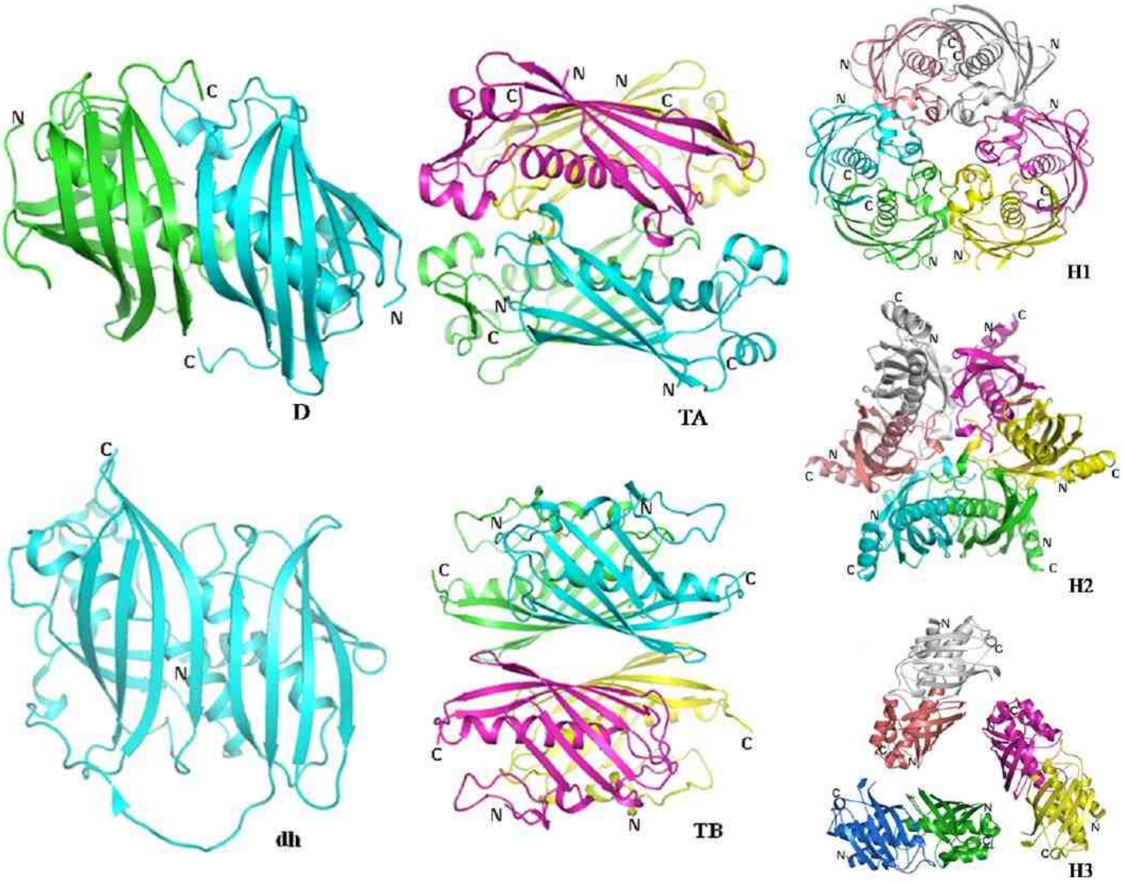


**Figure 1.2:** General types of quaternary structures into which hotdog-fold members have been demonstrated to assemble. Briefly: dimer (D), double hotdog with dimer-like structure (dh), loop-interface tetramer similar to a dimer of double hotdogs (TA/DdhA), β-sheet-interface tetramer (TB/DdhB), loop-interface hexamer (H1), helix-interface hexamer similar to a trimer of double hotdogs (H2/Trdh), and end-to-end interface hexamer (H3). Image from (*46*).

### *1.4.2: Evolution and the hotdog-fold*

The high plasticity of the hotdog-fold is inferred from the large degree of sequence variation observed between orthologs: the hotdog-fold superfamily exhibits sequence identities as low as 10-15% despite strict conservation of structure (*47*). The rapid adaptation of the hotdog-fold enzyme to a novel substrate is attributed to an active site platform that supports the participation of conserved catalytic residues in different spatial configurations and in different roles (*46*).

### *1.4.3: Chemical reactions of the hotdog-fold superfamily*

Most of the hotdog-fold functions can be categorized as either dehydratases/hydratases, catalyzing elimination or addition at the β-carbon position, or thioesterases, catalyzing hydrolysis at the thioester moiety (Figure 1.3). Individual subgroups of the hotdog-fold family are discussed further in Chapter 2 but a general overview follows.

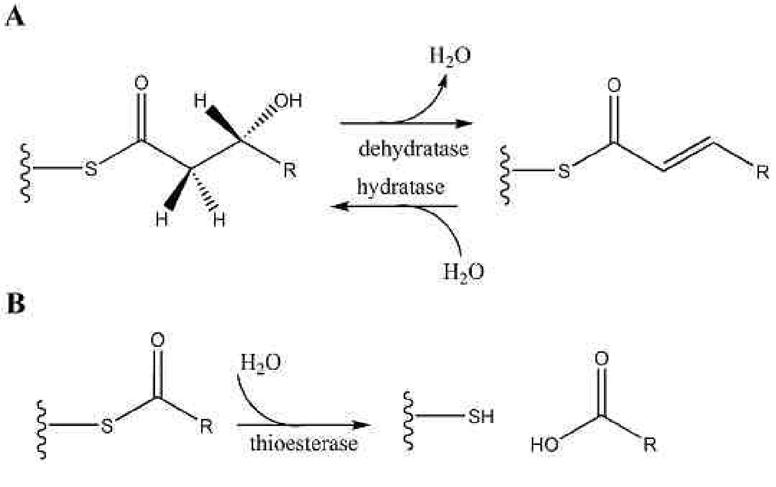


**Figure 1.3:** General reaction scheme for the dehydratase/hydrastase (A) and thioesterase (B) reactions typical of the hotdog-fold family.

Substrates for the hotdog-fold family are typically acylated or arylated Coenzyme A or *holo*-ACP, though other activities are possible (Figure 1.4). The hotdog bonding pocket is ideally suited for the pantetheine arm of CoA or ACP. It is a long, deep, primarily hydrophobic tunnel formed at the interface of the homodimer subunits (*41*). The binding pocket adapts based on the type and range of substrates catalyzed by each individual enzyme: enzymes catalyzing larger ranges of substrates tend to have a more open tunnel whereas it is more closed and defined for those with very limited substrate ranges or specific substrates. Thus, the hotdog-fold family is expected to, and does, carry out a wide range of reactions; it also tends toward promiscuity (*26*).

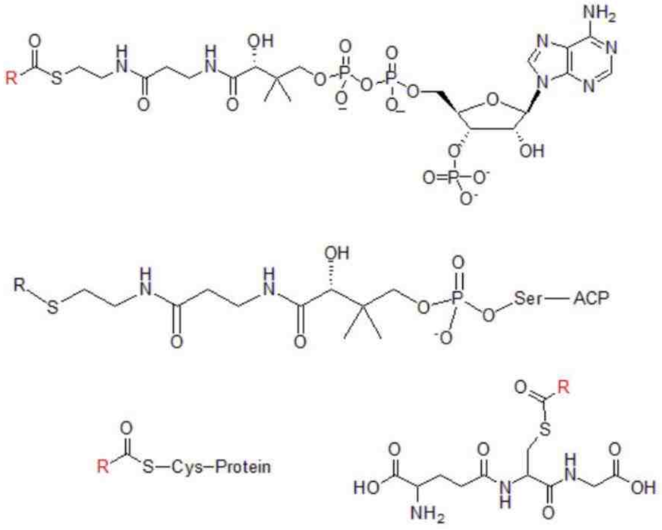


**Figure 1.4:** Biological thioesters. From top to bottom: Coenzyme A, pantetheine arm of *holo*-acyl carrier protein (ACP), modified cysteines of proteins, and glutathione. R groups are various acylated or aromatic compounds.

The dehydratases are used in the third step of type II fatty acid biosynthesis (FAS): conversion of β-hydroxyacyl-ACP to trans-2-acyl-ACP, preceded by condensation of malonyl-ACP by ß-ketoacyl-ACP synthase and reduction of the ß-ketoester by ß-ketoacyl-ACP reductase (*48*). A similar process is used in polyketide biosynthesis (PKS) as well, in which hotdog dehydratases also function. Hotdog-fold members also catalyze the backwards hydration reaction.

Hotdog-fold thioesterases hydrolyze the thioester bond between fatty acids and CoA or acyl carrier protein, resulting in free thiol and free carboxylic acid, which varies in size, shape, and polarity. Thioesters play a significant role in metabolism, membrane synthesis, signal transduction, and gene regulation within the cell (*49*). Thioesters are converted from carboxylic acids for myriad uses, including polyketide biosynthesis (*50*) and protein modification such as palmitoylation of cysteine for signaling (*51*). Thioesterases also plays a terminating role in fatty acid synthesis, in addition to its dehydratase role described above, by cleaving the fatty acid-ACP bond, releasing the fatty acid (*48*).

## 1.5: The Haloacid dehalogenase superfamily (HADSF)

### *1.5.1: Background and structure of the HADSF*

The Haloacid Dehalogenase Superfamily (HADSF) is a large, highly successful superfamily (>120,000 unique sequences), appearing across all three domains of life and typically represented by several members within a given organism, including 183 in *Homo sapiens* and 28 in *E. coli (52, 53)*. While its founding member is a dehalogenase and its members catalyze diverse reactions (*54, 55*), the majority of HADSF members catalyze phosphoryl transfer reactions (Figure 1.5) occurring through an aspartylphosphate intermediate (*52*).

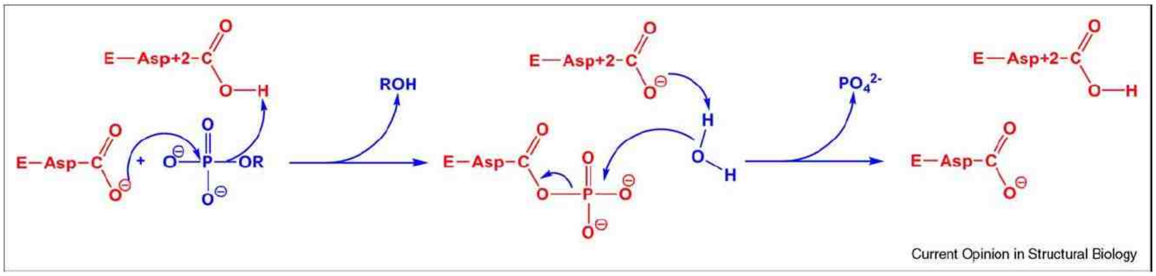


**Figure 1.5:** The aspartylphosphate intermediate catalytic mechanism used by phosphatase members of the HADSF. Image from (*52*).

The canonical structure for HADSF members is built around the catalytic core, which takes the structure of a Rossmannoid fold containing a phosphoryl transfer active site (*56*). Within the catalytic site, four key motifs are highly conserved, as shown in Figure 1.6. In Loop 1, the first Asp serves as a nucleophile while the second functions as a general acid/base (*57*). The second Asp first binds and protonates the leaving group of the substrate and subsequently deprotonates the nucleophile (*54, 57*). The residues of Loops 2 and 3 stabilize the aspartyl intermediate via hydrogen bonding (*56, 58*). The phosphatase members of the HADSF require a magnesium ion cofactor, which is positioned by the DxxxD motif of Loop 4 as well as the carboxylate of the first Asp and the C=O backbone of the second Asp in Loop 1 (*59*).

Three general cap types may be inserted at one of two insertion points and provide much of the basis for substrate recognition (*54, 60*), while catalysis is limited to the core residues described above (*52, 55, 60*). These cap domains are generally believed to participate in substrate binding—the cap can close to desolvate the active site and individual cap residues typically interact with the substrate (*57, 60-66*).

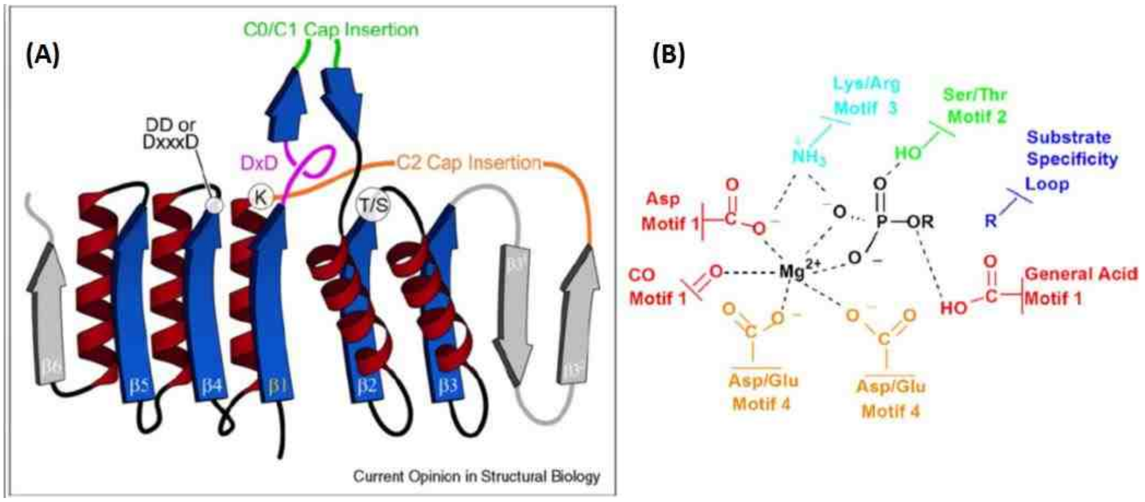


**Figure 1.6:** (A) The canonical HADSF Rossmann core domain with the four conserved motifs noted in black and pink and the variable cap insertion points noted in green and orange; image from (*52*). (B) Positioning the phosphate group and magnesium ion within the active site relative to the conserved active site residues and motifs/loops; image from (*54*).

HADSF members are categorized according to what cap type they possess (Figure 1.7). The C1 and C2 cap types are those that fold into distinct subdomains that are distinct from the core catalytic domains—they can be distinguished from each other based on their insert location—whereas the C0 cap types are inserts at either insert point that form small loops insufficient to be considered a domain distinct from the core (*54*). C1 caps are inserted in the middle of the β-hairpin of the flap motif; they can be further classified as α-helical vs α+β fold caps, though the latter are seen only in P-type ATPases. C2 types are inserted at the linker position after Loop 2; they can be further divided into two large, unrelated α+β with core β-sheet domains and a smaller flap-like structure. The cap domains are particularly interesting because the core fold can essentially operate on its own—the cap appears to be unnecessary for fundamental catalytic activity (*67*).

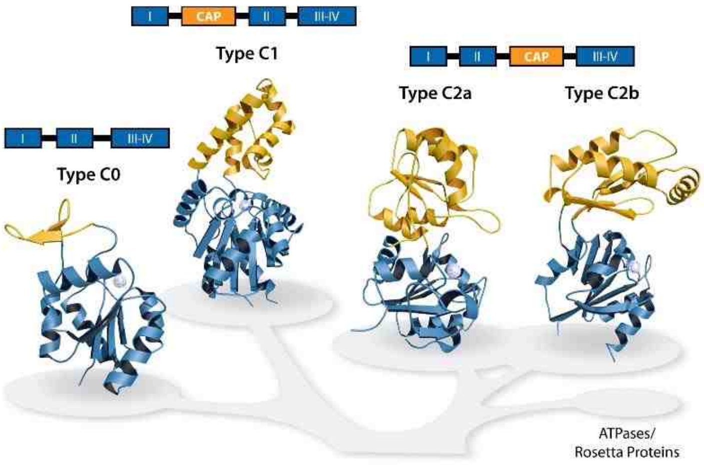


**Figure 1.7:** The four cap-based subclasses of HADSF members. Image acquired May 2015 from http://chemweb.bu.edu/groups/allengroup/efi.html.

### *1.5.2: Chemistry catalyzed by the HADSF*

Nearly 80% of the HADSF is comprised of phosphatases, with most of the remainder comprised of ATPases (*52*). Dephosphorylation reactions are highly in demand in the cell (*68*); indeed, 35-40% of the *E. coli* metabolome contains a phosphoryl group (*69*). The HADSF is a central player in catalyzing these reactions, which are used in myriad functions such as essential metabolic roles, regulation, proofreading, scavenging, and general housekeeping (*26, 52, 54*).

### *1.5.3: Evolvability of the HADSF*

As discussed above, the HADSF catalyzes the lion's share of crucial dephosphorylation reactions. That it catalyzes such important reactions may explain some degree of its success as a superfamily, but were it solely responsible, we would expect other families with phosphotransferase activity to have an equal share in the range of phosphoryl group catalysis. However, the HADSF outnumbers other protein families in this function space (*70, 71*). Thus, some other factors likely contribute to the HADSF's success. One such factor may be the inherent evolvability of the HADSF (*36, 52, 62, 72*)—a highly evolvable protein family would be able to accrue the many subtly different phosphotransferase activities of the HADSF without a significant stability penalty or deleterious mutations.

The HADSF is believed to be particularly well-suited for evolution and evolution-based studies for a number of reasons. Firstly, the bulk of the enzyme, and the location of the catalytic site, take the form of the particularly stable Rossmann-like fold (*11, 73*). This strong structure stability lends the enzymes' structural robustness, thus allowing them sequence plasticity; the fold persists across all of the members of the HADSF, despite family members routinely sharing sequence identities less than 15% (*52, 58*). This high stability suggests that enzymes can tolerate mutations that might otherwise destabilize the enzyme, allowing for the introduction of mutations that may not have an immediate evolutionary advantage, but may in the future contribute to competitive advantage in different conditions or in the event of gene duplication (*74, 75*).

Secondly, members of the HADSF are in possession of varying cap domains, described above. It is believed that the introduction of the cap may be the primary attribute of the HADSF that introduces substrate specificity, both by limiting the types of substrates that can access the catalytic site simply due to size and hindrance, but also by providing a surface for substrate-specificity-conferring residues. The cap domain introduces opportunities for modifying substrate specificity. Thus, these cap domains are rife with evolutionary opportunity— because the core fold is sufficient for catalytic activity, changing the residues on the cap that interact with substrate may be enough to change substrate specificity of the entire enzyme. So the cap may act as a sort of substrate specificity pegboard, taking on new function with very simple add-or-remove changes. This concept suggests a straightforward pathway for evolution. High cap plasticity, paired with the robustness and stability of the core Rossmann fold, would allow for the HADSF to rapidly alter substrate specificity with a minimal number of residue changes; indeed, it could help explain the ubiquity of the superfamily and the wide variety of phosphoryl transfer reactions it can catalyze.

Even compared to other superfamilies, the HADSF has low internal sequence identity, using E-values as a proxy in which smaller is better (*14*). E-values corresponding to an average of 30-40% sequence identity tend to be very small (stringent) for other superfamilies— 35% sequence identity corresponds to E-values $<10^{-90}$ for the enolase superfamily (*37*) and $<10^{-55}$ for the proline racemase family (*76*) — but, for the HADs, is a much larger (less stringent) $<10^{-20}$ (*53*). The varied type and location of cap inserts found in the HADSF may be partly to blame for this unusually low internal sequence identity and difficulty in automatically assigning functions (*61*).

### 1.6: Bioinformatic goals

For my doctoral work, I have focused on mining and manipulating the vast amounts of data available in public databases in order to identify and explore relationships among members in the hotdog-fold and Haloacid dehalogenase superfamilies. I used a combination of automatic and manual function assignment and techniques targeted to specific aspects of the two superfamilies of interest. Specifically, I combine various methods of manual gene context, biological range determination, and externally-conducted high-throughput screens with generation of large, homology-clustered sequence similarity networks to explore sequence-structure-function landscapes and assign tentative functions to previous unannotated enzymes.

In the hotdog-fold superfamily, I have focused on identifying general trends applied across the entire superfamily. Due to the large size of the hotdog-fold, relatively small amounts of data must be applied to the entire sequence and structure space. Nonetheless, this information can be used to annotate previously unannotated regions and identify under-characterized areas that would make good candidates for future work. Results of high-throughput screens indicate that much of the hotdog-fold sequence space has promiscuous activity.

In the HADSF, I have explored the members belonging to the Firmicutes phylum in order to better understand the evolutionary relationship across the phylum, including the appearance of fusion proteins within single domain clusters, possibilities of gene transfer among Firmicutes members and across other taxonomic groups, and the diversity of HADSF members across the phylum. I also explored gene contexts and biological ranges to identify previously uncharacterized functions.

### 1.7: References

1. Ohno, S., *Evolution by gene duplication*. Allen & Unwin; Springer-Verlag: London, New York, 1970.
2. Ohno, S.; Wolf, U.; Atkin, N. B., Evolution from fish to mammals by gene duplication. *Hereditas* **1968,** *59*, 169-87.
3. Bergthorsson, U.; Andersson, D. I.; Roth, J. R., Ohno's dilemma: evolution of new genes under continuous selection. *Proc Natl Acad Sci U S A* **2007,** *104*, 17004-9.
4. Hughes, A. L., The evolution of functionally novel proteins after gene duplication. *Proc Biol Sci* **1994,** *256*, 119-24.
5. Todd, A. E.; Orengo, C. A.; Thornton, J. M., Evolution of function in protein superfamilies, from a structural perspective. *J Mol Biol* **2001,** *307*, 1113-43.
6. Fitch, W. M., Distinguishing homologous from analogous proteins. *Systematic zoology* **1970,** *19*, 99-113.
7. Henikoff, S.; Greene, E. A.; Pietrokovski, S.; Bork, P.; Attwood, T. K.; Hood, L., Gene families: the taxonomy of protein paralogs and chimeras. *Science* **1997,** *278*, 609-14.

8. Demuth, J. P.; Hahn, M. W., The life and death of gene families. *BioEssays : news and reviews in molecular, cellular and developmental biology* **2009,** *31*, 29-39.

9. Thornton, J. W.; DeSalle, R., Gene family evolution and homology: genomics meets phylogenetics. *Annual review of genomics and human genetics* **2000,** *1*, 41-73.

10. Tatusov, R. L.; Koonin, E. V.; Lipman, D. J., A genomic perspective on protein families. *Science* **1997,** *278*, 631-7.

11. Russell, R. B.; Sasieni, P. D.; Sternberg, M. J., Supersites within superfolds. Binding site similarity in the absence of homology. *J Mol Biol* **1998,** *282*, 903-18.

12. Bork, P.; Koonin, E. V., Protein sequence motifs. *Curr Opin Struct Biol* **1996,** *6*, 366-76.

13. Dayhoff, M. O., The origin and evolution of protein superfamilies. *Fed Proc* **1976,** *35*, 2132-8.

14. Pearson, W. R., An introduction to sequence similarity ("homology") searching. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* **2013,** *Chapter 3*, Unit3 1.

15. Wilson, C. A.; Kreychman, J.; Gerstein, M., Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J Mol Biol* **2000,** *297*, 233-49.

16. Perona, J. J.; Craik, C. S., Evolutionary divergence of substrate specificity within the chymotrypsin-like serine protease fold. *J Biol Chem* **1997,** *272*, 29987-90.

17. Glasner, M. E.; Gerlt, J. A.; Babbitt, P. C., Evolution of enzyme superfamilies. *Current opinion in chemical biology* **2006,** *10*, 492-7.

18. Babbitt, P. C.; Gerlt, J. A., Understanding enzyme superfamilies. Chemistry As the fundamental determinant in the evolution of new catalytic activities. *J Biol Chem* **1997,** *272*, 30591-4.

19. Chothia, C.; Lesk, A. M., The relation between the divergence of sequence and structure in proteins. *The EMBO journal* **1986,** *5*, 823-6.

20. Hegyi, H.; Gerstein, M., The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J Mol Biol* **1999,** *288*, 147-64.

21. Tian, W.; Skolnick, J., How well is enzyme function conserved as a function of pairwise sequence identity? *J Mol Biol* **2003,** *333*, 863-82.

22. Jensen, R. A., Enzyme recruitment in evolution of new function. *Annu Rev Microbiol* **1976,** *30*, 409-25.

23. Copley, S. D., Enzymes with extra talents: moonlighting functions and catalytic promiscuity. *Current opinion in chemical biology* **2003,** *7*, 265-72.

24. Babtie, A.; Tokuriki, N.; Hollfelder, F., What makes an enzyme promiscuous? *Current opinion in chemical biology* **2010,** *14*, 200-7.

25. Khersonsky, O.; Tawfik, D. S., Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annu Rev Biochem* **2010,** *79*, 471-505.

26. Pandya, C.; Farelli, J. D.; Dunaway-Mariano, D.; Allen, K. N., Enzyme promiscuity: engine of evolutionary innovation. *J Biol Chem* **2014,** *289*, 30229-36.

27. O'Brien, P. J.; Herschlag, D., Catalytic promiscuity and the evolution of new enzymatic activities. *Chem Biol* **1999,** *6*, R91-R105.

28. Bornberg-Bauer, E.; Kramer, L., Robustness versus evolvability: a paradigm revisited. *Hfsp J* **2010,** *4*, 105-8.

29. Khersonsky, O.; Tawfik, D. S., Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annu Rev Biochem* **2010,** *79*, 471-505.

30. Wagner, A., Robustness and evolvability: a paradox resolved. *Proc Biol Sci* **2008,** *275*, 91-100.

31. Bloom, J. D.; Romero, P. A.; Lu, Z.; Arnold, F. H., Neutral genetic drift can alter promiscuous protein functions, potentially aiding functional evolution. *Biol Direct* **2007,** *2*, 17.

32. Kazlauskas, R. J., Enhancing catalytic promiscuity for biocatalysis. *Current opinion in chemical biology* **2005,** *9*, 195-201.

33. Li, W.; Godzik, A., Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **2006,** *22*, 1658-9.

34. Schnoes, A. M.; Brown, S. D.; Dodevski, I.; Babbitt, P. C., Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS computational biology* **2009,** *5*, e1000605.

35. Overbeek, R.; Fonstein, M.; D'Souza, M.; Pusch, G. D.; Maltsev, N., The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A* **1999,** *96*, 2896-901.

36. Nguyen, H. H.; Wang, L.; Huang, H.; Peisach, E.; Dunaway-Mariano, D.; Allen, K. N., Structural determinants of substrate recognition in the HAD superfamily member D-glycero-D-manno-heptose-1,7-bisphosphate phosphatase (GmhB). *Biochemistry* **2010,** *49*, 1082-92.

37. Gerlt, J. A.; Babbitt, P. C.; Jacobson, M. P.; Almo, S. C., Divergent evolution in enolase superfamily: strategies for assigning functions. *J Biol Chem* **2012,** *287*, 29-34.

38. Daughtry, K. D.; Huang, H.; Malashkevich, V.; Patskovsky, Y.; Liu, W.; Ramagopal, U.; Sauder, J. M.; Burley, S. K.; Almo, S. C.; Dunaway-Mariano, D.; Allen, K. N., Structural basis for the divergence of substrate specificity and biological function within HAD phosphatases in lipopolysaccharide and sialic acid biosynthesis. *Biochemistry* **2013,** *52*, 5372-86.

39. Robertson, H. M., Multiple Mariner transposons in flatworms and hydras are related to those of insects. *The Journal of heredity* **1997,** *88*, 195-201.

40. Gerlt, J. A.; Allen, K. N.; Almo, S. C.; Armstrong, R. N.; Babbitt, P. C.; Cronan, J. E.; Dunaway-Mariano, D.; Imker, H. J.; Jacobson, M. P.; Minor, W.; Poulter, C. D.; Raushel, F. M.; Sali, A.; Shoichet, B. K.; Sweedler, J. V., The Enzyme Function Initiative. *Biochemistry* **2011,** *50*, 9950-62.

41. Leesong, M.; Henderson, B. S.; Gillig, J. R.; Schwab, J. M.; Smith, J. L., Structure of a dehydratase-isomerase from the bacterial pathway for biosynthesis of unsaturated fatty acids: two catalytic activities in one active site. *Structure* **1996,** *4*, 253-64.

42. Dillon, S. C.; Bateman, A., The Hotdog fold: wrapping up a superfamily of thioesterases and dehydratases. *BMC bioinformatics* **2004,** *5*, 109.

43. Wu, R.; Latham, J. A.; Chen, D.; Farelli, J.; Zhao, H.; Matthews, K.; Allen, K. N.; Dunaway-Mariano, D., Structure and catalysis in the Escherichia coli hotdog-fold thioesterase paralogs YdiI and YbdB. *Biochemistry* **2014,** *53*, 4788-805.

44. Song, F.; Zhuang, Z.; Finci, L.; Dunaway-Mariano, D.; Kniewel, R.; Buglino, J. A.; Solorzano, V.; Wu, J.; Lima, C. D., Structure, function, and mechanism of the phenylacetate pathway hot dog-fold thioesterase PaaI. *J Biol Chem* **2006,** *281*, 11028-38.

45. Pettersen, E.; Goddard, T.; Huang, C.; Couch, G.; Greenblatt, D.; Meng, E.; Ferrin, T., UCSF Chimera-- a visualization system for explroatory research and analysis. *J Comput Chem* **2004,** *25*, 1605-1612.

46. Pidugu, L. S.; Maity, K.; Ramaswamy, K.; Surolia, N.; Suguna, K., Analysis of proteins with the 'hot dog' fold: prediction of function and identification of catalytic residues of hypothetical proteins. *BMC structural biology* **2009,** *9*, 37.

47. Zhuang, Z.; Song, F.; Zhao, H.; Li, L.; Cao, J.; Eisenstein, E.; Herzberg, O.; Dunaway-Mariano, D., Divergence of function in the hot dog fold

enzyme superfamily: the bacterial thioesterase YciA. *Biochemistry* **2008,** *47*, 2789-96.

48. Magnuson, K.; Jackowski, S.; Rock, C. O.; Cronan, J. E., Jr., Regulation of fatty acid biosynthesis in Escherichia coli. *Microbiological reviews* **1993,** *57*, 522-42.

49. Hunt, M. C.; Alexson, S. E., The role Acyl-CoA thioesterases play in mediating intracellular lipid metabolism. *Progress in lipid research* **2002,** *41*, 99-130.

50. Katz, L.; Donadio, S., Polyketide synthesis: prospects for hybrid antibiotics. *Annu Rev Microbiol* **1993,** *47*, 875-912.

51. Smotrys, J. E.; Linder, M. E., Palmitoylation of intracellular signaling proteins: regulation and function. *Annu Rev Biochem* **2004,** *73*, 559-87.

52. Allen, K. N.; Dunaway-Mariano, D., Markers of fitness in a successful enzyme superfamily. *Curr Opin Struct Biol* **2009,** *19*, 658-65.

53. Huang, H.; Pandya, C.; Liu, C.; Al-Obaidi, N. F.; Wang, M.; Zheng, L.; Toews Keating, S.; Aono, M.; Love, J. D.; Evans, B.; Seidel, R. D.; Hillerich, B. S.; Garforth, S. J.; Almo, S. C.; Mariano, P. S.; Dunaway-Mariano, D.; Allen, K. N.; Farelli, J. D., Panoramic view of a superfamily of phosphatases through substrate profiling. *Proc Natl Acad Sci U S A* **2015**.

54. Burroughs, A. M.; Allen, K. N.; Dunaway-Mariano, D.; Aravind, L., Evolutionary genomics of the HAD superfamily: understanding the structural adaptations and catalytic diversity in a superfamily of phosphoesterases and allied enzymes. *J Mol Biol* **2006,** *361*, 1003-34.

55. Koonin, E. V.; Tatusov, R. L., Computer analysis of bacterial haloacid dehalogenases defines a large superfamily of hydrolases with diverse specificity. Application of an iterative approach to database search. *J Mol Biol* **1994,** *244*, 125-32.

56. Lu, Z.; Dunaway-Mariano, D.; Allen, K. N., The catalytic scaffold of the haloalkanoic acid dehalogenase enzyme superfamily acts as a mold for the trigonal bipyramidal transition state. *Proceedings of the National Academy of Sciences of the United States of America* **2008,** *105*, 5687-92.

57. Lahiri, S. D.; Zhang, G.; Dunaway-Mariano, D.; Allen, K. N., Caught in the act: the structure of phosphorylated beta-phosphoglucomutase from Lactococcus lactis. *Biochemistry* **2002,** *41*, 8351-9.

58. Allen, K. N.; Dunaway-Mariano, D., Phosphoryl group transfer: evolution of a catalytic scaffold. *Trends Biochem Sci* **2004,** *29*, 495-503.

59. Zhang, G.; Morais, M. C.; Dai, J.; Zhang, W.; Dunaway-Mariano, D.; Allen, K. N., Investigation of metal ion binding in phosphonoacetaldehyde hydrolase identifies sequence markers for metal-activated enzymes of the HAD enzyme superfamily. *Biochemistry* **2004,** *43*, 4990-7.

60. Lahiri, S. D.; Zhang, G.; Dai, J.; Dunaway-Mariano, D.; Allen, K. N., Analysis of the substrate specificity loop of the HAD superfamily cap domain. *Biochemistry* **2004,** *43*, 2812-20.

61. Pandya, C.; Dunaway-Mariano, D.; Xia, Y.; Allen, K. N., Structure-guided approach for detecting large domain inserts in protein sequences as

illustrated using the haloacid dehalogenase superfamily. *Proteins* **2014,** *82*, 1896-906.

62. Huang, H.; Patskovsky, Y.; Toro, R.; Farelli, J. D.; Pandya, C.; Almo, S. C.; Allen, K. N.; Dunaway-Mariano, D., Divergence of structure and function in the haloacid dehalogenase enzyme superfamily: Bacteroides thetaiotaomicron BT2127 is an inorganic pyrophosphatase. *Biochemistry* **2011,** *50*, 8937-49.

63. Peeraer, Y.; Rabijns, A.; Verboven, C.; Collet, J. F.; Van Schaftingen, E.; De Ranter, C., High-resolution structure of human phosphoserine phosphatase in open conformation. *Acta Crystallogr D Biol Crystallogr* **2003,** *59*, 971-7.

64. Kim, H. Y.; Heo, Y. S.; Kim, J. H.; Park, M. H.; Moon, J.; Kim, E.; Kwon, D.; Yoon, J.; Shin, D.; Jeong, E. J.; Park, S. Y.; Lee, T. G.; Jeon, Y. H.; Ro, S.; Cho, J. M.; Hwang, K. Y., Molecular basis for the local conformational rearrangement of human phosphoserine phosphatase. *J Biol Chem* **2002,** *277*, 46651-8.

65. Wang, W.; Cho, H. S.; Kim, R.; Jancarik, J.; Yokota, H.; Nguyen, H. H.; Grigoriev, I. V.; Wemmer, D. E.; Kim, S. H., Structural characterization of the reaction pathway in phosphoserine phosphatase: crystallographic "snapshots" of intermediate states. *J Mol Biol* **2002,** *319*, 421-31.

66. Zhang, G.; Mazurkie, A. S.; Dunaway-Mariano, D.; Allen, K. N., Kinetic evidence for a substrate-induced fit in phosphonoacetaldehyde hydrolase catalysis. *Biochemistry* **2002,** *41*, 13370-7.

67. Fortpied, J.; Maliekal, P.; Vertommen, D.; Van Schaftingen, E., Magnesium-dependent phosphatase-1 is a protein-fructosamine-6-phosphatase potentially involved in glycation repair. *J Biol Chem* **2006,** *281*, 18378-85.

68. Dzeja, P. P.; Terzic, A., Phosphotransfer networks and cellular energetics. *J Exp Biol* **2003,** *206*, 2039-47.

69. Nobeli, I.; Ponstingl, H.; Krissinel, E. B.; Thornton, J. M., A structure-based anatomy of the E.coli metabolome. *J Mol Biol* **2003,** *334*, 697-719.

70. Vincent, J. B.; Crowder, M. W.; Averill, B. A., Hydrolysis of phosphate monoesters: a biological problem with multiple chemical solutions. *Trends Biochem Sci* **1992,** *17*, 105-10.

71. Vetter, I. R.; Wittinghofer, A., Nucleoside triphosphate-binding proteins: different scaffolds to achieve phosphoryl transfer. *Quarterly reviews of biophysics* **1999,** *32*, 1-56.

72. Rangarajan, E. S.; Proteau, A.; Wagner, J.; Hung, M. N.; Matte, A.; Cygler, M., Structural snapshots of Escherichia coli histidinol phosphate phosphatase along the reaction pathway. *J Biol Chem* **2006,** *281*, 37930-41.

73. Mirny, L. A.; Shakhnovich, E. I., Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J Mol Biol* **1999,** *291*, 177-96.

74. DePristo, M. A.; Weinreich, D. M.; Hartl, D. L., Missense meanderings in sequence space: a biophysical view of protein evolution. *Nat Rev Genet* **2005,** *6*, 678-87.

75. Aharoni, A.; Gaidukov, L.; Khersonsky, O.; Mc, Q. G. S.; Roodveldt, C.; Tawfik, D. S., The 'evolvability' of promiscuous protein functions. *Nature genetics* **2005,** *37*, 73-6.

76. Zhao, S.; Sakai, A.; Zhang, X.; Vetting, M. W.; Kumar, R.; Hillerich, B.; San Francisco, B.; Solbiati, J.; Steves, A.; Brown, S.; Akiva, E.; Barber, A.; Seidel, R. D.; Babbitt, P. C.; Almo, S. C.; Gerlt, J. A.; Jacobson, M. P., Prediction and characterization of enzymatic activities guided by sequence similarity and genome neighborhood networks. *eLife* **2014,** *3*.

77. Radivojac, P.; Clark, W. T.; Oron, T. R.; Schnoes, A. M.; Wittkop, T.; Sokolov, A.; Graim, K.; Funk, C.; Verspoor, K.; Ben-Hur, A.; Pandey, G.; Yunes, J. M.; Talwalkar, A. S.; Repo, S.; Souza, M. L.; Piovesan, D.; Casadio, R.; Wang, Z.; Cheng, J.; Fang, H.; Gough, J.; Koskinen, P.; Toronen, P.; Nokso-Koivisto, J.; Holm, L.; Cozzetto, D.; Buchan, D. W.; Bryson, K.; Jones, D. T.; Limaye, B.; Inamdar, H.; Datta, A.; Manjari, S. K.; Joshi, R.; Chitale, M.; Kihara, D.; Lisewski, A. M.; Erdin, S.; Venner, E.; Lichtarge, O.; Rentzsch, R.; Yang, H.; Romero, A. E.; Bhat, P.; Paccanaro, A.; Hamp, T.; Kassner, R.; Seemayer, S.; Vicedo, E.; Schaefer, C.; Achten, D.; Auer, F.; Boehm, A.; Braun, T.; Hecht, M.; Heron, M.; Honigschmid, P.; Hopf, T. A.; Kaufmann, S.; Kiening, M.; Krompass, D.; Landerer, C.; Mahlich, Y.; Roos, M.; Bjorne, J.; Salakoski,

T.; Wong, A.; Shatkay, H.; Gatzmann, F.; Sommer, I.; Wass, M. N.; Sternberg, M. J.; Skunca, N.; Supek, F.; Bosnjak, M.; Panov, P.; Dzeroski, S.; Smuc, T.; Kourmpetis, Y. A.; van Dijk, A. D.; ter Braak, C. J.; Zhou, Y.; Gong, Q.; Dong, X.; Tian, W.; Falda, M.; Fontana, P.; Lavezzo, E.; Di Camillo, B.; Toppo, S.; Lan, L.; Djuric, N.; Guo, Y.; Vucetic, S.; Bairoch, A.; Linial, M.; Babbitt, P. C.; Brenner, S. E.; Orengo, C.; Rost, B.; Mooney, S. D.; Friedberg, I., A large-scale evaluation of computational protein function prediction. *Nature methods* 2013, **10**, 221-7.

78. Devos, D.; Valencia, A., Intrinsic errors in genome annotation. *Trends in genetics : TIG* **2001,** *17*, 429-31.

# CHAPTER 2

# EXPLORATION OF DIVERGENCE OF SEQUENCE AND FUNCTION WITHIN THE HOTDOG-FOLD ENZYME SUPERFAMILY

## 2.1: Introduction

### *2.1.1: Subfamilies within the hotdog-fold superfamily*

Some members of the hotdog-fold superfamily are known to associate with other domains which, together with a variety of catalyzed chemistry described below, lead to a number of subfamily divisions within the larger superfamily. Currently, Pfam divides the hotdog clade into 13 subfamilies, which are themselves associated with InterPro subgroupings, as illustrated in Table 2.1.

| PFAM members of Hot Dog clan (CL0050) | PFAM name | IPRO associated with PFAM member | IPRO name | Master IPRO group to which this IPRO group belongs | IPRO groups subordinate to this IPRO group |
|---|---|---|---|---|---|
| PF02551 | Acyl-CoA thioesterase II domain | IPR025652 | Acyl-CoA thioesterase (double hot dog) | IPR029069 | none |
| PF03061 | 4HBT | IPR006683 | 4hbt | IPR029069 | IPR003736 |
| PF09500 | YiiD c-term | IPR012660 | Thioesterase, putative | IPR029069 | none |
| PF10862 | FcoT-like thioesterase domain | IPR022598 | Long-chain fatty acyl-CoA thioesterase, Rv0098-like | none | none |
| PF13279 | 4HBT_2 | n/a | n/a | n/a | n/a |
| PF03756 | A-factor biosynthesis hotdog domain | IPR005509 | A-factor biosynthesis hotdog domain | none | none |

| | | | | | |
|---|---|---|---|---|---|
| PF13452 | N-terminal half of MaoC dehydratase | n/a | n/a | n/a | n/a |
| PF13622 | 4HBT_3 | n/a | n/a | n/a | n/a |
| PF14539 | DUF4442 | n/a | n/a | n/a | n/a |
| PF01575 | MaoC like domain | IPR002539 | MaoC-like domain | IPR029069 | none |
| PF01643 | Acyl-ACP thioesterase | IPR002864 | Acyl-ACP thioesterase | none | none |
| PF07977 | FabA-like domain | IPR013114 | Beta-hydroxydecanoyl thiol ester dehydrase, FabA/FabZ | IPR029069 | none |
| P14765 | polyketide synthase dehydratase | n/a | n/a | n/a | n/a |
| | | **Additional IPRO groups** | **IPRO name** | **Master IPRO group to which this IPRO group belongs** | **IPRO subgroups to this IPRO group** |
| n/a | n/a | IPR025540 | Fluoroacetyl-CoA thioesterase | none | none |
| n/a | n/a | IPR029069 | Hot Dog domain | none | IPR025652,<br>IPR006683,<br>IPR012660,<br>IPR002539,<br>IPR013114 |

**Table 2.1:** Subfamilies of the hotdog-fold superfamily, as categorized by the Pfam and InterPro databases (*1, 2*). On the Pfam website, members of the hotdog clade are linked to corresponding InterPro groups, some of which belong to a separate, 'master' InterPro group encompassing additional hotdog domain sequences. Accessed October 10, 2014.

In 2004, Dillon and Bateman expanded on the Pfam categorizations described above and further categorized the hotdog-fold superfamily into 17 distinct subfamilies with varying degrees of characterization (*3*). Additional reviews address the comparative biological structure assembles (*4*) and catalytic architectures (*5*), of the hotdog-fold superfamily. A brief overview of the subfamilies and their associated functions follows in Section 2.1.1.1.

#### *2.1.1.1 Dehydratases/hydratases*

FabZ-like dehydratases are involved in type II fatty acid biosynthesis, specifically the third step in fatty acid elongation, conversion of β-hydroxyacyl-ACP to trans-2-acyl-ACP (*3, 6*). They function on short chain β-hydroxyacyl-ACPs and long chain saturated and unsaturated β-hydroxyacyl-ACPs (*6*). A subgroup of this subfamily is a coronafacid acid (CFA) dehydratase involved in coronatine, a virulence factor in the plant pathogen *Pseudomonas syringae;* this particular function is not associated with any additional domains whereas the above sometimes associated with LpxC domains (*3*).

FabA, like FabZ, catalyzes the third step in type II fatty acid biosynthesis but is alone in its 2-decenoyl-ACP isomerase activity, allowing it to initiate unsaturated fatty acid biosynthesis. It is most active on intermediate chain length β-hydroxyacyl-ACPs and also possesses significant activity toward both short and long chain saturated β-hydroxyacyl-ACPs, but not long chain unsaturated (*3, 6*). A subsection of FabA-like proteins are involved in the polyunsaturated fatty acid (PFA) biosynthesis, very similar to fatty acid synthesis. PFA biosynthesis proteins may contain two hotdog domains in addition to β-keto-acyl synthase (BKAS) domains and, sometimes, an acyl-transferase domain (*3*).

The MaoC hydratase-like subfamily consists of (R)-specific enoyl-coA hydratases. These catalyze the hydration of trans-2-enoyl-CoA to (R)-3-hydroxyacyl-CoA, supplying it from the beta-oxidation pathway to the PHA biosynthetic pathway (*7*). MaoC enzymes typically present with an N-terminal short-chain dehydrogenase domain (*3*). A subgroup of the MaoC dehydratese-like subfamily is the NodN-like group, which are involved in production of single molecules for root hair deformation in *Rhizobium* species (*8*).

#### *2.1.1.2 Thioesterases*

The acyl-CoA thioesterase family is the largest hotdog-fold family member; it catalyzes the hydrolysis of acyl-CoA thieosters to free fatty acids plus CoA-SH, a functionality associated with fatty acid metabolism. It contains members with specific activities for medium and long chain acyl-CoAs (*3*). In mammals, brown-fat-inducible thioesterase (BFIT) and cytoplasmic acetyl-CoA hydrolase (CACH) both contain StAR-related lipid-transfer (START) domain; brain acyl-CoA hydrolase has a duplicate of the hotdog domain (*9*).

The YbgC-like subfamily has been shown to hydrolyze conflicting acyl-CoA thioesters, both short-chain aliphatic acyl-CoA thioesters (*10*) and long chain (*11*). It is hypothesized to be involved in cell envelope maintenance due to its inclusion in the tol-pal cluster, the contents of which are believed to be involved in septation ring formation during cell division (*12*), but its specific function is still unclear.

The fat subfamily acyl-ACP thioesterases, which may be grouped into A (high activity with oleoyl-ACP) and B (high activity with palmitoyl-ACP) subgroups, catalyze the terminal fatty acid synthesis step in plants, breaking the thioester-ACP bond (*3, 13*).

The tesB-like subfamily, involved in fatty acid metabolism, acts on medium chain acyl-CoA thioesterases and is also known as human thioesterase II; it hydrolyzes palmitoyl-CoA to palmitate and CoA. It contains two hotdog domains and, on occasion, a cNMP domain (*3*).

The 4-hydroxybenzoyl (4HBT) subfamily, notable for its role in degradation of 4-chlorobenzoate as a carbon source, is broken into two groups I and II. The groups differ in the orientation of their active site residues and whether their α-helices are inwards- or outwards-facing in the tetramer-from-dimers structure. Some 4HBT-II members contain additional HAD domains (*3*).

Members of the PaaI subfamily are part of the phenylacetic acid (PA) catabolic pathway. It is believed to rescue CoA from phenylacetyl CoA if a downstream enzyme stalls; also rescues CoA from dead-end products (*14*).

3-hydroxyacyl-CoA dehydrogenase (3HCDH)-associated thioesterases are specific to short chain fatty acids of fatty acid metabolism and are typically fused to 3HCDH C-terminal and NAD-binding domains (*3*). The dehydrogenase region catalyzes the reduction of 3-hydroxyacyl-CoA to 3-oxoacyl-CoA (*15*); the combination of dehydrogenase and thiosterase regions may allow for substrate transportation.

#### *2.1.1.3 Other*

The FapR subfamily contains transcriptional regulators that control gene expression in type II fatty acid and phospholipid biosynthesis. It is controlled by malonyl-CoA and is associated with an HTH domain (*3*).

### *2.1.2: Goals*

We explored the sequence and structure space of the hotdog-fold family by combining sequence similarity networks with targeted high-throughput screening and literature reviews.  We chose a diverse number of target sequences to undergo expression and HTS results in order to characterize as much of the network as possible.  Ultimately, mapping HTS and published results allows us to assign function and structures to yet-uncharacterized proteins by virtue of their sequence and structure similarity to proteins of known function and highlight sequence spaces without annotatable function as areas for future study. It also paves the way for further annotations upon characterization of current areas of interest.

## **2.2: Methods**

### *2.2.1: Sequence similarity networks*

Sequence similarity networks (SSNs) have arisen as a recent tool used to qualitatively view relationships among a large number of sequences (*16-23*). They are particularly useful when considering a large enough number of sequences that viewing a multiple sequence alignment would be visually cumbersome, if not impossible, to meaningfully interpret.  SSNs are constructed by running an all-by-all BLAST for the sequences of interest; that is, a BLAST is run and an E-value computed for each query sequence against every other query sequence in the collection.  Once each sequence has an E-value relating it to every single other sequence, each sequence is represented as a node connected to other nodes by ‘edges’—lines representing the E-value relationship between every two nodes (Figure 2.1). An E-value threshold can be selected, below which sequence-sequence relationships (represented by edges and their E-values) will not be displayed. Thus, any remaining edges are known to represent relationships between nodes that are at or above the E-value threshold.

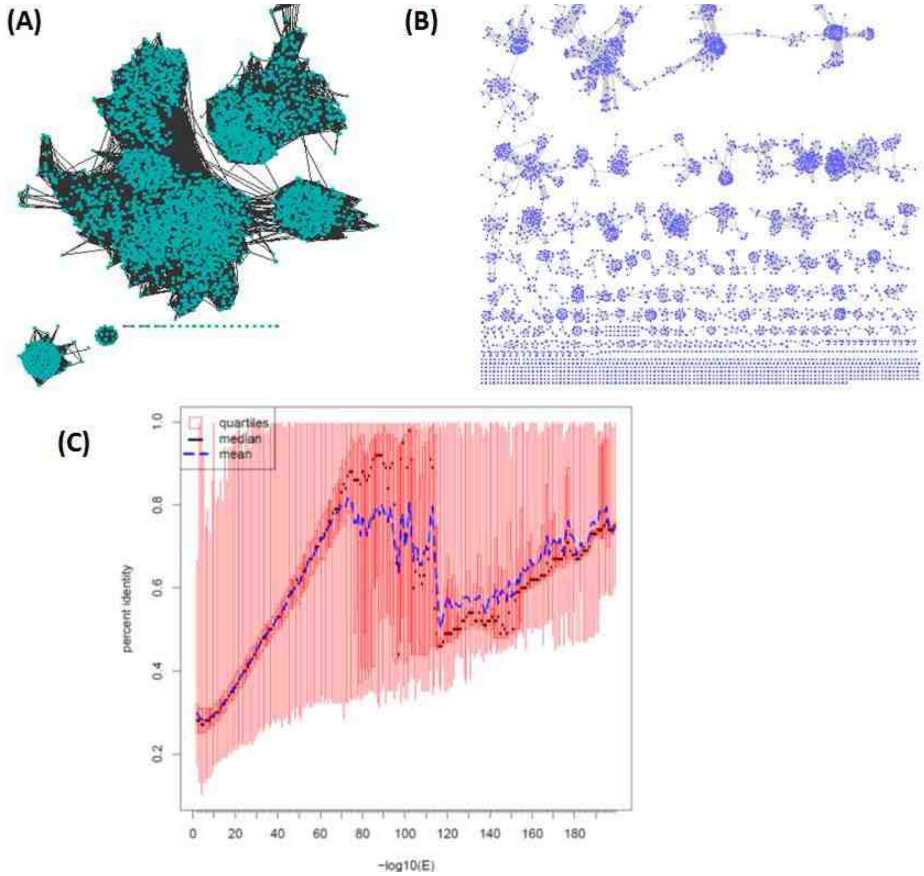


**Figure 2.1:** Representative node networks generated for the InterPro thioesterase family collection of sequences (IPR006683, August 2013) with 4,103 representative nodes clustered at >60% sequence identity. (A) E-value cutoff of $10^{-10}$, resulting in 405,613 edges. (B) A more stringent E-value cutoff of $10^{-30}$, resulting in 28,970 edges. (C) quartile plot for network generation, in which $10^{-30}$ corresponds to an average sequence identity of ~40% whereas $10^{-10}$ corresponds to average sequence identity of ~30%.

We generated SSNs using scripts provided by the Enzyme Function Initiative (*24*) which draw sequences from InterPro and Pfam memberships given as input (*1, 2*). For the hotdog-fold superfamily, this included all members of the Pfam Hotdog Clan (CL0050), as well as any associated or subordinate InterPro groups; see Table 2.1 for a list of all the sequence sources used in generating the SSN. Both Pfam and InterPro groups, as well as any further subordinate InterPro groups, were used in generating the SSN.

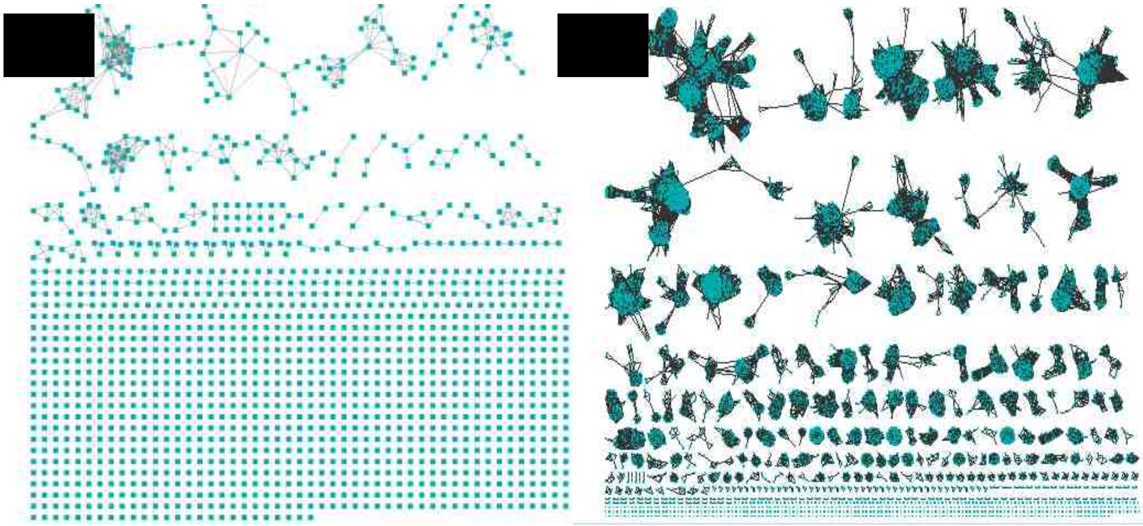
The biocluster on which the scripts were run— hosted by the Institute for Genomic Biology at the University of Illinois at Urbana-Champaign— was accessed using the PuTTy terminal emulator (http://www.putty.org/), the Xming X-window client (http://www.straightrunning.com/XmingNotes/), and WinSCP (http://winscp.net/) for accessing files saved on the server. Both Cytoscape 2.8 and 3.1 were used to visualize and edit protein networks (*25, 26*); images were exported using Cytoscape 2.8.

After the initial network generation via all-by-all BLAST, quartile plots were generated depicting the average and quartile relationships among: percent identity vs E-value, alignment length vs E-value, number of edges vs E-value, and sequence length. For these networks, we used the sequence identity vs E-value quartile plots to identify the E-value cutoff below which sequences would not be considered related. The criteria for choosing a sequence identity/E-value vary depending on the superfamily of interest, but are generally chosen to provide sufficiently high sequence identity suggestive of potentially related function and sufficiently high E-value to result in distinct clustering in the network

(Figure 2.1). Sequence identities of 35-40% are frequently used to cluster isofunctional protein sequences (*17*); in this case, we an average sequence identity of 35%.

### *2.2.2: Representative node networks*

Due to the large number of sequences that can be involved in these networks, Representative Node Networks (RNNs) are often used in place of full SSNs. Full SSNs result in a node for every sequence in the network; thus, if there are five sequence that are all identical, a full SSN would include each sequence as an individual node, each having identical edge relationships to other members of the network. But in a RNN, a percent identity threshold and the clustering program CD-HIT (*27*) are used to cluster 'similar' sequences together into meta-nodes, 'similar' being defined as 'sequence identities above the given threshold. So a 100% RNN would, in the case described above, represent the five identical sequences in a single meta-node. A single meta-node in a 80% RNN would contain an identifying sequence as well as all other sequence IDs that have sequence identities of 80% or higher (Figure 2.2). In general, RNNs are used to simplify very large SSNs that may be too memory-intensive for even high-end computers or in which there are enough very similar or identical sequences that it presents misleadingly large clusters of similar sequences.



[redacted] [redacted]

**Figure 2.2:** Representative sequence similarity network generated for the InterPro thioesterase superfamily collection of sequences (IPR006683, August 2013). A) Representative nodes based on 40% sequence identity clustering. B) Representative nodes based on 80% sequence identity clustering. The full network was too large to be visualized.

### *2.2.3: Network annotation*

Much of the utility of sequence similarity networks and their representative node variants is in the simplicity and speed with which complicated relationships can be visually inspected. This is enhanced by the ability to colorfully annotate these networks with myriad different types of data. For example, the same network can be painted according to taxonomic lineage, experimental function, number of domains, etc.; the only limitation is the information available.

Because the size of the SSNs and the data clustering in the RNNs described in this chapter exceed the data capacities of programs like Microsoft Excel, we developed a Python program, AssignAttributes, to map user-generated annotations to the raw data and keyIDs from a network (see Appendix 2.4). We used this mapped annotation data to paint and/or filter the network.

### *2.2.4: Sequence similarity network annotation data collection*

Activity annotations were acquired using the UniProtKB's manually curated Swiss-Prot database of protein sequences (*28-30*), generally considered a reliable database source of functional annotation (*31*). Swiss-Prot and the UniProtKB's and automatically annotated TrEMBL databases were also used to collect up-to-date information on associated PDB structures and taxonomic lineages. In cases where UniProtKB database information was used for annotations, available annotations were limited to those taken from the ~80,000 up-to-date records, not the entire ~200,000 mapped records, the latter of which included a very large number of records deleted due to redundancy (*32*).

Yajun Wu, a former member of the Dunaway-Mariano lab, conducted a literature search for hotdog-fold thioesterases with experimentally verified function. Yajun identified 58 thioesterases, which were later organized into overall reaction types and mapped; for an overview of the literature search results, see Appendix 1.1.

In addition to the literature search, sequence and structural categorizations done by Dillon and Bateman in 2005 and Pidugu *et al* in 2009, respectively, were used to inform function assignment and/or subfamily membership (*3, 4*). These categorization types were combined with the Swiss-Prot manually curated annotations to develop a new list of ~1600 hotdog-fold enzymes with functional or subfamily assignment. If a single UniProt entry was given conflicting function/subfamily assignment from different categorization types (sequence vs. PDB vs. Swiss-Prot), it was noted as a conflicted entry and was only given a function/subfamily assignment if contextual clues from the sequence similarity network allowed (e.g., it appeared in a large cluster of exclusively PaaI members).

A number of hotdog-fold members were annotated as fluoroacetyl CoA thioesterases (FLK) based on the work of Lucas Zimney, a member of the Dunaway-Mariano lab. Putative FLKs were inspected for key conserved residues and motifs (*33*); those matching FLK criteria were retained as "probable FLKs."

Pfam domain annotations were collected for sequences with current (May 2015) UniProtKB records. Sequence records were inspected for Pfam domain annotations and, if present, the number of different associated Pfam domains. Nodes containing multiple numbers of associated Pfam domains were annotated as such, unless the combination consisted of a single Pfam domain and a single type of multiple Pfam domain (e.g., 1 and 3, 1 and 4, but not 2 and 3).

### *2.2.5: Target selection for high-throughput screening*

The first network generated was an 80% RNN for the thioesterase family (IPR006683, accessed August 2013, Figure 2.2). We used this limited network for target selection in order to maintain a narrower range of probable substrates for HTS screening. We compared the species of each representative node against the 359 taxonomic IDs available for cloning. Targets were refined by eliminating any sequences associated with known function according the literature search above or PDB structures having assigned function, leaving uncharacterized sequences or PDB structures without functional assignments (the latter are referred to as SNFs). This list of target proteins was sent to be synthesized by the EFI protein core lab, headed by Dr. Steve Almo at the Albert Einstein College of Medicine.

Successfully purified proteins underwent high-throughput screening against 50 substrates (Table 2.2) conducted by Tianyang Ji in Dr. Karen Allen's lab in Boston (*34*). HTS results were mapped onto the SSN containing all hotdog-fold members according to the following criteria: low activity = no observable activity, specific activities= activity with 5 or fewer substrates, promiscuous = activity with 6-20 substrates; very promiscuous = activity with 21-47 substrates .

**High-throughput screen substrates**

| **Short chain saturated fatty acids** | **Long chain saturated fatty acids** |
|---|---|
| Acetyl CoA | Palmitoyl CoA |
| n-Propionyl CoA | n-Heptadecanoyl CoA |
| Butyryl CoA | Stearoyl CoA |
| Hexanoyl CoA | Nonadecanoyl CoA |
| **Branched fatty acids** | Arachidoyl CoA |
| Acetoacetyl CoA | Henarachidoyl CoA |
| DL-3-Hydroxy-3-methylglutaryl CoA | Docosanoyl CoA |
| DL-β-Hydroxybutyryl CoA | Tricosanoyl CoA |
| Glutaryl CoA | Arachidonoyl CoA |
| Isobutyryl CoA | Pentacosanoyl CoA |
| Isovaleryl CoA | Hexacosanoyl CoA |
| Malonyl CoA | Diphytanoyl CoA |
| Methylmalonyl CoA | α-hydroxy octadecanoyl CoA |
| Succinyl CoA | **Long chain unsaturated fatty acids** |

β-Methylcrotonyl CoA

**Short chain unsaturated fatty acid**

Crotonoyl CoA

**Medium chain saturated fatty acids**

Decanoyl CoA

Lauroyl CoA

Myristoyl CoA

Octanoyl CoA

Tridecanoyl CoA

Pentadecanoyl CoA

**Derivatives**

04:0 Pyrene CoA

12:0 Biotinyl CoA

16-NBD-16:0 CoA

Palmitoleoyl CoA

(10Z-heptadecenoyl) CoA

(6Z-octadecenoyl) CoA

(9Z-octadecenoyl) CoA

(11Z-octadecenoyl) CoA

Linoleoyl CoA

(9Z, 12Z, 15Z-octadecatrienoyl) CoA

(6Z,9Z,12Z-octadecatrienoyl) CoA

(5Z,8Z,11Z,14Z-eicosatetraenoyl) CoA

(5Z,8Z,11Z,14Z,17Z-eicosapentaenoyl) CoA

Docosahexaenoyl CoA

(15Z-tetracosenoyl) CoA

**Aromatic**

Benzoyl CoA

Phenylacetyl CoA

**Table 2.2:** Coenzyme A substrates used in the high-throughput screenings, sorted according to substrate type.

### *2.2.6: Biological range and gene context of selected proteins*

At the time of this writing, if there is an NCBI BLAST result referencing multiple identical proteins, any records with WC_XX accession numbers are preferentially the first and primary result. To avoid incorrect automatic neighbor calculations, we checked each query sequence for identical sequences in the NCBI database, either by accession number or protein sequence. In the event of an identical record with a WC_XX accession number in the same species (but not necessarily the same strain), the WC_XX record was subsequently used as the query protein; barring an identical WC_XX record, the original query accession number was used as the query protein.  Given the accession number of the query, the 10 numerically adjacent accession numbers were defined as neighbors; e.g., if WC_15 was the query, the neighbors were WC_05-WC_14 and WC_16-WC_25.

An in-house program called ContextBLAST was written using the Biopython package for Python 2.7 (*35*) to determine biological range of the query protein by running a BLAST on the query protein (Appendix 2.3).  In this program, only results above a given percent query coverage (calculated by dividing aligned length by the original query's length) and percent sequence identity are retained.  Default parameters were 30% sequence identity, 70% query coverage, and a limit of 5000 sequences due to computational time; results that did not taper to 30% sequence identity on the 5000th result were expanded to 10000 results.  ContextBLAST then compiles a list of result species based on retained results.  ContextBLAST determines gene context by running a BLAST on each neighbor of the query protein; only neighbor results that matched the query list of result species are retained, and then only above a given percent query coverage and percent sequence identity (again, 70% and 30%, respectively).  For all retained results, ContextBLAST calculates neighbor distance by subtracting the accession numbers of the query and neighbor (e.g., WC_15 and WC_10 are 5 genes apart).  An illustration of this process is shown in Figure 2.3.

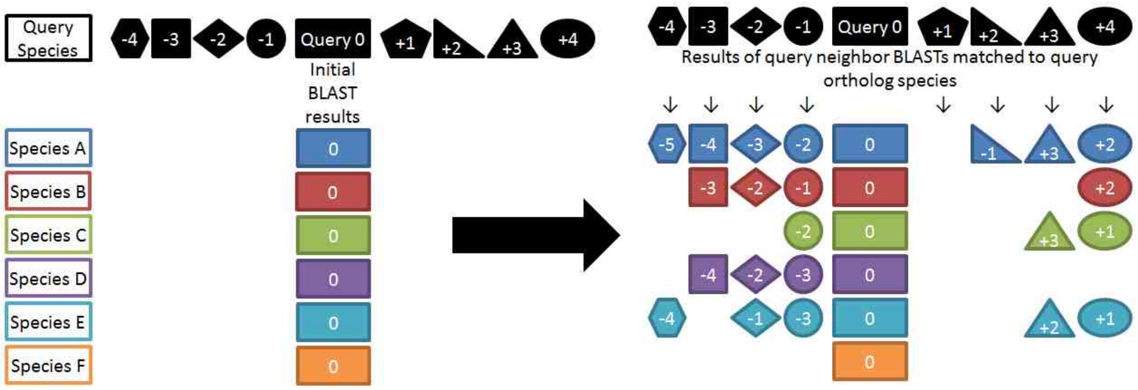


**Figure 2.3:** Initial BLAST results of a query protein result in a list of species containing putative orthologs. Neighbors to the original query each undergo their own BLAST search; any neighbor orthologs belonging to a query ortholog species is compared to the query ortholog in that species to determine whether the two orthologs are still neighbors.

Finally, ContextBLAST compiles BLAST results for all neighbors of the query into a single file and assigns taxonomic lineages via the gi2taxid2lineage program described in Section 2.2.7. An Excel macro imported the neighborhood files for all queries into a single file, in which results were manually color coded based on whether each species contained a potential neighbor ortholog. This composite file was manually inspected for potential gene context. In cases with potential conserved gene context, we assigned gene function based on top hits or consensus (*36*).

As genome sequencing costs decrease, more and more information is available and uploaded to the NCBI databases, including protein sequence information for multiple strains of the same species. As a result, BLAST results include all matching strains of a given species which, for biological range and gene neighborhood purposes, is redundant and may deceptively weight the biological range in favor of the multiple-strained species (current as of May

2015). To combat this, we manually removed such multiple strain cases from our results. We retained the species strain containing the highest SI match to the original query protein and removed all other strains of the same species; if this highest SI strain contained multiple hits, we retained all of the hits.

For visualization and inspection, we generated bar graphs depicting conserved biological range for each query that appeared to have potentially conserved gene context. Unless otherwise noted, biological range and context conservation graphs display only those taxonomic groups with potential neighbor orthologs. If a taxonomic group contains a query ortholog with no neighbor orthologs, that taxonomic group is not displayed.

### *2.2.7: Parsing taxonomic lineages for selected proteins*

Taxonomic lineages were generated from the NCBI taxonomy database; at the time of download (May 13, 2014), it contained taxonomic information for more than 160,000 organisms (*37*). We wrote a Python program called gi2taxid2linaege (Appendix 2.2) to generate taxonomies from the downloadable complete databases of names-to-taxids (names.dmp) and pair-wise relationships between taxonomic ids (nodes.dmp). gi2taxid2lineage mines the names.dmp file to determine whether input queries were represented in the NCBI taxonomy database; specifically, it searches for taxonomic id in the case of sequence similarity networks or for species names and pairs them to taxonomic ids in the case of gene contexts. Because the nodes.dmp file contains only pairwise parent-child relationships (e.g., homo : sapiens, hominidae : homo, primates : hominidae, mammalia : primates, chordata : mammalia), gi2taxid2lineage generates taxonomic lineages by  seeking the parent-child relationship for the query taxonomic id, followed by the grandparent – parent relationship, followed by the great grandparent – grandparent relationship, etc.  The resulting taxonomic lineage lists great grandparent-grandparent-parent-child relationships. Finally, gi2taxid2lineage tabulates taxonomic lineage with the input queries and any data associated with the queries.  We adapted gi2taxid2lineage as necessary to assign taxonomic lineages in other programs described in this manuscript.

### *2.2.8: Outliers in iso-taxonomic clusters*

In cases where a member of one taxonomic group (e.g., Bacteria) appeared in a cluster overwhelmingly belonging to a distant taxonomic group (e.g., Eukaryota), the outlier sequence was further pursued.  If the outlier was a member of a representative node containing other sequences, its co-members were inspected for UniProtKB functional annotations or Pfam membership; any majority or plurality annotations were noted.  The SSN was filtered to remove all edges below 50% sequence identity and UniProtKB annotations were collected for immediate neighbor nodes above this more stringent threshold.  Outlier node co-members and >50% SI neighboring nodes were also inspected for manually curated annotations from the UniProtKB Swiss-Prot database.  The outlier sequence underwent a BLAST search against the NCBI non-redundant sequences database and inspected for high sequence identity relationships to members of its own taxonomic group as well as the dominant taxonomic group of its cluster.

In the case of the multi-domain sequence B8BGK6, its two domains underwent separate BLAST searches (D-Tyr tRNA  deacylase domain = 1-90 aa, hotdog domain(s) = 70-370 aa) to collect taxonomic distribution for each domain individually.  This sequence also underwent a full-length BLAST search limited to its node co-members.

## 2.3: Results and discussion

### *2.3.1:  The general sequence similarity network for the hotdog-fold superfamily*

The hotdog-fold sequence similarity network was preceded by a pilot SSN produced for one of its member families, the thioesterase superfamily (Pfam03061/IPR006683) to identify initial screening targets and network parameters.  For the subsequent SSN of the entire hotdog-fold superfamily, we used an E-value cutoff of $10^{-27}$, corresponding to an average sequence identity of 40%.  This threshold was chosen to reflect an average sequence identity above which clusters are likely to be isofunctional as well as to ensure that the network would exhibit distinct clustering—in the preceding SSN of the thioesterase subfamily, sequences isolated into individual clusters at an E-value threshold $10^{-27}$ whereas less stringent thresholds resulted in ‘hairball’ arrangements, as seen in Figure 2.1 above.

For visualization purposes within this manuscript, a 65% representative node network of the hotdog-fold superfamily is used.  It contains 17,311 representative nodes and 518,447 edges, compared the full SSN with a total of 231,380 nodes and 462,360,392 edges.  Hereafter, the same network will be displayed multiple times, painted according to different annotation schemes.  Due to the large size of the sequence similarity network and for ease of illustration within this manuscript, the SSN is divided into subnetworks A and B. Both subnetworks will be presented as on consecutive pages, except in cases where the results and discussion apply only to clusters within subnetwork A, in which case only subnetwork A will be presented. All clusters are numbered according to their position within the subnetwork; due to space constraints, some clusters are not visibly labeled with their number assignment.

Of the 223,540 InterPro and Pfam sequences represented in the hotdog-fold family network, 1,057 had manually curated annotations in Swiss-Prot and an additional 79,303 belonged to the automatically annotated TrEMBL database (See Table 2.3, current as of May 2015). The rest were either not in the UniProtKB or had been removed, the latter largely due to proteome consolidation efforts—all but 684 of the deleted entries were deleted on or shortly after 4/1/2015, corresponding to the release of UniProtKB's first database version using automatic protein redundancy detection, version 2015_04 (*32*). Of the Swiss-Prot annotations, 101 were annotated only generally (putative esterases or uncharacterized proteins) and were removed from the curated annotation database. Thus, only 1.2% of the consolidated UniProtKB hotdog members have verified or experimentally supported functional annotations; this number is reduced to 0.42% when considering all sequences included in the SSN. These curated functions are mapped onto the SSN in 2.3.3.

| | Number of UniProtKB members |
|---|---|
| **UniProtKB status** | |
| Swiss-Prot | 1,057 |
| TrEMBL | 79,303 |
| Deleted from UniProtKB | 143,118 |
| Not found in UniProtKB | 47 |
| **Domain or kingdom classification** | |
| *Bacteria* | 69,587 |
| *Eukaryota* | 8,540 |
| *Fungi* | 4,720 |
| *Metazoa* | 1,565 |
| *Viridiplantae* | 1,490 |
| *Archaea* | 1,654 |
| Virus | 3 |
| Not specified | 576 |

**Table 2.3:** Distribution of hotdog-fold members according to database membership and Swiss-Prot/TrEMBL taxonomic assignments.

### *2.3.2: Subfamily segregation and domain overlap*

Many of the ~80,000 hotdog-fold sequences with UniProtKB records contained annotations describing the Pfam family/families to which they belong. The most commonly occurring of these family memberships, defined as the 22 families with >200 sequences and >20 representative nodes attributed (Figure 2.4), were mapped onto the representative node network. Nodes containing sequences with membership in multiple commonly occurring families were also noted and mapped.

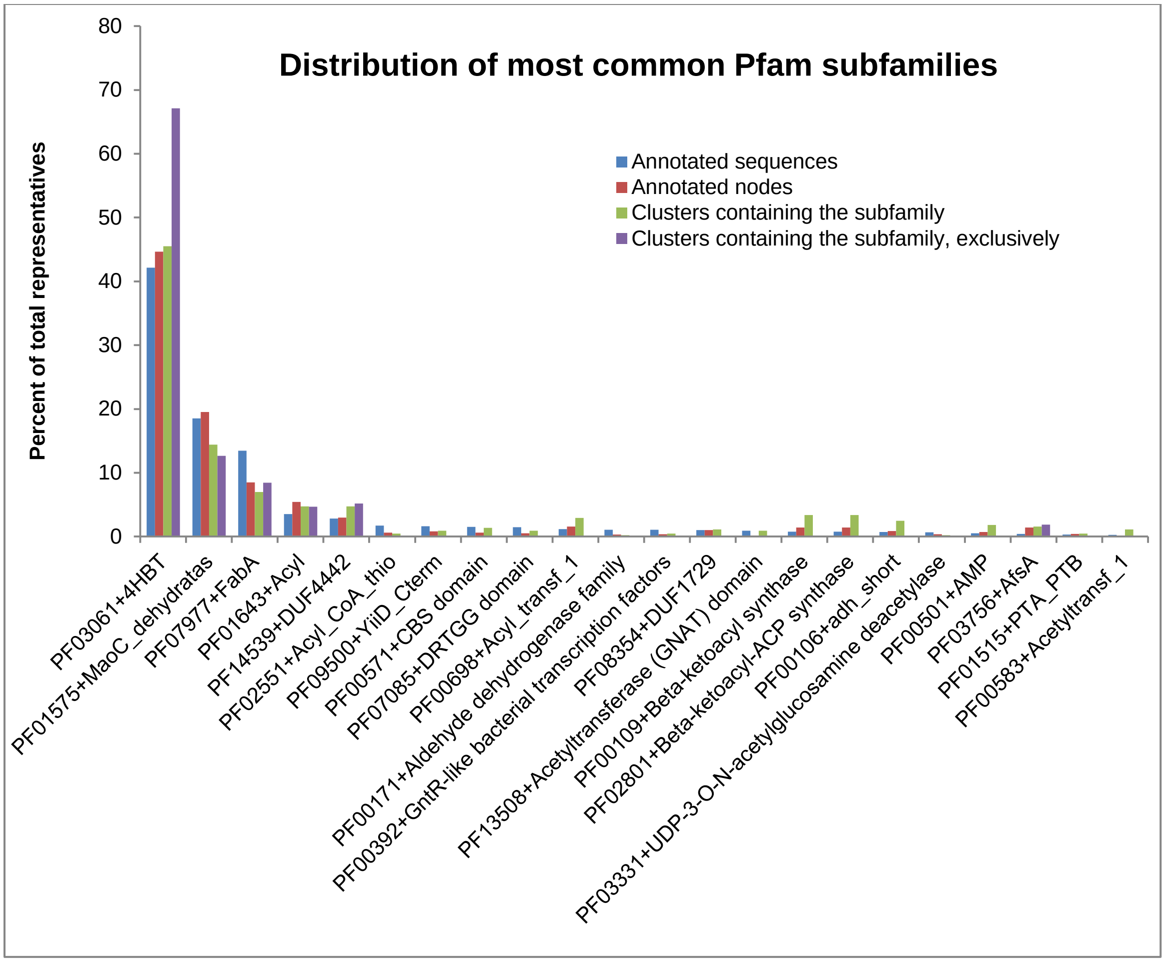
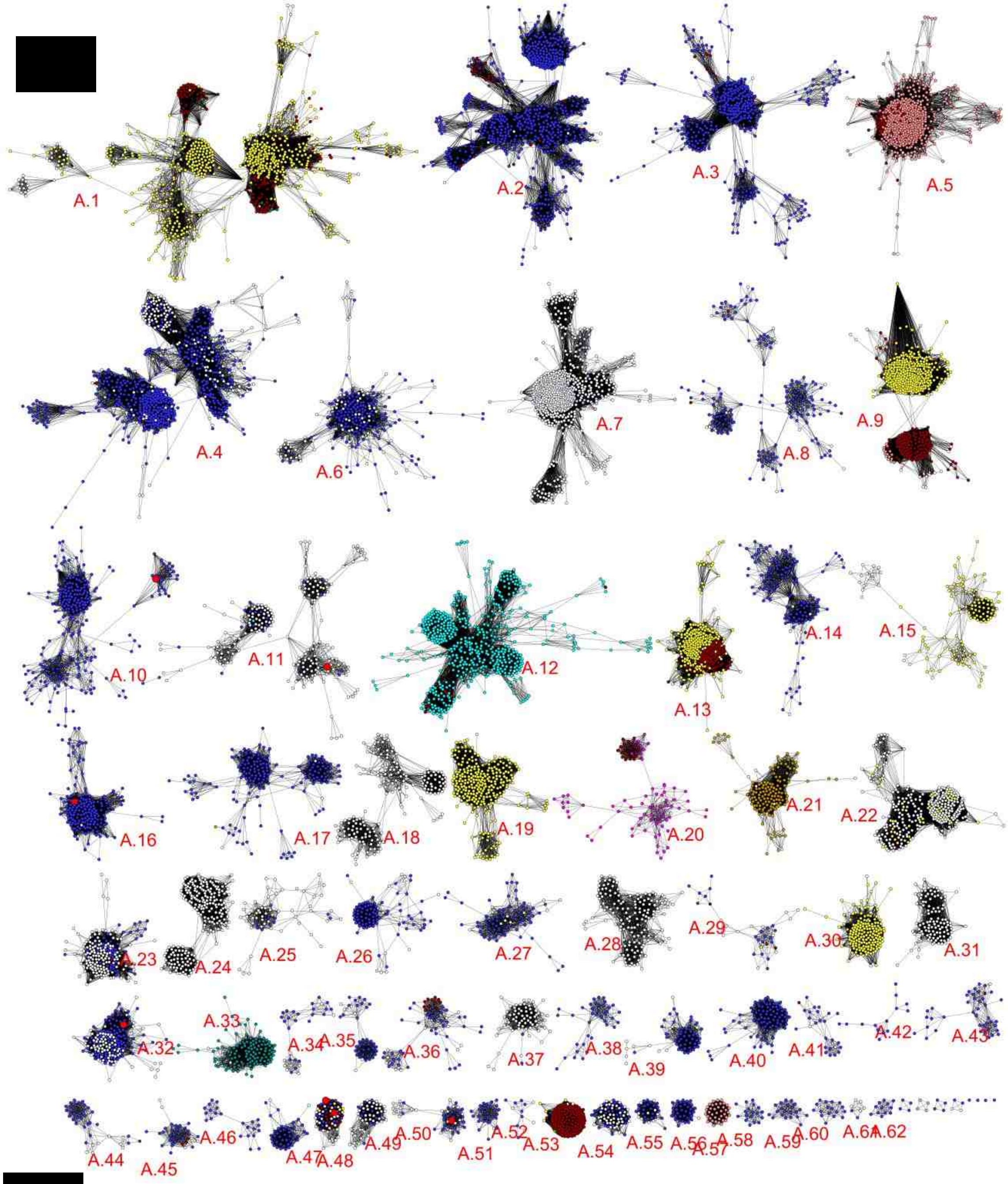


**Figure 2.4:** The 22 most commonly-attributed Pfam families from all UniProtKB records with Pfam annotations. Values are reported as percent of all records with the given Pfam identifier out of all records with any Pfam identifiers. Values are reported for: all sequences with annotations (blue), all representative nodes with annotations (red), clusters containing the subfamily of interest and any additional subfamilies (green), and clusters containing only the subfamily of interest and no additional subfamilies (purple).

The resulting network (Figure 2.5) reveals that most clusters can be assigned to a single subfamily and that several individual families have members spread across multiple clusters. However, there are nonetheless several clusters in which multiple Pfam subfamilies co-occur.

[redacted]



[redacted]

| | | | |
|---|---|---|---|
| ■ | PF14539-- DUF4442 | ■ | PF01515-- PTA_PTB |
| ■ | PF09500-- YiiD C-terminal | ■ | PF00698-- Acyl transferase 1 |
| ■ | PF07977-- FabA | ■ | PF00583-- Acetyltransferase 1 |
| ■ | PF03756-- AfsA | ■ | PF00501-- AMP |
| ■ | PF03061-- 4HBT | ■ | PF00106-- Adh short |
| ■ | PF02551-- Acyl CoA thioesterase | ■ | Contains sequences with which multiple Pfam annotations |
| ■ | PF01643-- Acyl | □ | No Pfam domain annotations |
| ■ | PF01575-- MaoC dehydratase | Enlarged | Contains sequences with single, different Pfam annotations |

[redacted]

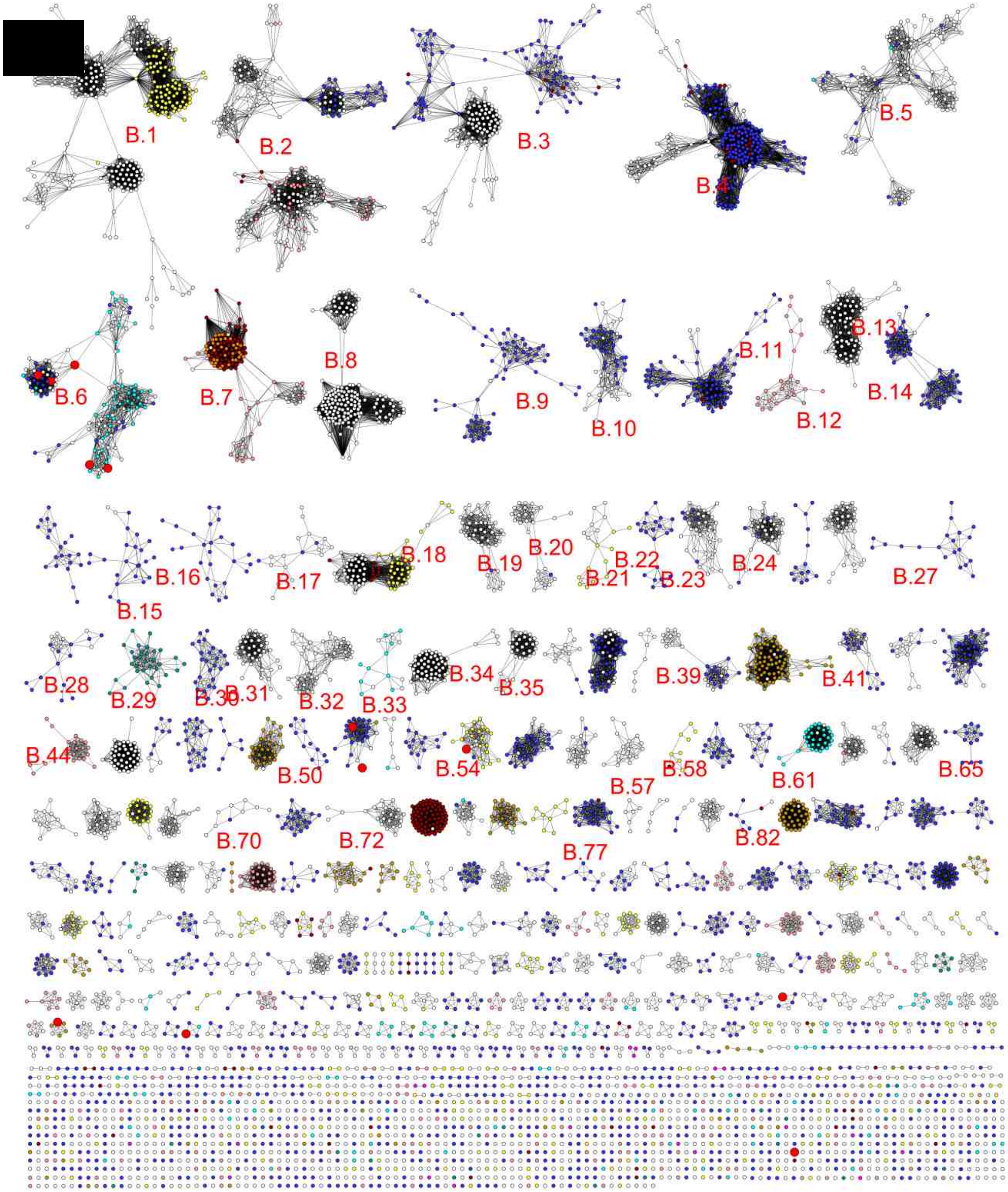


**Figure 2.5:** The hotdog-fold SSN painted according to Pfam domain annotations acquired from the UniProtKB; see the key (B) for color assignments). Only the 22 top most commonly attributed are displayed. Nodes containing multiple sequences each with a different Pfam annotation are bright red and enlarged; nodes containing sequences each with multiple Pfam annotations are

dark red and normal-sized. Nodes containing no sequences with Pfam annotations are not colored. Subnetworks A and B are represented in images (A) and (C), respectively.

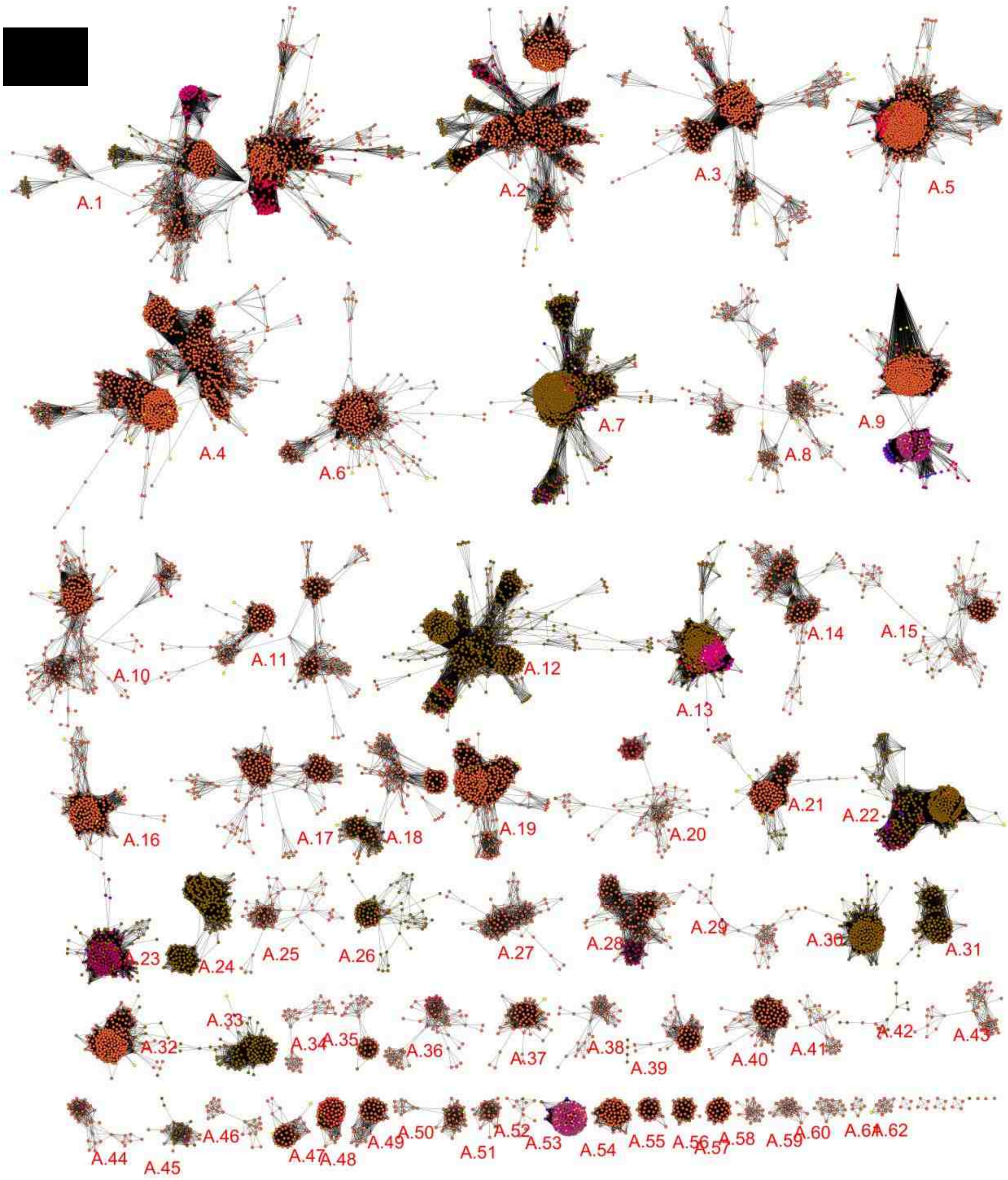
The most widely represented subfamily is the 4HBT family (PF03061), applied to sequences in 202 distinct clusters, including 143 clusters containing only sequences with 4HBT subfamily annotations; the latter may all be annotated as belonging to this subfamily (77,053 sequences total). The next most widespread family is the MaoC dehydratase family, applied to sequences in 64 clusters, including 27 clusters containing only sequences with this annotation, allowing 25,603 sequences to be assigned this subfamily. The FabA family (PF07977) is applied to 30 clusters, including 18 clusters (5,997 sequences) which may be annotated as belonging exclusively to the FabA subfamily. After these subfamilies, there is a sharp drop-off in subfamily assignment: the Domain of Unknown Function (DUF) 4442 and acyl-ACP thioesterase families are each applied to 21 separate clusters (~10 of which contain only annotations for these families) and all subsequent subfamilies are applied to fewer clusters yet.

Unfortunately, most of the aforementioned subfamilies with significant membership are too general to be used for function inference: the 4HBT, MaoC, and FabA subfamilies can all be further subdivided into more specific functional assignments (*3-5*) and DUF4442 has no known function associated. Section 2.3.3: will further demonstrate this generality by assigning more specific functions based on additional literature data. Thus, while applying these general Pfam family memberships to previously uncharacterized sequences is useful in that it narrows the field of membership and potential function, it is not sufficient to assign detailed function. In such cases, additional information is necessary to provide more detailed functions.

The fairly common acyl-ACP thioesterase subfamily (PF01643) contains thioesterases acting on acyl-ACPs; such thioesterases terminate fatty acyl synthesis by hydrolyzing an acyl group on a fatty acid. Clusters containing primarily or exclusively sequences with this subfamily annotation (A.12 being the largest such cluster) may be assigned fatty acid synthesis biological function and should be expected to have strong activity with acyl-ACPs. The AfsA subfamily may also be used to immediately assign biological function: clusters A.33, B.29, B.91, and B.188 contain sequences with this subfamily annotation exclusively. Members annotated with this family, including the aforementioned clusters, may be assigned biological function related to A-factor biosynthesis (*38*).

Several clusters contain only sparse subfamily annotations from the Pfam database. Internal length comparisons among cluster members can be used to verify the likelihood that cluster members share subfamily assignments (Figure 2.6). Because clusters necessarily share sequence identities above 40% due to the cutoff used for the SSN creation, similar sequence lengths indicate additional homology. Thus, as long as all members of a cluster are of similar lengths and the nodes with family annotations are in consensus, it is possible to infer that all members will share the annotated subfamily membership.

[redacted]



[redacted]

wide range of lengths
<100 aa
+100 aa
+200 aa
+400 aa
+600 aa
+1000 aa
+2000 aa
+4000 aa
+8000 aa

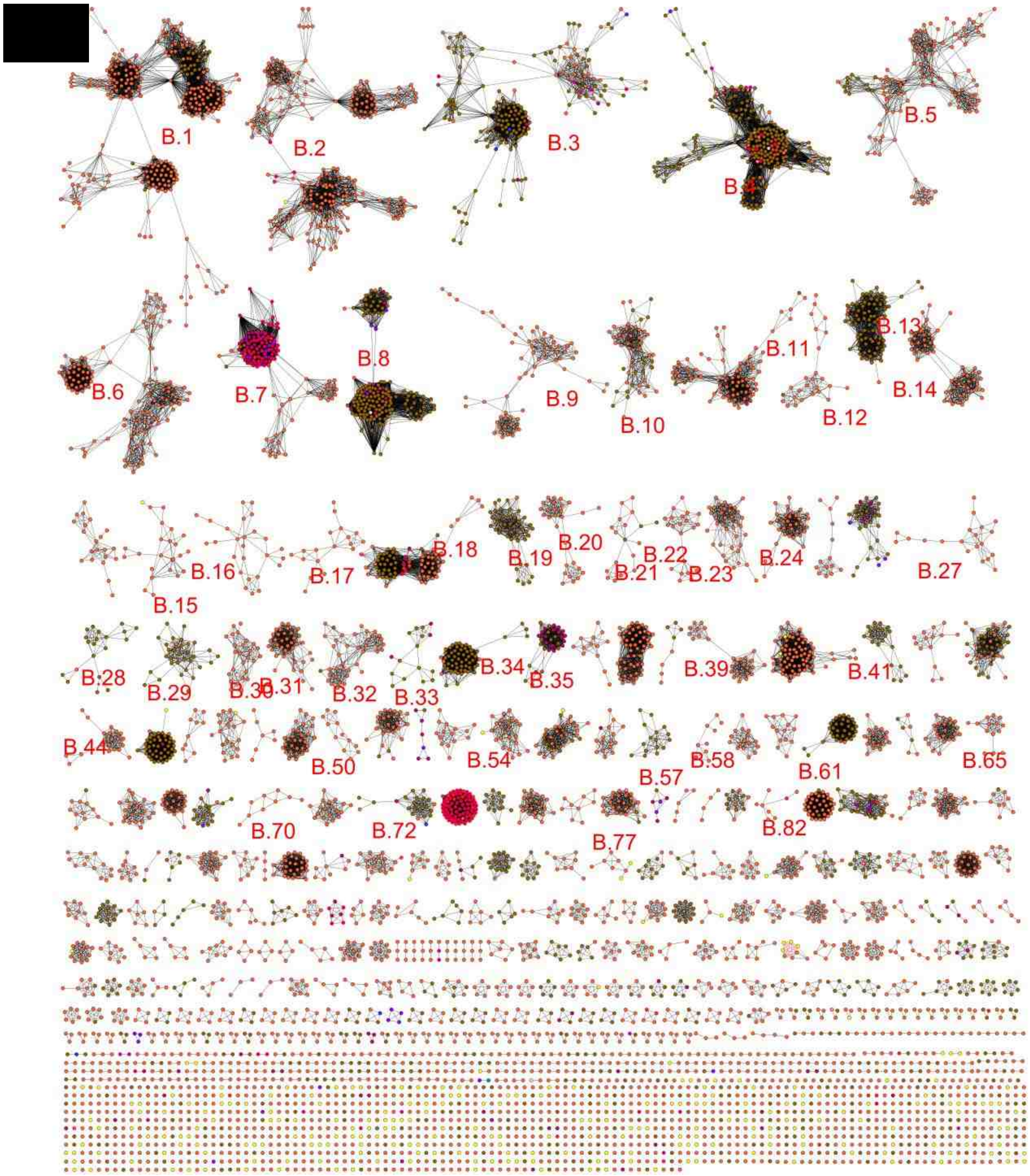

[redacted]

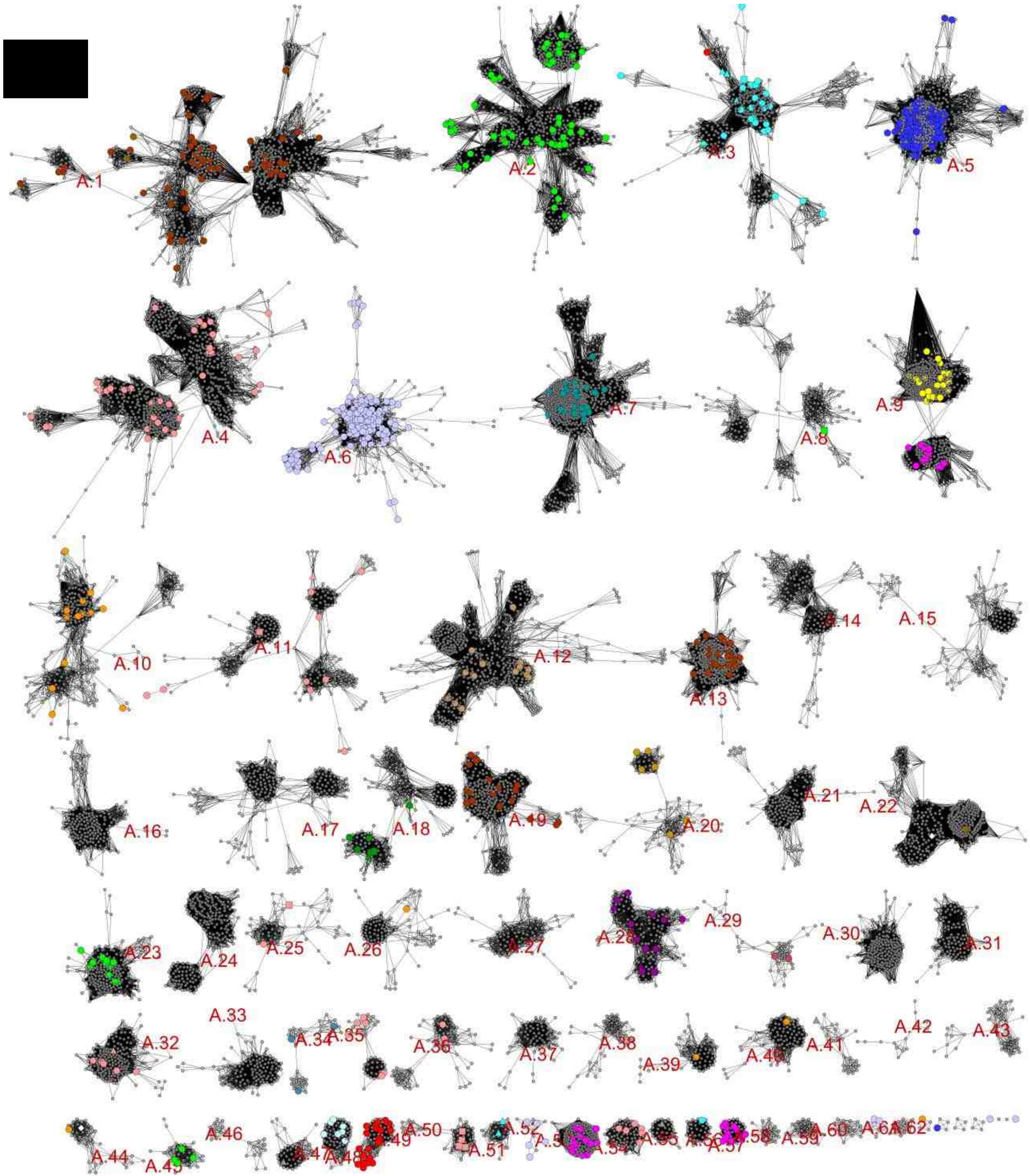
**Figure 2.6:** The hotdog-fold SSN painted according to average length of node contents, based on length annotations for each sequence generated upon network creation; see key (B) for color assignments. Nodes containing multiple Pfam domains have thickened, magenta borders. Subnetworks A and B are represented in images (A) and (C), respectively.

The remaining major subfamilies co-occur with other domains. As described above, hotdog-fold subfamilies are known to associate with other domains in fusion proteins; this tendency is borne out in the network, with a number of clusters containing regions of sub-clusters with multiple Pfam family annotations. This domain co-occurrence can be used to infer function if the domains are known to be associated with a particular biological function. Such instances and their use for functional annotation will be discussed later in Section 2.3.3.

### *2.3.3: Mapping published results to predict subfamily, structure, and function*

As described above, in 2005, Dillon and Bateman categorized proteins known to belong to the hotdog family into 17 subfamilies (948 hotdog members assigned) and 85 distinct clusters (including an additional 345 hotdog members in 66 clusters without subfamily assignments) based on sequence similarities and hidden Markov modeling (*3*). Pidugu *et al* expanded on this categorization and demonstrated that the then-known hotdog-fold structures (~60 in 2009) could be categorized into several general quaternary structures (*4*). We combined these categorizations with the manually annotated Swiss-Prot function annotations described above, literature search results, and in-house fluoroacetyl-CoA annotations, resulting in what we will refer to as “consensus annotations” of function and/or subfamily for 2,057 members of the hotdog-fold SSN. These consensus annotations, completely independent from the UniProtKB Pfam subfamily annotations described in Section 2.3.2, are painted on Figure 2.7 and discussed below.

[redacted]



- 🟥 [redacted]inobutyryl-CoA ammonia lyase
- 🟦 4HBT-I
- 🟦 4HBT-II (triangle with DHNA-CoA, square with EntH)
- 🟨 Acetyltransferase
- 🟩 Acyl-CoA thioesterase (square with PaaI)
- 🟥 AMP-binding subfamily
- 🟥 DHNA-CoA
- 🟪 FabA
- 🟦 FabZ
- 🟥 YbgC-like, YbaW-like (square), or both (triangle)
- 🟪 FLK
- 🟪 Hydroxyacyl-CoA dehydrogenase, L-carnitine dehydrogenase (square), or both (triangle)
- 🟫 MaoC-like
- 🟩 Mesenchymal stem cell protein, THEM6 (square), or both (triangle)
- 🟫 Mesaconyl-CoA hydratase
- 🟨 NodN
- Other
- 🟧 PaaI
- 🟦 TesB
- 🟫 Fat subfamily (acyl-ACP thieosterases). Dodecanoyl-specific (vee), oleoyl-specific (rectangle), palmitoyl-specific (triangle)

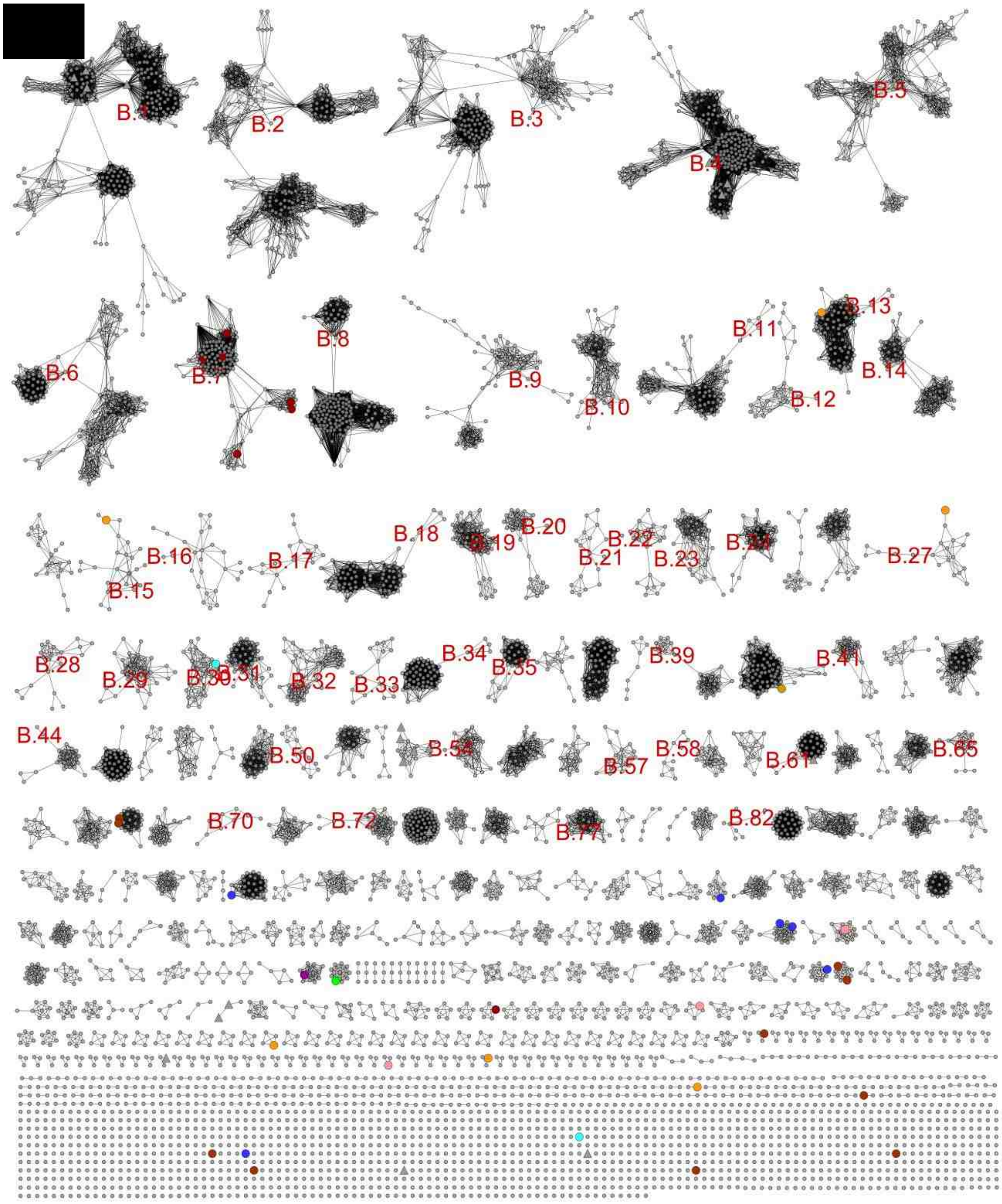

[redacted]

**Figure 2.7:** The hotdog-fold SSN painted according to consensus annotations from the combination of literature searches, the Dillon/Bateman and Pidugu et al reviews, general categorization of Swiss-Prot annotations, and in-house FLK assignments. All nodes with consensus annotations are enlarged; see the key (B) for color and node shape assignments. Subnetworks A and B are represented in images (A) and (C), respectively.

This consensus annotation SSN further demonstrates the subfamily co-clustering behavior initially described in Section 2.3.2. Furthermore, this level of annotation provides more detail than the initial Pfam-based annotation. For example, several of the clusters originally annotated as belonging to subfamily 4HBT are shown to have their own distinct functions. Indeed, of the clusters with both consensus annotations and Pfam annotations of 4HBT, only clusters A.3, A.34, A.51 and A.56 retain 4HBT annotations and even their 4HBT annotations have been further refined (A.3 and A.56 to 4HBT-II, A.34 to 4HBT-I, and A.51 to 4HBT-II with a 1,4-dihydroxy-2-naphthoate-Coenzyme A, or DHNA-CoA, domain). The other clusters are given more specific annotations, as described in Appendix 1.2.

As with the Pfam subfamily annotations, most consensus annotations cluster exclusively with their fellows. However, there are several islands of multi-domain or multi-consensus sequences within clusters. Likewise, instances of different-domain clustering tend to be contained to small, tightly localized areas, several of which can be used to infer function when the additional domains have Pfam- or consensus-assigned subfamilies. Prime examples of this can be found in clusters A.1, A.2, A.3, and A.9.

There are several interesting cases where neither of the above are true and domain organization is chaotic. In these cases, multiple different classifications of domains and numbers of involved families cluster together, disallowing application of a single annotation across the entire cluster. Clusters A.48. B.6, and B.18 are prime examples of this.

#### *2.3.3.1 Automatic subfamily annotation lacks some nuance and additional subfamily data*

While Pfam subfamily annotations can be used to annotate previously uncharacterized sequences, attempting to do so for the entire network reveals significant gaps in the automatic annotations from the UniProtKB. 52 clusters containing 8,704 sequences have no Pfam annotation data whatsoever and additional clusters contain regions lacking Pfam annotation. While some of these regions can be annotated by published data as described in Section 2.3.3.3 (clusters A.18, A.24, A.31, B.13, B.143, and B.158), the majority of these remain completely uncharacterized. This may be because they belong to heretofore uncharacterized subfamilies. However, as indicated in the cases where they may be annotated based on other data, this emphasizes that the Pfam database does not have the 'full picture' of the hotdog-fold superfamily and its subfamilies. This is particularly noted for cases such as FLK, DHNA-CoA, and NodN, which are not assignable based on Pfam but are readily assignable from literature results.

The mapping of consensus annotation also reveals that Pfam's subfamily annotations are too broad. As noted above, much of the network is annotated as belonging to the 4HBT (PF03061) or MaoC-dehydratase (PF 01575) subfamilies. However, several clusters with these annotations can be given narrower annotations based on literature data, discussed below. Thus, more specific annotations can be made, even automatically, based on existing data, as shown; the result of this can be seen in the full table of assignments in Appendix 1.2.

Dillon and Bateman's 2005 categorization and creation of a Hidden Markov model (HMM) library for subfamily determination would be particularly useful for providing Pfam annotations with additional nuance. However, most superfamilies do not benefit from such large-scale analyses; thus, the overly general Pfam annotations would nonetheless persist on a global scale.

#### *2.3.3.2 Subfamily, function, or structure assignment from single annotation types*

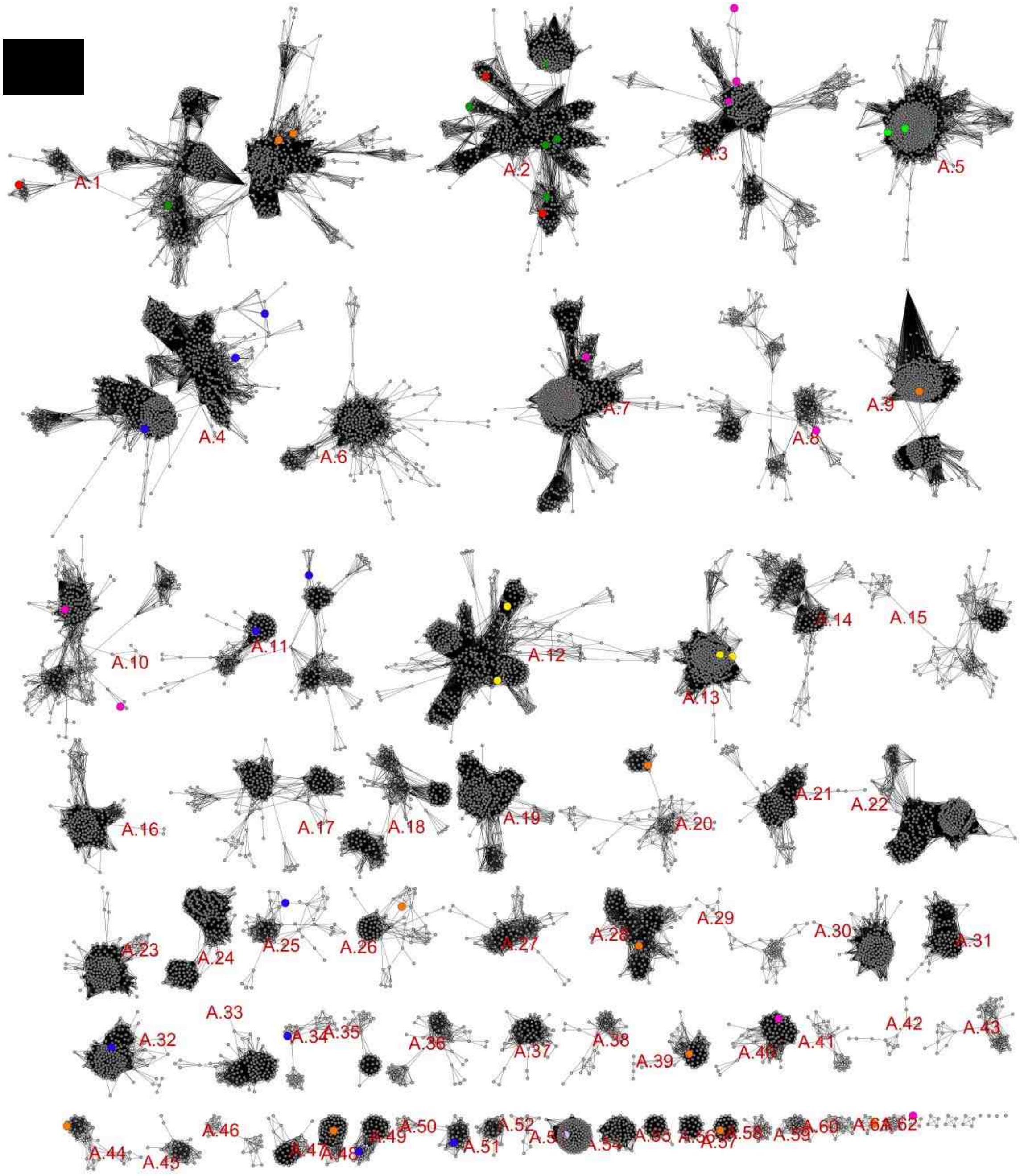
The sequence similarity network was generated using an E-value cutoff of $10^{-27}$ corresponding to ~40% sequence identity, which is often used as a lower bound for identifying isofunctional sequences. As such, if no discrepancies among cluster members exist, annotations to a single sequence in a cluster may be reasonably applied to unannotated members of the same cluster. We can thus use the subfamily and structural assignments to predictively assign subfamily and structure memberships to entire clusters. In this manner, we were able to tentatively assign subfamilies to 9 of the 66 'unknown' subfamily clusters derived by Dillon and Bateman and apply the 85 HMM subfamilies to an additional ~163,000 sequences. Though Pfam subfamily annotations are automatically applied and are typically more general than annotations acquired from the consensus, as discussed above, some clusters possess no other types of annotation data. In these cases, the cluster is annotated based on its Pfam subfamily alone, the identity of which is frequently either 4HBA or MaoC-like. Appendix 1.2 summarizes function, subfamily, and quaternary structure annotations applied to clusters or cluster portions according to the available annotation sources.

Certain discrepancies call such annotations into question and result in the outlier nodes not being annotated with the rest of the cluster or, if enough discrepancies occur with enough of a cluster's nodes, result in the cluster remaining unannotated.  Length discrepancies are the most common issue and occur when a cluster contains nodes of notably varying length (Figure 2.6). Hotdog-fold members are known to fuse with other domains as well as contain duplicate hotdog sequences, which results in length increases (*5*).  Significantly varying sequence lengths may indicate such an event which may itself result in acquisition of new function, changes in quaternary structure, or regulatory modifications, any of which would invalidate annotation applications.  For the same reason, varying numbers of domains within a cluster also disqualify a sequence from being annotated with its cluster. Conflicting Swiss-Prot, literature, or quaternary structure annotations also result in an annotation not being applied to an entire cluster, unless the conflicting annotations are shown to occur in distinct sub-clusters within the cluster as a whole; Cluster A.1 is a prime example in which subfamily co-occurrence varies across the entire cluster but is internally consistent within smaller sub-clusters.

Because of the strong relationship between structure and function, sequences with high sequence identity are expected to not only be isofunctional but also largely isostructural as well.  Internally consistent numbers of Pfam domains (Appendix 1.3) and sequence length within a cluster further suggest minimal variation in sequence lengths and domain types, as discussed above. Thus, unless otherwise specified, quaternary structure assignments (from Pidugu *et* al) of a single node are applied to the entire cluster to which the annotated node belongs, as noted in Appendix 1.2. The exception to this quaternary structure annotation application is instances in which a single cluster has multiple quaternary structure annotations or is itself divided into clear sub-clusters. In the first case, clusters A.1 and A.2 have multiple different quaternary structure annotations that are seen to apply to specific sub-clusters; thus, the quaternary structure annotation is applied only to those select sub-clusters. In the second case, clusters A.9, A.20, and A.34 are also divided into sub-clusters but have only one quaternary structure annotation; thus, the single annotation is applied only to its fellow sub-cluster members.

Cluster A.2 is particularly interesting for its multiple quaternary structures. While most of the quaternary structure-assigned nodes localize in individual sub-clusters, two different quaternary structures co-localize in the same lower branch. However, the two quaternary structures both result in thehelix-interface hexamer with the general structure of H2, the only difference being that one is a hexamer formed from homodimers whereas the other (Trdh) is a trimer formed of double hotdogs. Cluster A.2 has regions in which the homodimer has been fused to a single sequence. The same occurs in Cluster A.7 with the ß-sheet-interface tetramer formed from homodimers vs doublehotdog structures.

[redacted]



[redacted]

| | | | |
|---|---|---|---|
| 🟥 | Trdh | 🟩 | H2 |
| 🟪 | TB | 🟩 | H1 |
| 🟦 | TA | 🟪 | DdhB |
| ⬜ | T+ | 🟨 | DdhA |
| 🟦 | H3 | 🟧 | D |

[redacted]

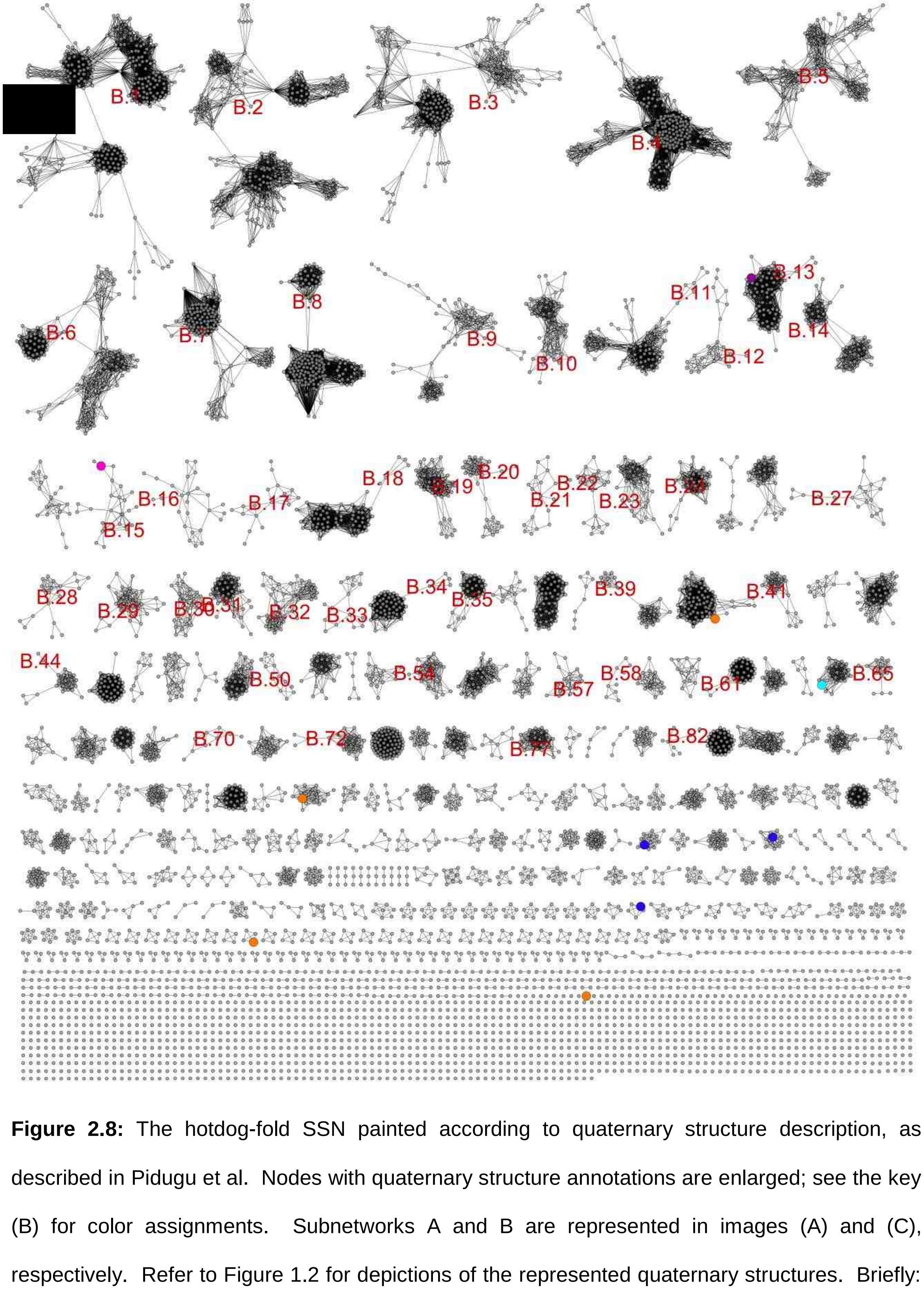


**Figure 2.8:** The hotdog-fold SSN painted according to quaternary structure description, as described in Pidugu et al. Nodes with quaternary structure annotations are enlarged; see the key (B) for color assignments. Subnetworks A and B are represented in images (A) and (C), respectively. Refer to Figure 1.2 for depictions of the represented quaternary structures. Briefly: dimer (D), double hotdog with dimer-like structure (dh), loop-interface tetramer similar to a dimer

of double hotdogs (TA/DdhA), β-sheet-interface tetramer (TB/DdhB), loop-interface hexamer (H1), helix-interface hexamer similar to a trimer of double hotdogs (H2/Trdh), and end-to-end interface hexamer (H3).

### *2.3.3.3 Subfamily and function assignment from multiple sources*

As described above, some clusters are annotated based on co-clustering with Pfam subfamilies or already characterized subgroups from our assembled consensus annotations. Described herein are clusters given assignments based on multiple sources, especially the inclusion of Swiss-Prot annotations (Figure 2.9). The full table of assignments for all clusters, including the number of nodes and sequences affected, can be found in Appendix 1.2.

Cluster A.1 is subdivided into multiple regions, though all sequences belong to the general MaoC-like subfamily. Two small offshoots are assigned mesaconyl-CoA hydratase function based on Swiss-Prot annotations and internally consistent lengths. An upper region co-occurs with Aldehyde dehydrogenase (PF00171) domains, suggesting that it is involved in PHA biosynthesis (*3*); this is further supported by the presence of a Swiss-Prot-annotated PaaZ protein in this sub-cluster. The lowest region on the right-most sub-cluster co-occurs with phosphate acetyl/butaryl transferase (PF01515) domains, which are involved in transfers of acetyl or butaryl groups onto orthophosphate (*2*). A small subsection of the right-most cluster co-occurs with short-chain dehydrogenase (PF00106) domains, which would suggest involvement in hormone biosynthesis if sterol carrier protein domains (PF00188) also co-occurred (*3*); lacking the SCP domain, the biological function of this subsection is unclear.

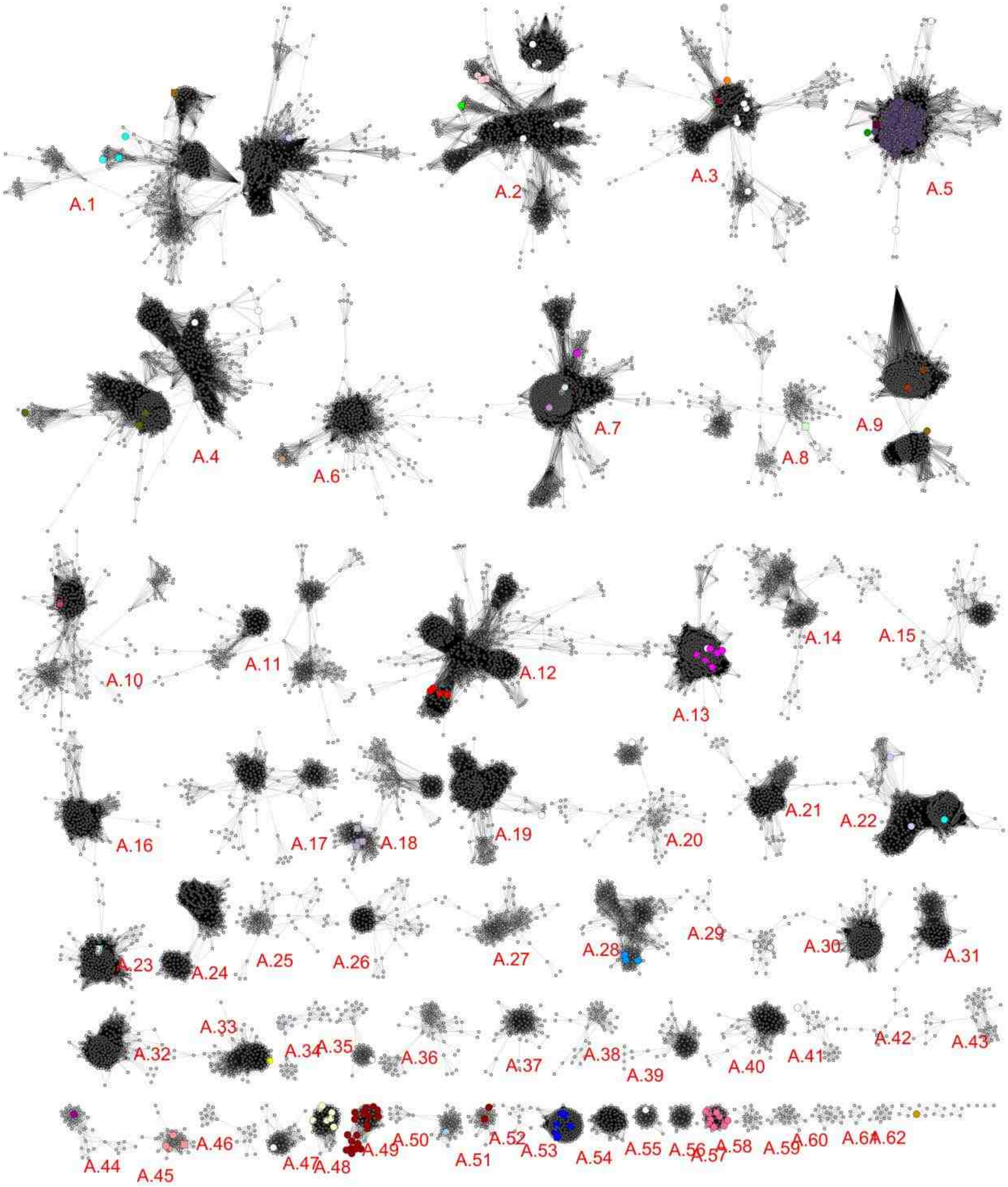


**Figure 2.9:** The hotdog-fold SSN painted according to curated Swiss-Prot annotation, excluding general annotations such as "putative esterase" and "uncharacterized protein"; only Subnetwork A is shown as Subnetwork B contains no meaningful Swiss-Prot annotations. Annotations are condensed when applicable; e.g., putative NodN and NodN annotations are both given a "NodN" annotation. Nodes with Swiss-Prot annotations are enlarged; color and shape assignments are described in Table 2.4, below.

|  |  |  |  |
|---|---|---|---|
| ■ | DHNA-CoA hydrolase | ■ | Cytosolic acyl coenzyme A thioester hydrolase |
| ■ | Acyl-coenzyme A thioesterase PaaI | ■ | Dodecanoyl/oleoyl/palmitoyl-ACP hydrolase, chloroplastic |
| ■ | Beta-methylmalyl-CoA dehydratase | ■ | Peroxisomal enoyl CoA, epimerase, or multifunctional |
| ■ | Bifunctional enzyme LpxC with FabZ | ■ | Hydroxyacyl-thioester dehydratase II mitochondrial 3-hydroxyacyl-ACP dehydratase |
| ■ | Mesaconyl-CoA hydratase | ■ | Methylthioribose-1-phosphate isomerase |
| ■ | Coronafacic acid dehydratase | ■ | Outer membrane protein assembly factor BamA |
| ■ | FadM | ○/□ | Polyketide synthase PksN / MaoC protein with bifunctional PaaZ |
| ■ | L-carnitine dehydrogenase | ■ | Probable A-factor biosynthesis enzyme |
| ■ | Nodulation protein N | ■ | Proofreading thioesterase EntH |
| ■ | Fluoroacetyl-CoA thioesterase | ○/□ | FAS-β / sterigmatocystin biosynthesis FAS-β |
| ○/□ | THEM4/5 | ○/□ | Acyl-coenzyme A thioesterase 12/BFIT |
| □ | THEM6 | ○/□ | Acyl-coenzyme A thioesterase 8/13 |
| ■ | Transcription factor FapR | □/○ | Mitochondrial acyl-coenzyme A thioesterase 9/9 and 10 |
| ■ | YbgC | ■ | 3-aminobutyryl-CoA ammonia lyase |
| ■ | 4-HBA-CoA thioesterase | ■ | 3-hydroxyacyl-ACP dehydratase FabZ |
| ■ | Acyl-CoA thioesterase 2 | ■ | 3-hydroxydecanoyl-ACP dehydratase |

**Table 2.4:** Key of node coloration and shape for Figure 2.9.

Cluster A.2 is annotated as containing acyl-CoA thioesterases, a general subfamily. The small, upper-left sub-cluster belonging to Metazoan species contains cytoplasmic acetyl-CoA hydrolases (CACH), including brown fat-inducible thieosterases (BFIT). This is confirmed by several Swiss-Prot annotations, the presence of expected additional START-domains (PF01851), and a lack of bacterial or archaeal sequences, as is expected for enzymes limited to mammals. Another small Metazoan cluster is annotated as brain acyl-CoA thioesterase (BACH) or acyl-CoA thioesterase (ACOT) 7 on the basis of taxonomy and a Swiss-Prot annotation.

Cluster A.3 contains general 4HBT-II sequences, with the exception of the small off-shoot in the top left section. This small section contains sequences with an additional HAD domain (PF08282), which has been observed in 4HBT-II sequences belonging to *B. thetaiotaomicron* (*3*). This small sub-cluster also contains a small number of multi-domain sequences combining 4HBT-II and DHNA-CoA domains, the function of which is unclear.

Cluster A.5 contains FabZ dehydratases involved in fatty acid biosynthesis, corroborated by a large number of FabZ Swiss-Prot annotations (568 sequences out of 17572 total sequences in the cluster). Additionally, several proteins in this cluster co-occur with LpxC domain (PF03331), expected of FabZ dehydratases, specifically those involved in fatty acid biosynthesis (as opposed to coronafacic acid dehydratase, discussed below in A.64). The H1 quaternary structure can be applied to the majority of the cluster except the multi-domain LpxC region due to its multi-domain component.

Cluster A.6 contains the majority of FLKs from the hotdog-fold superfamily, as determined by Swiss-Prot annotations and in-house investigation. Two other small clusters contain the rest of the hotdog-fold FLKs (A.52 and A.61).

Cluster A.9 contains two large sub-clusters belonging to different subfamilies. The top cluster contains single domain NodN sequences. The bottom cluster contains FabA sequences, specifically associated with polyunsaturated fatty acid biosynthesis. The latter is confirmed by the presence of BKAS-N- and C-terminal domains and various other domains—this cluster contains myriad different domains, but the BKAS domains are the most common ones in addition to the basic FabA hotdog domain.

Cluster A.10 contains PaaI proteins involved in phenylacetic acid metabolism. In addition to a large number of members belonging to the Dillon PaaI cluster, there is also a PaaI Swiss-Prot annotation.

Cluster A.12 contains fat subfamily acyl-ACP thioesterases, confirmed by several Swiss-Prot annotations as well as many members belonging to associated Dillon cluster 6. Interestingly, the Swiss-Prot annotations are the only ones with an additional domain (acyl ATP thioesterases associated with Swiss-Prot chloroplastic proteins, PF 12590) and also belong to Viridiplantae; the rest belong to bacteria.

Cluster A.13 contains MaoC dehydgrogenases, likely involved in hormone biosynthesis. The function annotation is based on co-occurrence with ADH short and SCP domains (PF00106 and 02036) and annotations of peroxisomal hydratase dehydrogenase epimerase, at least on the right part of the cluster belonging to eukaryotes.

Cluster A.18 contains mesenchymal stem cell proteins and/or THEM6, at least in the Eukaryotic sections. The latter is verified by Swiss-Prot. Interestingly, only the Eukaryotic section has longer sequences—the rest is all ~100 aa, but the eukaryotic piece is longer.

Cluster A.22 is tentatively annotated as containing mesaconyl-CoA hydratases in the right hemisphere and mitochondrial hydroxyacyl-thioester dehydratase type 2 on the left, both from Swiss-Prot. However, these are tentative assignments and should be confirmed by more information—there is not enough Swiss-Prot, Pfam, or literature annotation information to assign these, especially given that this cluster is multi-taxonomic.

Cluster A.23 contains acyl-CoA thioesterases, particularly mitochondrial ACOT9 and 10 from Swiss-Prot annotations. Extra domains are likely due to different eukaryote kingdoms but may still receive the same annotation due to the additional annotations.

Cluster A.28 contains 3-hydroxyacyl-CoA hydrogenases (3HCDH), involved in fatty acid metabolism, as well as a small sub-cluster containing more specialized L-carnitine dehydrogenases. The latter activity is assigned based on Swiss-Prot annotations in the lower sub-cluster. The general 3HCDH activity is further confirmed by the association of additional 3HCDH NAD-binding domains (PF02737) for the sequences with Pfam annotation data, which is expected of this class of enzymes (*3*).

Cluster A.29 contains 3-aminobutyryl-CoA ammonia lyases, based entirely on Swiss-Pro annotations. This annotation is applied across the entire cluster because it is internally consistent with regards to length, taxonomy, and domain content.

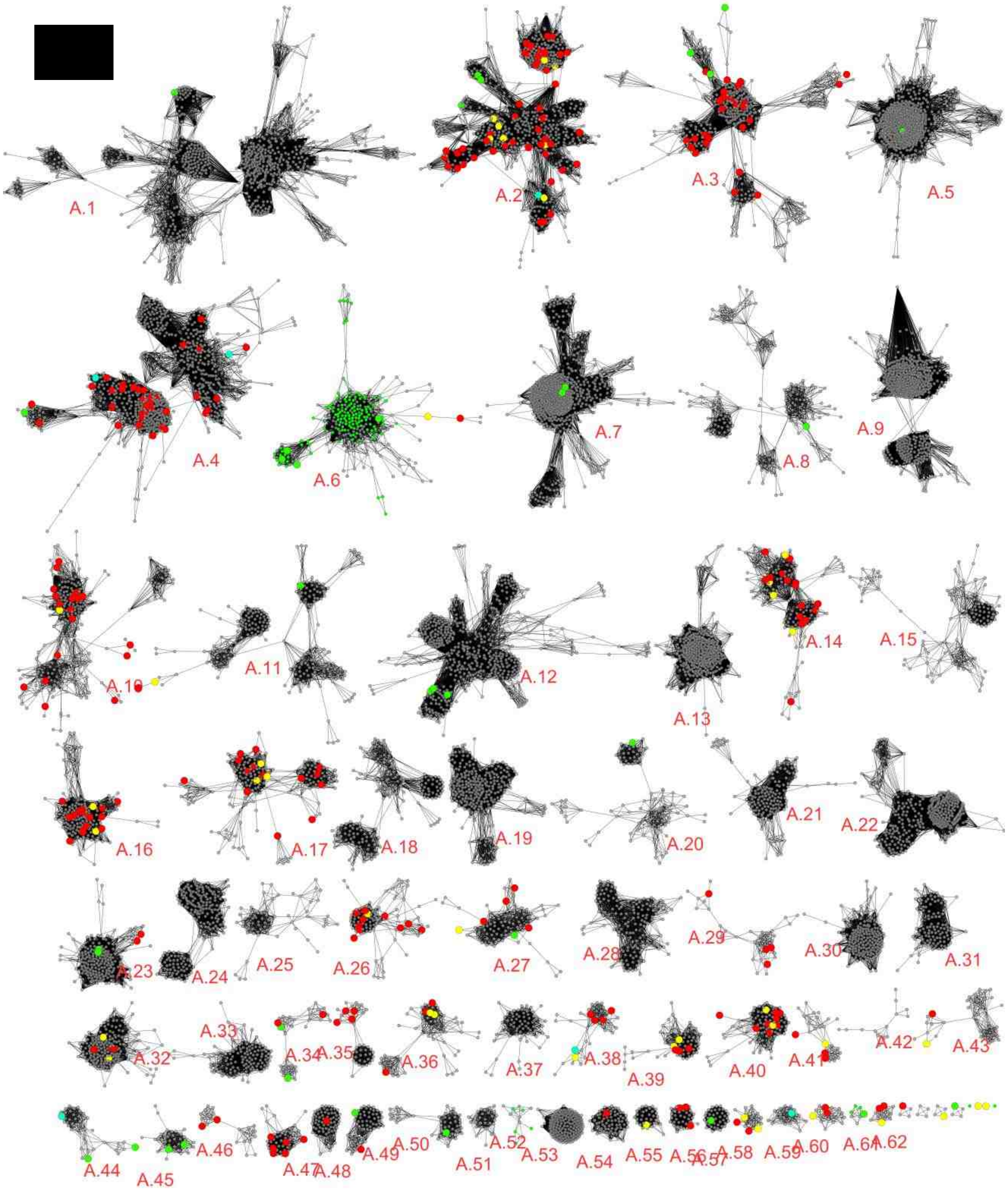
Cluster A.33 contains A-factor biosynthesis enzymes, based on Pfam and Swiss-Prot annotations. These are essential for streptomycin production and resistance *(38)*. Cluster A.36 contains sequences assigned as ybgC-like according to application of the Dillon and Bateman cluster assignments. A small region co-occurs with acetyltransferase 1 (PF00583) or 7 (PF13508) domains.

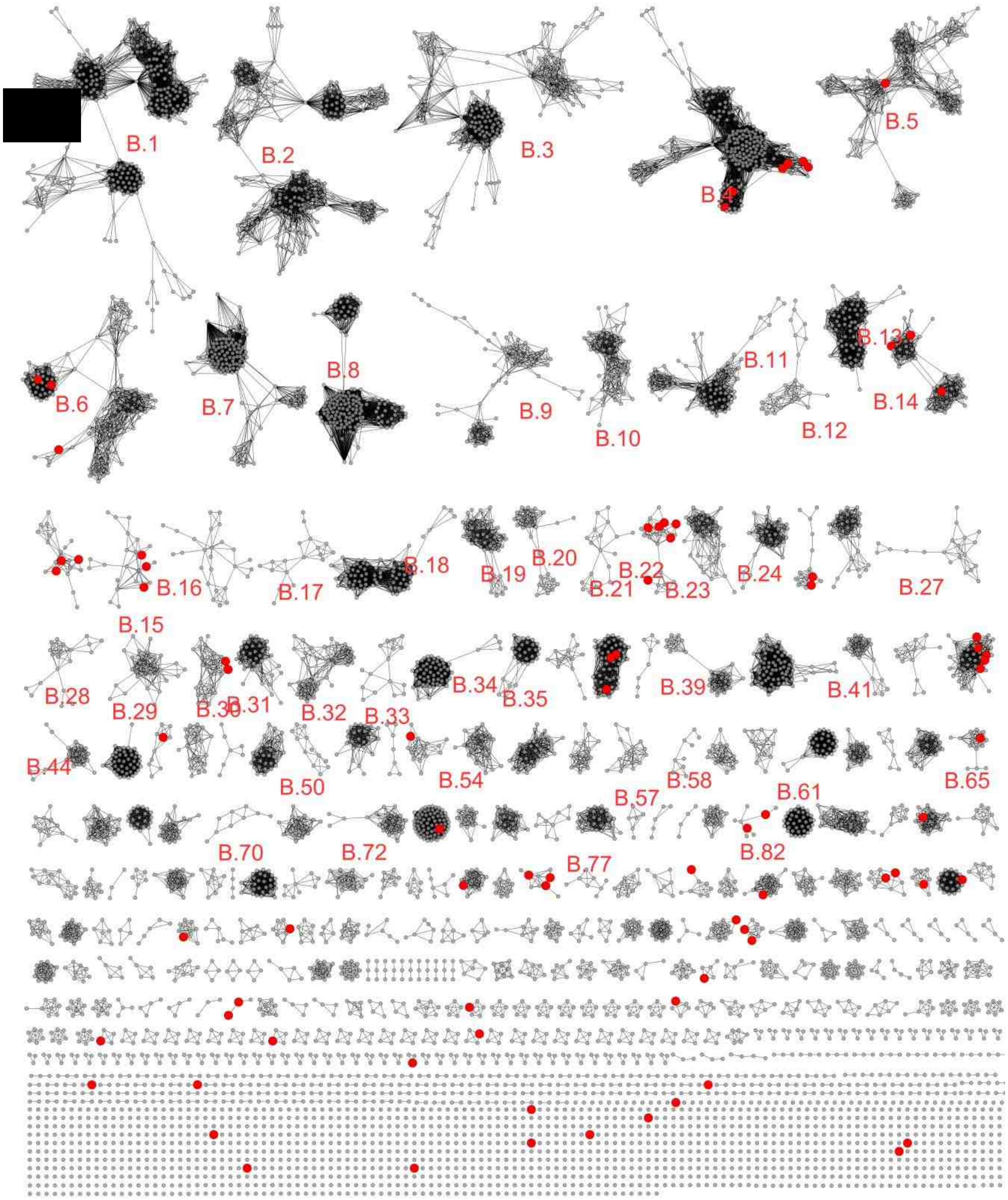
Cluster A.49 represents the vast majority of DHNA-CoA hydrolases of the hotdog-fold superfamily. Nearly one quarter of the cluster's sequences were already annotated as DHNA-CoA hydrolases in Swiss-Prot. This cluster corresponds to Dillon and Bateman's uncharacterized group 37.

The small cluster A.64 contains coronafacic acid (CFA) dehydratases involved in CFA biosynthesis. This is confirmed by Swiss-Prot annotations as well as the lack of LpxC domains that are expected to co-occur with other FabZ-like dehydratase, of which CFA dehydratase is a subset (*3*).

### *2.3.4: Mapping high-throughput screens*

In order to characterize the hotdog-fold network, we selected diverse targets for HTS screening. Because these targets were chosen based on the thioesterase subfamily of the hotdog-fold superfamily, not all clusters are represented in the target list; indeed, this presents a good direction for future research. Nonetheless, many clusters are represented by the ultimate target list, with distribution across several of the larger nodes in particular. In total, 465 sequences in 105 Bacterial species were selected, 41 of which successfully underwent HTS screening (Figure 2.10).


[redacted]
A.1
A.2
A.3
A.5
A.4
A.6
A.7
A.8
A.9
A.10
A.11
A.12
A.13
A.14
A.15
A.16
A.17
A.18
A.19
A.20
A.21
A.22
A.23
A.24
A.25
A.26
A.27
A.28
A.29
A.30
A.31
A.32
A.33
A.34
A.35
A.36
A.37
A.38
A.39
A.40
A.41
A.42
A.43
A.44
A.45
A.46
A.47
A.48
A.49
A.50
A.51
A.52
A.53
A.54
A.55
A.56
A.57
A.58
A.59
A.60
A.61
A.62



[redacted]

**Figure 2.10:** The hotdog-fold SSN painted according to EFI HTS status (enlarged). HTS results are further subdivided by degree of promiscuity: low activity (triangle), specific activity (rectangle), promiscuous or very promiscuous activity (diamond). Nodes are colored thusly: not selected due to known literature/FLK function (green), selected as a target with no structural information (red), selected as a target with structural information/SNF (cyan), target with successful protein

purification and HTS screening (yellow), not a target and no known literature function (grey). Subnetworks A and B are represented in images (A) and (B), respectively.

Targets for which HTS screens were successfully performed were categorized based on the degree of activity they showed, ranging from low activity to specific activity for certain substrates to promiscuous activity for multiple substrates. The HTS and literature search function summaries were combined and used to paint the SSN according to overall substrate type (Figure 2.11)

The networks show a wide distribution of HTS-assigned specificity spread across the network—21 clusters contain sequences with HTS results. Most of the HTS results indicate promiscuity of some degree or another applied to a large number of the HTS result-containing clusters: 13 of the 21 HTS-containing clusters contain sequences with promiscuous activity, 9 contain only sequences with promiscuous activity. This underscores the inherent promiscuity in hotdogs.

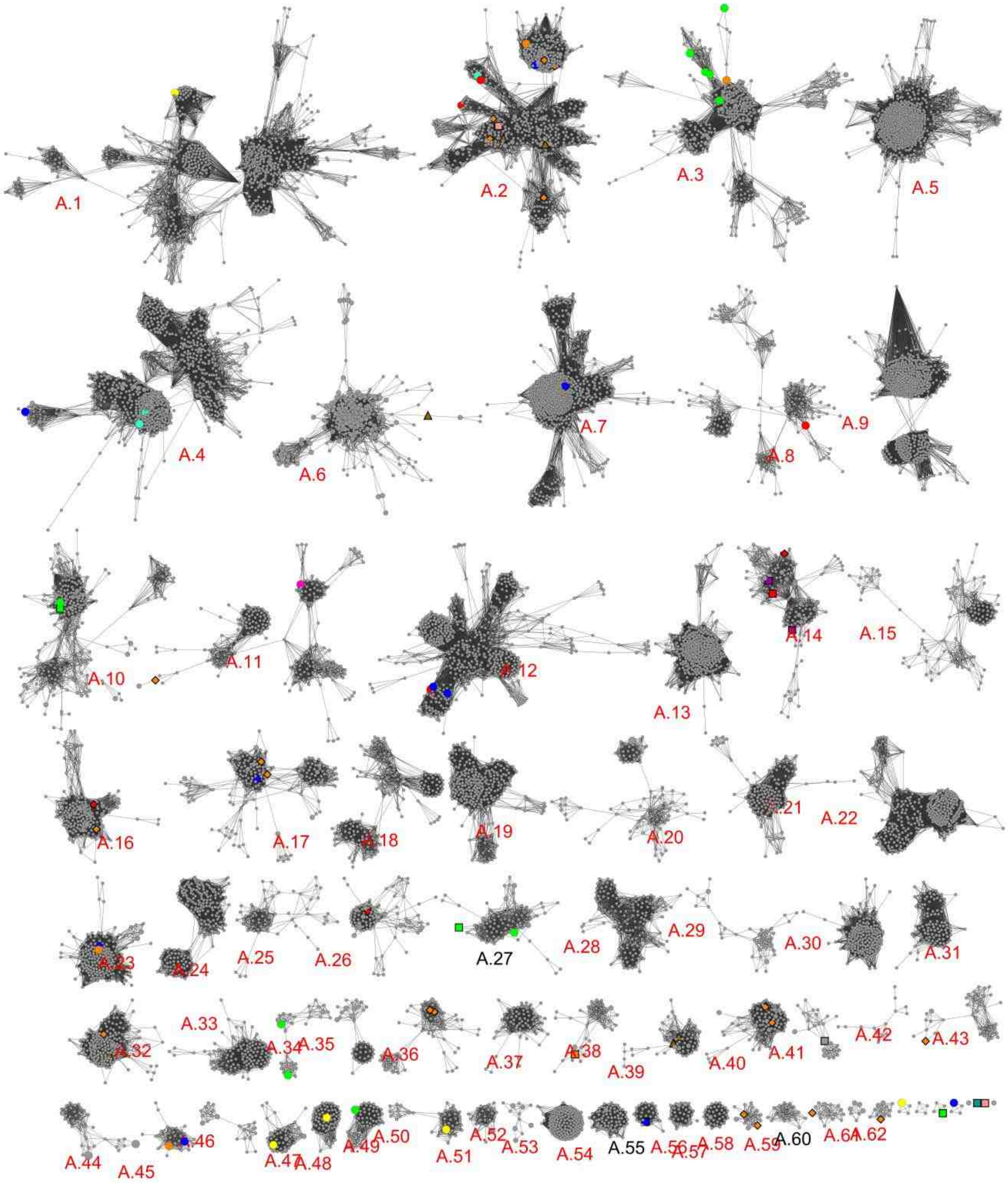


**Figure 2.11:** The hotdog-fold SSN painted according to HTS result and literature known function. Enlarged grey nodes represent targets for which HTS has not yet been completed. Node coloration is: broad range (orange), fatty acyl (yellow), aromatic (green), branched (pink), long chain and aromatic activities (purple), long chain (blue), medium chain (magenta), medium to long chain (red), short chain (cyan), short chain and aromatic (turquoise), no specific HTS activity

(brown). HTS results are subdivided by degree of promiscuity: low activity (triangle), specific activity (rectangle), promiscuous or very promiscuous activity (diamond).

### *2.3.5: Gene contexts of HTS targets*

Targets for which HTS screens were successfully performed underwent additional bioinformatics analysis in the form of gene context determination, by which most were found to have minimally informative gene contexts (Table 2.5).

| UniProtKB ID | Gene context summary | Cluster location | Active substrates, if specific | Activity class |
|---|---|---|---|---|
| **D2QSK4** | no neighbor conservation | A.2 | Low activity | low activity |
| **Q49YS3** | no neighbor conservation | A.2 | Low activity | low activity |
| **Q48BL7** | no neighbor conservation | A.2 | Promiscuous | short sat; medium sat; aromatic; deriv; |
| **Q9RZL9** | no neighbor conservation | A.2 | Promiscuous | short sat; branched; aromatic; |
| **A3M371** | insufficient data | A.2 | Succinyl CoA; | branched; |
| **Q11QP9** | no neighbor conservation | A.2 | Very promiscuous | short sat; branched; medium sat; long sat; long unsat; deriv; |
| **Q15YX3** | no neighbor conservation | A.2 | Very promiscuous | short sat; branched; short unsat; medium sat; long sat; long unsat; aromatic; deriv; |
| **Q5LWA2** | no neighbor conservation | A.2 | Very promiscuous | short sat; branched; medium sat; long sat; long unsat; aromatic; deriv; |
| **Q47SH7** | no neighbor conservation | A.6 | Low activity | low activity |
| **A5W3A3** | Conserved context (Paa) | A.10 | Phenylacetyl CoA; | aromatic; |
| **Q97AV4** | no neighbor conservation | A.11 | Promiscuous | branched; medium sat; long sat; long unsat; |
| **Q3J4C7** | insufficient data | A.14 | Promiscuous | medium sat; long sat; long unsat; deriv; |
| **A5W133** | insufficient data | A.14 | pentacosanoyl CoA; Benzoyl CoA; | long sat; aromatic |
| **A1TZH5** | insufficient data | A.14 | tridecanoyl CoA; pentadecanoyl CoA; Stearoyl CoA; (9Z_ 12Z_ 15Z-octadecatrienoyl) CoA; | medium sat; long sat_ long unsat; |
| **A1TY75** | insufficient data | A.14 | Linoleoyl CoA; Benzoyl CoA; | long unsat; aromatic; |
| **Q5LP35** | no neighbor conservation | A.16 | Promiscuous | medium sat; long sat; long unsat; aromatic; |
| **A3M7N5** | insufficient data | A.16 | Promiscuous | medium sat; long sat; long unsat; |
| **A3PJA8** | no neighbor conservation | A.17 | Promiscuous | short sat; branched; short unsat; long sat; long unsat; aromatic; |
| **Q5LMG0** | no neighbor conservation | A.17 | hexacosanoyl CoA; pentacosanoyl CoA; | long sat; |

| | | | | |
|---|---|---|---|---|
| **Q0C266** | no neighbor conservation | A.17 | Very promiscuous | short sat; branched; medium sat; long sat; long unsat; aromatic; deriv; |
| **A0QY86** | no neighbor conservation | A.26 | Promiscuous | medium sat; long sat; long unsat; deriv; |
| **Q12FZ4** | Conserved context | A.27 | Benzoyl CoA; Benzoyl CoA; | aromatic; |
| **Q73TX1** | no neighbor conservation | A.32 | Low activity | low activity |
| **Q15YT2** | insufficient data | A.32 | Very promiscuous | short sat; branched; short unsat; medium sat; long sat; long unsat; aromatic; deriv; |
| **Q0KBD3** | no neighbor conservation | A.36 | Promiscuous | short sat; medium sat; long sat; long unsat; deriv; |
| **Q12AK1** | Conserved context | A.36 | Very promiscuous | short sat; medium sat; long sat; long unsat; deriv; |
| **Q0KF28** | no neighbor conservation | A.38 | Hexanoyl CoA; Decanoyl CoA; Octanoyl CoA; hexacosanoyl CoA; 12:0 Biotinyl CoA; | short sat; medium sat; long sat; deriv; |
| **Q21SC3** | no neighbor conservation | A.39 | Low activity | low activity |
| **Q7MVA3** | no neighbor conservation | A.39 | Low activity | low activity |
| **A1U2I8** | no neighbor conservation | A.40 | Promiscuous | short sat; branched; short unsat; medium sat; deriv; |
| **A3SAI8** | insufficient data | A.40 | Very promiscuous | short sat; branched; short unsat; medium sat; long sat; long unsat; aromatic; deriv; |
| **A5ES38** | no neighbor conservation | A.41 | 04:0 Pyrene CoA; | deriv; |
| **Q0JZY5** | no neighbor conservation | A.43 | Promiscuous | short sat; branched; medium sat; deriv; |
| **Q9K8B6** | insufficient data | A.55 | docosahexaenoyl CoA; | long unsat; |
| **Q0BYF3** | insufficient data | A.58 | Promiscuous | medium sat; long unsat; aromatic; deriv; |
| **A5EMI2** | no neighbor conservation | A.58 | Very promiscuous | short sat; branched; medium sat; long sat; long unsat; aromatic; deriv; |
| **Q11TP9** | insufficient data | A.60 | Very promiscuous | short sat; branched; medium sat; long sat; long unsat; aromatic; deriv; |
| **Q11WY5** | no neighbor conservation | A.62 | Promiscuous | short sat; branched; medium sat; aromatic; deriv; |
| **Q7MS67** | no neighbor conservation | A.65 | Phenylacetyl CoA; | aromatic; |
| **Q0C3Y4** | insufficient data | Singleton | Acetyl CoA; Phenylacetyl CoA; | short sat; aromatic; |
| **B1M6X7** | insufficient data | Singleton | Glutaryl CoA; | branched; |

**Table 2.5:** HTS results, gene context summary, and SSN mapping for all targets having successfully undergone HTS screening. Insufficient data indicates that there was some degree of conserved context but there were insufficient data to make inferences. No neighbor conservation indicates that there was no or minimal conservation of neighbors. For the activity class, sat = saturated, unsat = unsaturated, deriv = derivatives, referring to the class of acyl-CoA substrate described in Table 2.1.

Most targets without illuminating gene contexts simply did not have sufficiently conserved neighbors or did not have sufficient annotation from which to infer operon or function information. In the case of Q0KBD3 (Figure 2.12), for example, very few neighbors are conserved within the query's order and even fewer are conserved within other orders. Even in the order of the query protein, only three neighbors are conserved approximately in approximately one quarter of order members, and not always simultaneously. This lack of neighbor conservation is typical of the majority of the query target HTS Proteins and is denoted by "no neighbor conservation" in Table 2.5.

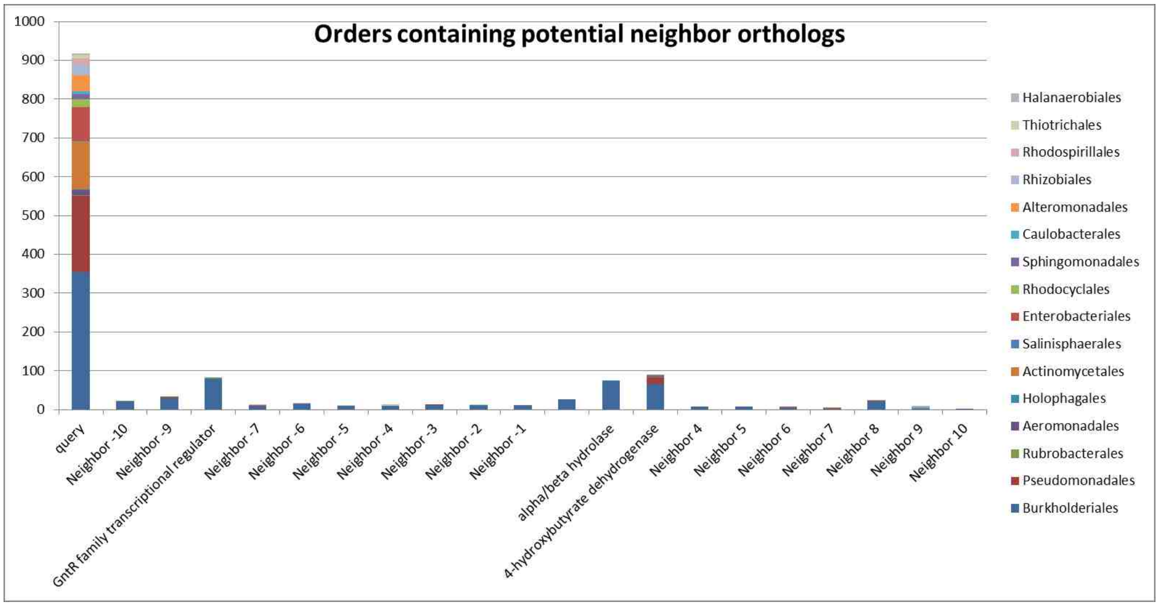


**Figure 2.12:** The order-level gene context for Q0KBD3 suggests no recurring gene context. Duplicate strains and subspecies were removed from the sample as described in Section 2.2.6; only those orders containing potential orthologs are displayed.

In the case of A1TZH5 (Figure 2.13), for another example, an adjacent protein was very well conserved within the query order and in other orders; two very close proteins were somewhat conserved on the order level. However, the three neighbor proteins were annotated very vaguely, as “thioesterase”, “heat shock protein 90”, and “membrane protein.” Literature searches in Google Scholar and PubMed using these three protein annotation as keywords in combination were unproductive. Without more specific annotations, the query and its neighbors cannot be assigned even a general or expected function. Other target proteins with similarly uninformative, conserved gene contexts are noted as “Insufficient data” in Table 2.5.

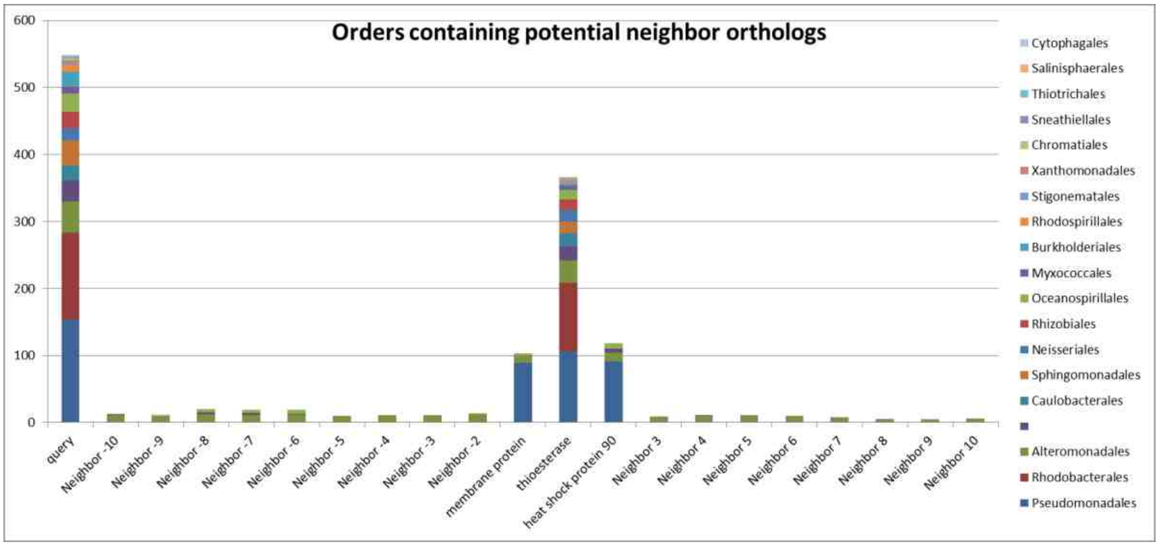


**Figure 2.13:** The order-level gene context for A1TZH5 suggests that the adjacent thioesterase is well-conserved; however, the query function cannot be guessed at based on this data, as no orthologs have annotation data more detailed than “thioesterase”.

A few targets did have sufficient conserved context and neighbor annotations to be considered as having “probable gene context; these are presented in Figure 2.14 and Figure 2.15.

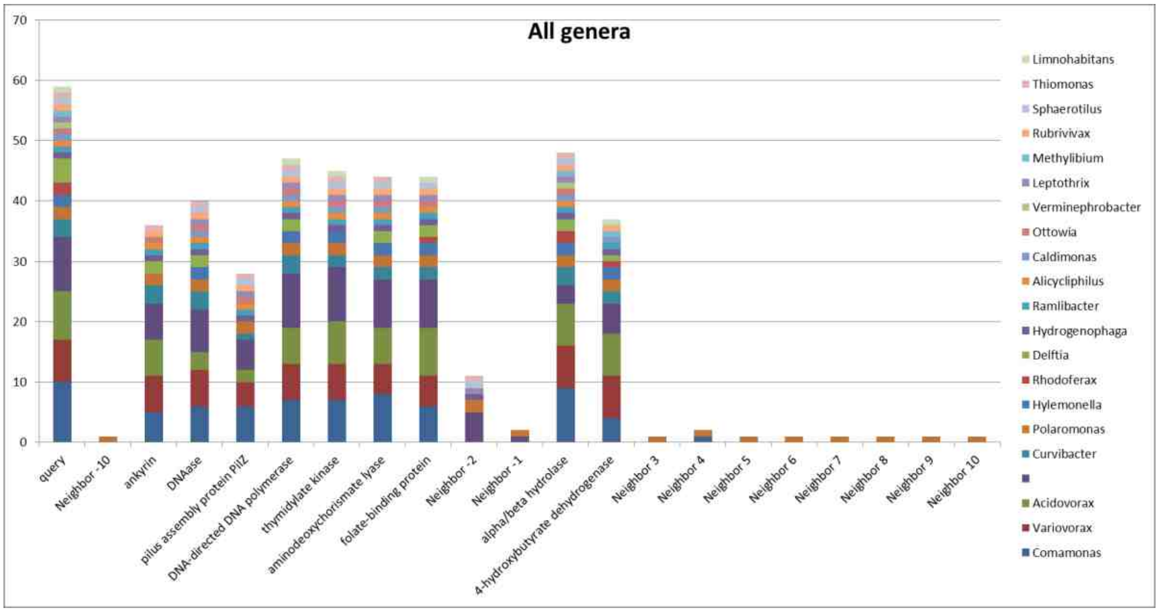


**Figure 2.14:** All genera for Q12AK1 (within Comamonadaceae). Context is largely conserved, but this is not unexpected necessarily across genera.

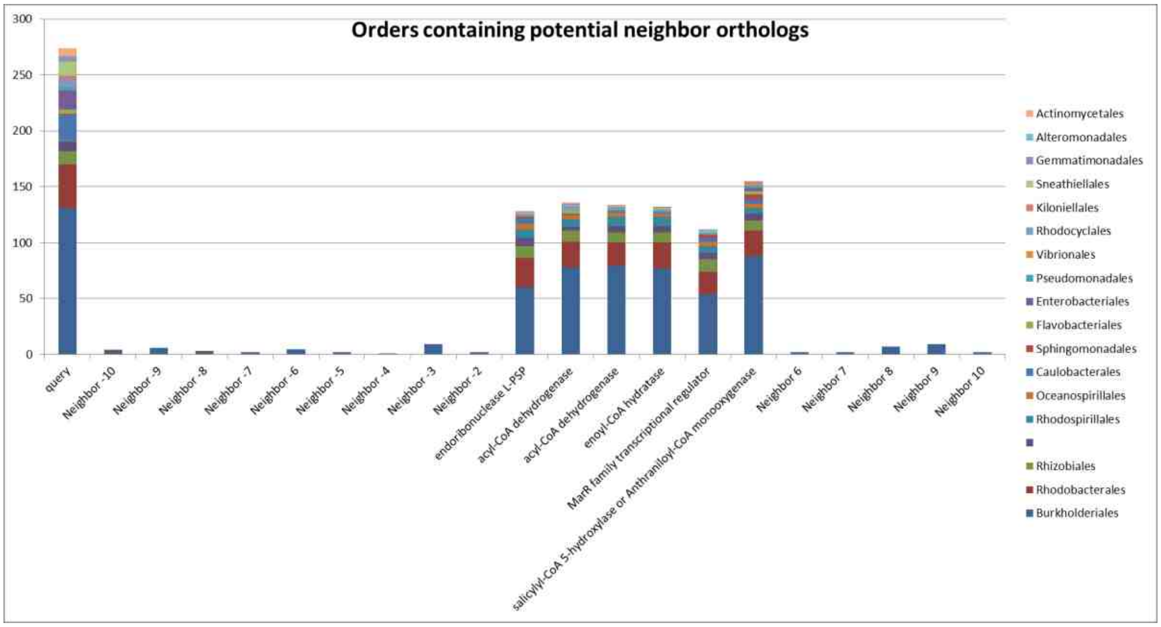


**Figure 2.15:** Order-level for Q12FZ4. Context is largely conserved.

### *2.3.5.1 Hotdog member A5W3A3 is a member of the phenylacetic acid degradation pathway*

Protein A5W3A3 from *Pseudomonas putida* was selected as one of the HTS target proteins; it was successfully expressed and underwent screening in Karen Allen’s lab, where it was determined to be specific for phenylacetyl CoA (*34*). Gene context determination reveals that it is frequently co-localized with members of the phenylacetic acid degradation pathway (Figure 2.16).

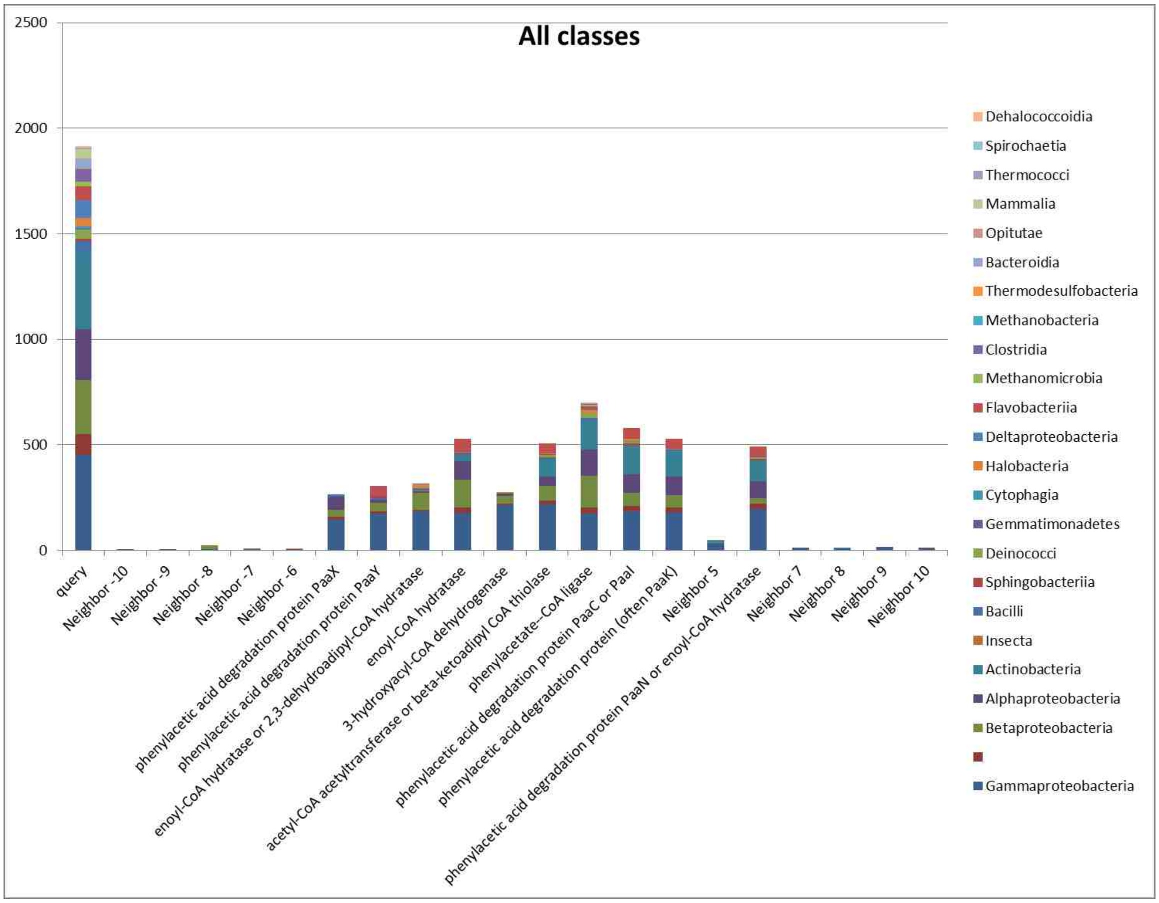


**Figure 2.16:** A5W3A3 Members of the phenylacetic acid degradation pathway. Based on the consensus of ortholog annotations, we can assign the 2,3 dehydroadipyl-CoA hydratase as PaaF, the adjacent enoyl-coA hydratase as PaaG, the 3-hydroxyacyl-coA dehydrogenase as PaaH, the query thioesterase (whose position falls between the dehydrogenase and the thiolase) as PaaI, the thiolase as PaaJ, and the ligase as possibly PaaK.

The phenylacetic acid degradation pathway (PAA pathway) is one pathway by which bacteria can use aromatic compounds as growth substrates *(39, 40)*. The operon composition and nomenclature vary in different species (*41*), but the primary composition is described in Figure 2.17.

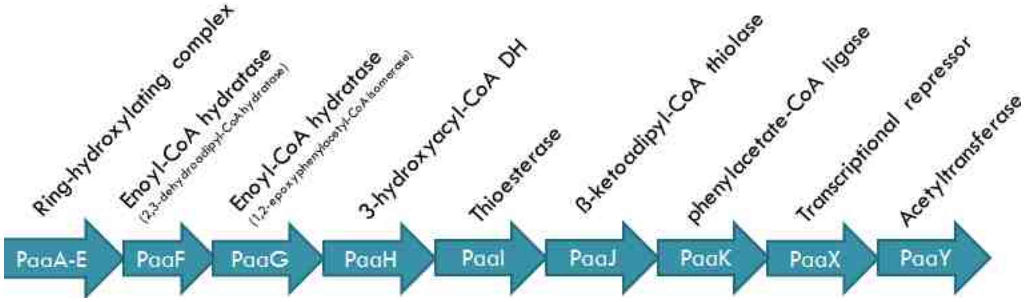


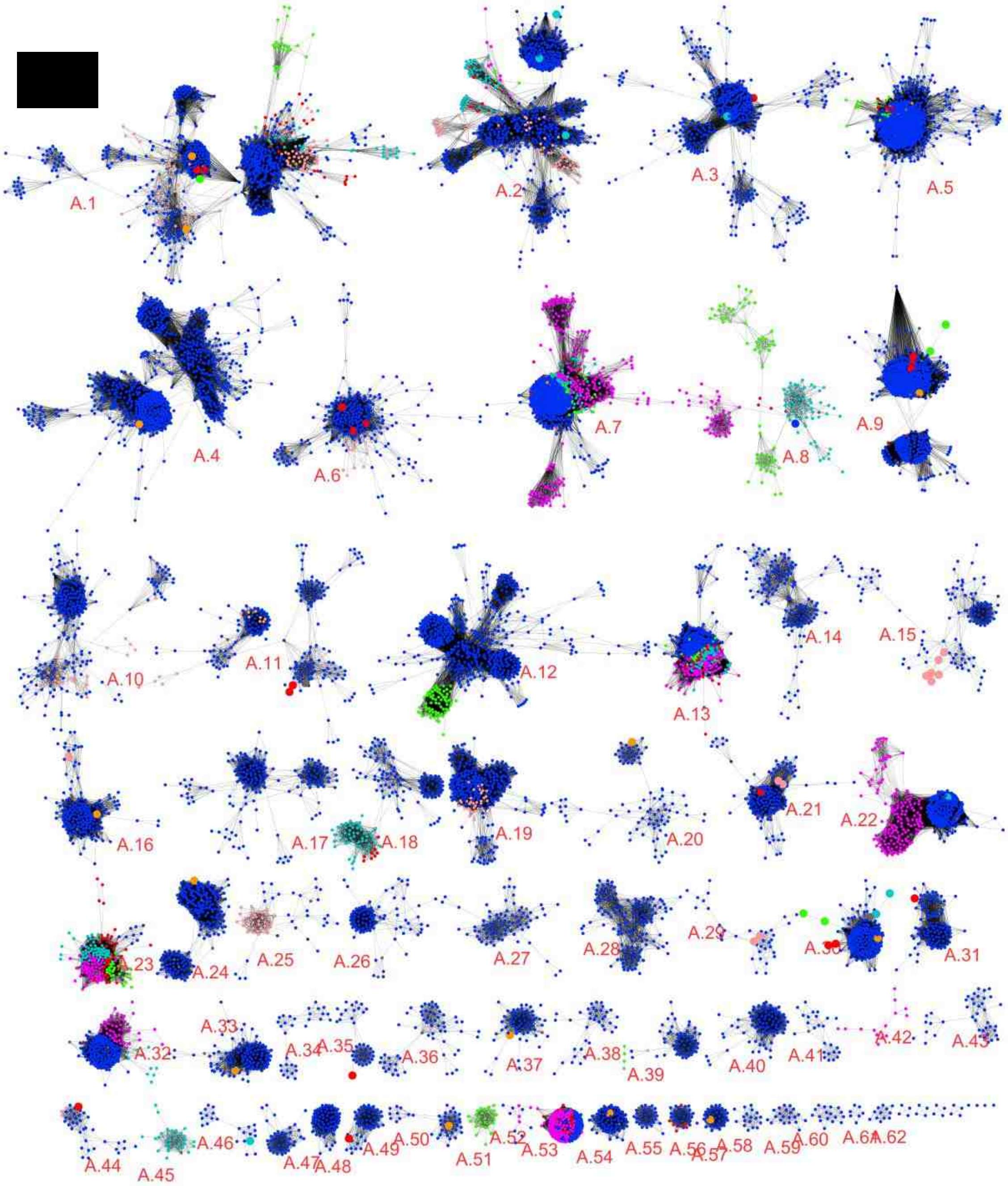
**Figure 2.17:** Composition of the phenylectic acid degradation operon in *E. coli* (39, 40).

The neighbor orthologs are not conserved across all classes containing A5W3A3 orthologs and nor are they completely conserved within Gammaproteobacteria. However, those members that are consistently conserved have consensus annotations and order that match the typical composition of the PAA pathway (Figures 2.16 and 2.17), suggesting membership within the pathway. Together, the HTS results and gene context confirm the automatic UniProt assignment of A5W3A3 to the PAA degradation pathway; specifically, the HTS and context support a specific annotation of PaaI. The order and annotations of the conserved gene context allow assignment of function to a number of the neighbors, as well: the 2,3 dehydroadipyl-CoA hydratase is PaaF, the adjacent enoyl-coA hydratase is PaaG, the 3-hydroxyacyl-coA dehydrogenase is PaaH, the thiolase is PaaJ, and the ligase as probably PaaK, though further investigation of the neighbor ortholog sometimes annotated as PaaK is warranted.

### *2.3.6: Diversity within domain- and phylum-level sequence similarity networks*

Hotdog-fold proteins are found across all domains of life, though most predominantly in bacteria (Table 2.3). A sequence similarity network painted according to taxonomic distribution at the Domain (and Kingdom level, for Eukaryotes) reveals that in many cases, hotdog-fold members cluster along domain or kingdom lines. Single domain or kingdom clusters are primarily the stuff of smaller clusters—in addition to Bacteria, Fungi and Viridiplantae frequently cluster into their own distinct, small clusters, though Archaea and Metazoa also have a small number of isolated clusters (Figure 2.18).

[redacted]


A.1
A.2
A.3
A.5
A.4
A.6
A.7
A.8
A.9
A.10
A.11
A.12
A.13
A.14
A.15
A.16
A.17
A.18
A.19
A.20
A.21
A.22
A.23
A.24
A.25
A.26
A.27
A.28
A.29
A.30
A.31
A.32
A.33
A.34
A.35
A.36
A.37
A.38
A.39
A.40
A.41
A.42
A.43
A.44
A.45
A.46
A.47
A.48
A.49
A.50
A.51
A.52
A.53
A.54
A.55
A.56
A.57
A.58
A.59
A.60
A.61
A.62

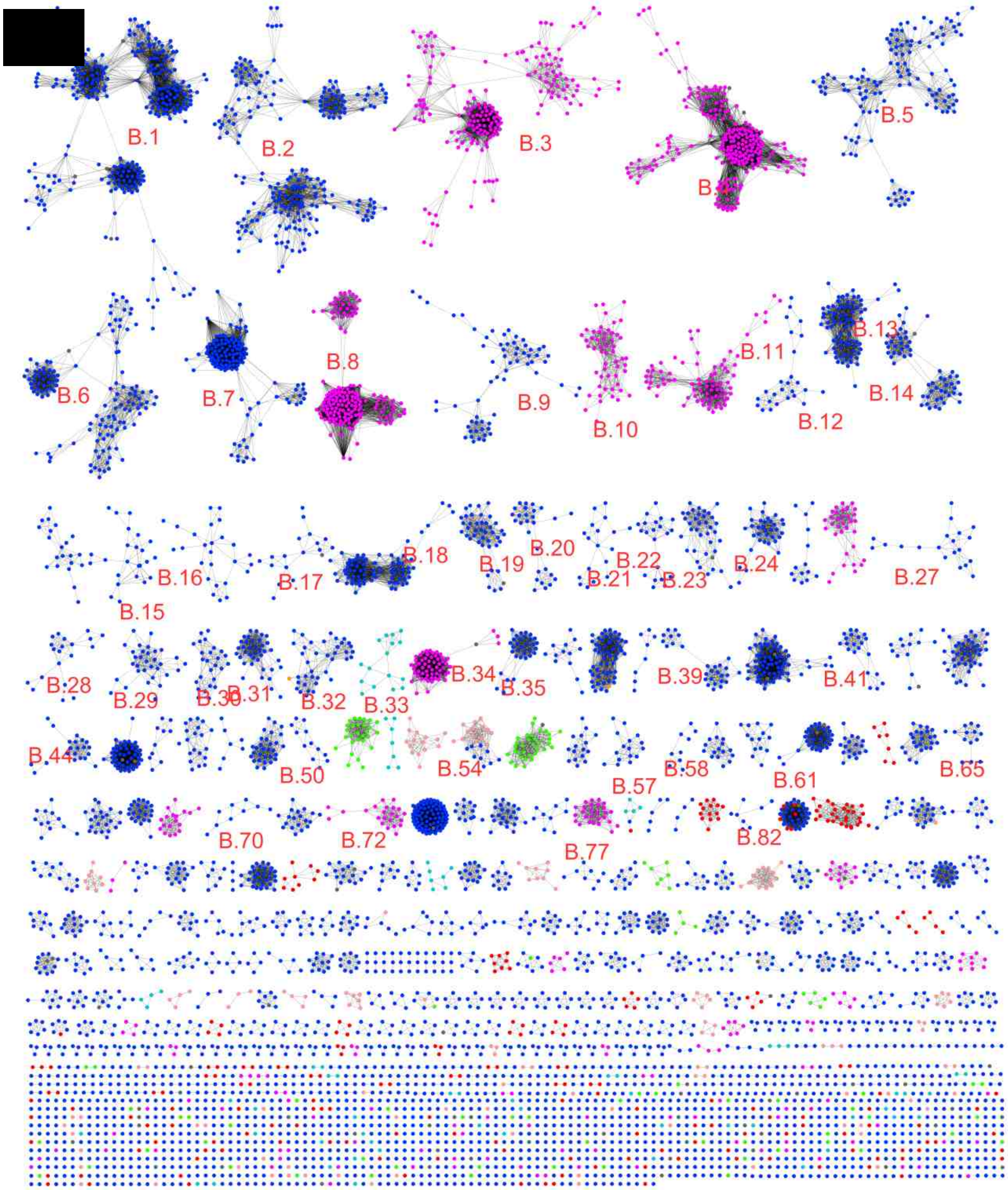


**Figure 2.18:** The hotdog-fold SSN painted according to taxonomic assignment (Domain, plus Kingdom for Eukaryotes) annotations in the UniProtKB. Color and node assignments are: Archaea (peach), Bacteria (blue), members from multiple groups (orange), Eukaryote (red), Eukaryote/Fungi (magenta), Eukaryote/Metazoa (cyan), Eukaryote/Viridiplantae (green). Nodes containing sequences with evidence of horizontal gene transfer are enlarged. Subnetworks A and B are represented in images (A) and (B), respectively.

Notably, however, several of the larger clusters boast members of multiple domains or kingdoms, both outliers within an otherwise iso-taxonomic cluster and well-populated subareas belonging to other domains or kingdoms.

In a number of these cases, different domains/kingdoms are arranged in distinct sub-clusters. In the case of A.1 and A.2, Metazoa and Archaea sequences form distinct 'sprays' away from the central, Bacteria cluster. This subdivision is clearer in A.7, where Fungi sequences with a sprinkling of Metazoa sequences form what appears to be a second hemisphere to the central, Bacteria cluster, with Fungi sequences branching out into their own 'arms'. Cluster A.12 and A.22 show similar patterns of having distinct offshoots for non-Bacteria sequences. In cluster A.8 and A.18, there is no central cluster—members of different taxonomic groups distinctly cluster on their own, tethered to each other by only one or two edges.

Clusters A.1, A.5, A.13, A.19, A.23, and A.53 are of particular interest because, unlike those described above, these clusters do not exhibit distinct sub-clustering patterns for sequences from different domains or kingdoms. In these clusters, while sequences from particular domains or kingdoms may group together like hemispheres or continents on a globe, they are nonetheless still distinctly part of the overall cluster—they share multiple edges with members from different taxonomic groups.

Hotdog-fold members are represented in all domains of life and they are also represented across the major bacterial phyla. In many ways, the distribution of hotdog-fold members mirrors the taxonomic distribution of proteins available from the UniProtKB: Firmicutes and Gamma Proteobacteria are best represented, followed by Actinobacteria, Alpha Proteobacteria, Bacteroides/Chlorobi, and Beta Proteobacteria (Figure 2.19). Unlike the domain distribution, the phylum bacterial distribution shows only minor clustering, though there is some degree of sub-clustering on a small scale (Appendix 1.4).

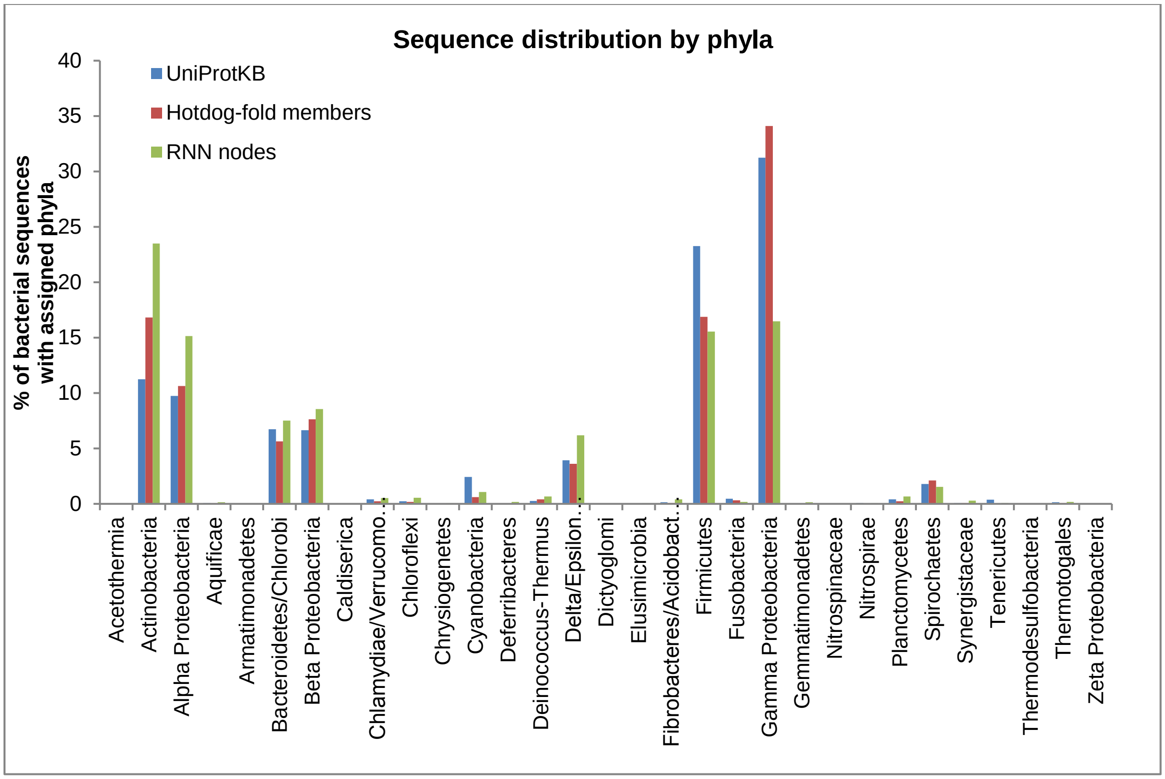


**Figure 2.19:** Distribution of Bacteria phyla (classes for Proteobacteria) within the UniProtKB (Accessed 5/29/15), SSN sequences with UniProt taxonomy information, and RNN nodes. Phylum membership is shown as percent of all sequences belonging to a given phylum within the dataset (UniProtKB N = 29494663; hotdog-fold members N = 69375; RNN nodes N = 13261).

### *2.3.7: Domain-level sequence similarity networks reveal evidence of gene transfer between domains*

The domain/kingdom-level SSN reveals that the majority of the hotdog-fold proteins cluster along domain/kingdom divisions, except as described above.

However, outliers within these iso-taxonomic clusters provide indicators of gene transfer events across taxa.  The majority of these outliers take the form of a single or very few proteins belonging to one or very few Eukaryotic species appearing in predominantly bacterial clusters (Figure 2.18).  However, there is also one instance of the reverse scenario and a few instances of isolated Archaeal species appearing amidst bacterial clusters Table 2.6.

In several cases noted in Table 2.6, these outliers have greater than 65% sequence identity to members of taxonomically distant groups with Swiss-Prot annotations while having no or very poor sequence similarity to members of their own taxonomic group.  The high sequence identity across taxa, the isolated existence of the outlier in its kingdom, and the large number of orthologs in other kingdoms all point to gene transfer into Eukaryotes.  One particular example of is particular note and is described below (Cluster A.20, B8BGK6 from *Oryza sativa subsp. Indica*).

| | Outlier species | Closely related (>50% SI) taxonomic group or [somewhat related (40-50% SI)] | Matched Swiss-Prot annotation | BLAST results to Swiss-Prot hotdogs | TrEMBL annotation consensus (>50% SI node members and neighbors) |
|---|---|---|---|---|---|
| **A.1** | Ricinus communis | Sphingomonadales | | | MaoC, oxidase regulatory protein, or acyl dehydratase |
| **A.1** | Aureococcus anophagefferens | Actinomycetales | | | MaoC, oxidase regulatory protein, or acyl dehydratase |
| **A.1** | Thalassiosira pseudonana | Actinomycetales | | | MaoC or acyl dehydratase |
| **A.1** | Necator americanus | Burkholderiales | | | MaoC, transcription regulatory protein, or acyl dehydratase |
| **A.1** | Ricinus communis | Firmicutes, protebacteria | | | MaoC dehydratase |
| **A.2** | Capitella teleta | [Alteromonadales, Oceanospirillales] | | | n/a |
| **A.2** | Acyrthosiphon pisum | Staphylococcus aureus | | | Uncharacterized protein |
| **A.2~** | Capitella teleta | Betaproteobacteria | yciA (acyl coa thioester hydrolase), palmitoyl and malonoyl coa | 98% with a single Endozoicomonas species, 70% with two others in same genus, ~60% with bacteria outside that genus. ~60% with the SwissProt proteins | yciA |
| **A.3** | Acanthamoeba castellanii | Candidatus Methylomirabilis oxyfera | | | no consensus |
| **A.3*** | Capitella teleta | Bacillales | putative esterase/ ydiI menI DHNA coA | ~56% with SwissProt hotdogs, as high as 62% with other hotdogs in same cluster, no bacteria with high SI | 4-hydroxybenzoyl-CoA thioesterase domain protein or DHNA coa (especially menI or ydiL) |
| **A.4*** | Rhodnius prolixus | Enterobacteriales | Acyl-CoA thioesterase YbgC | yes (~70%), but only with one domain of query protein | ybgc |
| **A.6** | Dictyostelium purpureum | Clostridiales | | | Putative uncharacterized protein |
| **A.6** | Dictyostelium discoideum | Clostridiales | | | Putative uncharacterized protein |
| **A.8** | Stigmatella aurantiaca | [Capsaspora owczarzaki, Amphimedon queenslandica] | | | n/a |
| **A.8** | Stigmatella aurantiaca | [Capsaspora owczarzaki, Amphimedon queenslandica] | | | n/a |
| **A.8** | Stigmatella aurantiaca | [Capsaspora owczarzaki, Amphimedon queenslandica] | | | n/a |
| **A.9** | Micromonas pusilla | [Moraxellaceae, Pseudonocardiaceae, Micromonosporaceae] | | | Putative uncharacterized protein |

|  | Outlier species | Closely related (>50% SI) taxonomic group or [somewhat related (40-50% SI)] | Matched Swiss-Prot annotation | BLAST results to Swiss-Prot hotdogs | TrEMBL annotation consensus (>50% SI node members and neighbors) |
|---|---|---|---|---|---|
| **A.9** | Emiliania huxleyi | [Proteobacteria, Firmicutes, Actinobacteria, Cyanobacteria, Bacteroidetes] |  |  | polyketide synthase |
| **A.9~** | Ricinus communis | Rhizobales | nodN/ probable enoyl coa hydratase 1 | ~70% with SwissProt bacterial hotdogs; as high 90% with non SwissProt bacterial hotdogs (bacterial hotdogs have a ~30 aa leading edge before aligning with eukaryote). |  |
| **A.9** | Dictyostelium discoideum | Actinomycetales |  |  | MaoC, nodN, enoyl coA, or acyl dehydratase |
| **A.9** | Dictyostelium purpureum | Actinomycetales |  |  | 3-hydroxyacyl-thioester dehydratase, enoyl coA hydratase, acyl dehydratase, MaoC |
| **A.9** | Acanthamoeba castellanii | Actinomycetales |  |  | MaoC, nodN, enoyl coA, or acyl dehydratase |
| **A.11** | Monosiga brevicollis | [Salpingoeca rosetta] |  |  | n/a |
| **A.11** | Capsaspora owczarzaki | [Vibrionales, Alteromonadales] |  |  | n/a |
| **A.11** | Salpingoeca rosetta | Vibrionales |  |  | n/a |
| **A.15** | various halobacteria | [Mycobacteriaceae, Nocardiaceae, Frankiaceae, Gordoniaceae] |  |  | MaoC domain containing protein dehydratase |
| **A.16** | Caenorhabditis remanei | Pseudomonadales |  |  | Phenylacetic acid degradation protein, |
| **A.16** | Halostagnicola larsenii | Desulfomonile tiedjei |  |  | Uncharacterized protein |
| **A.20~** | Oryza sativa subsp. Indica | Vibrionales | yiiD | 1 match with fungi (87% si, 100% query cover), excellent match with hotdog domain portion in bacteria (as high as 99% SI) | Galactoside O-acetyltransferase, putative YiiD or GNAT family acetyltransferase |
| **A.21** | Phaeodactylum tricornutum | Alcanivorax sp. W11-5 |  |  | Predicted protein |
| **A.21** | various Halorubrum | Myxococcales |  |  | uncharacterized protein (DUF4442) |
| **A.21** | Various Halobacteriaceae | Myxococcales |  |  | uncharacterized protein (DUF4442) |
| **A.24** | Necator americanus | Pseudomonas mandelii |  |  | Uncharacterized protein |
| **A.24** | Caenorhabditis remanei | Actinomycetales |  |  | tesB or thioesterase-like |

| | | | | | |
|---|---|---|---|---|---|
| | (Caenorhabditis vulgaris) | | | | |
| | **Outlier species** | **Closely related (>50% SI) taxonomic group or [somewhat related (40-50% SI)]** | **Matched Swiss-Prot annotation** | **BLAST results to Swiss-Prot hotdogs** | **TrEMBL annotation consensus (>50% SI node members and neighbors)** |
| **A.29*** | Nitrosopumilus maritimus | Thermotogales | 3-aminobutyryl-CoA ammonia-lyase | bacteria SI ~60%, higher hits in six other archaea | 3-aminobutyryl-CoA ammonia-lyase OR Beta-alanyl-CoA:ammonia lyase |
| **A.30** | Lottia gigantea | [Capitella teleta] | | | n/a |
| **A.30** | Capsaspora owczarzaki | [Corallococcus coralloides] | | | n/a |
| **A.30** | Volvox carteri | [Gamma proteobacterium HdN1] | | | n/a |
| **A.30** | Capitella teleta | [Moraxellaceae, Streptomycetaceae] | | | n/a |
| **A.30** | Nannochloropsis gaditana | [Other nannochloropsis] | | | n/a |
| **A.30** | Nannochloropsis gaditana | [Pseudonocardiaceae, Streptomycetaceae, Alteromonadaceae] | | | n/a |
| **A.30** | Trypanosoma congolense (strain IL3000) | Pseudomonas | | | MaoC dehydratase |
| **A.31** | Emiliania huxleyi | [Caulobacteraceae, Bradyrhizobiaceae] | | | n/a |
| **A.33** | Enterocytozoon bieneusi | Pseudomonas sp. RIT357 | | | A-factor biosynthesis hotdog domain |
| **A.35** | Cyanidioschyzon merolae | [Chroococcales] | | | n/a |
| **A.37** | Nematostella vectensis, Ricinus communis | Pseudomonadales | | | thioesterase |
| **A.37** | Ricinus communis | Pseudomonadales | | | thioesterase |
| **A.44*** | Emiliania huxleyi | Proteobacteria | Methylthioribose-1-phosphate isomerase | domain SI ~60% with bacterial, 100% with other eukaryote | Uncharacterized protein |
| **A.44** | Emiliania huxleyi | Proteobacteria | Methylthioribose-1-phosphate isomerase | domain SI ~60% with bacterial, 100% with other eukaryote | Uncharacterized protein |
| **A.46** | Amphimedon queenslandica | [Pseudonocardiaceae, Streptomycetaceae, Alteromonadaceae] | | | n/a |

| | | | | | |
|---|---|---|---|---|---|
| **A.49*** | Paulinella chromatophora | Synechococcus sp. RCC307 | DHNA-CoA hydrolase | max ~50% SI with bacteria | DHNA-CoA hydrolase |
| | **Outlier species** | **Closely related (>50% SI) taxonomic group or [somewhat related (40-50% SI)]** | **Matched Swiss-Prot annotation** | **BLAST results to Swiss-Prot hotdogs** | **TrEMBL annotation consensus (>50% SI node members and neighbors)** |
| **A.50*** | Acyrthosiphon pisum | Enterobacteriales | Long-chain acyl-CoA thioesterase FadM | Initial best BLAST hits are not hotdogs. | 4-hydroxybenzoyl-CoA thioesterase, YbgC/YbaW family, Long-chain acyl-CoA thioesterase tesC OR fadM, |
| **A.54** | Caenorhabditis remanei (Caenorhabditis vulgaris) | Pseudomonadales | | | 4-hydroxybenzoyl-CoA thioesterase, YbgC/YbaW family |
| **A.57~** | Rhodnius prolixus | Gammaproteobacteria | 3-hydroxy decanoyl-ACP dehydratase | y, >95% with bacterial hotdogs. BLAST hit with one other eukaryote (fungi Beauveria bassiana) at 90% SI | 3-hydroxydecanoyl-ACP dehydratase |

**Table 2.6:** Clusters containing one or a few sequences belonging to outlier species within a cluster predominantly of a different kingdom or domain (e.g., a eukaryotic species within a bacterial cluster). The outlier species is noted, along with the most closely related (greater than 50% SI to the outlying sequence) or somewhat related (40-50% SI to the outlying sequence, reported in brackets) members of the representative node or immediate neighbors within the cluster. Neighbors or representative node members with >50% SI to the outlier sequence were inspected for TrEMBL annotations, the consensus of which is reported, as well as any manually curated Swiss-Prot annotations. If a related Swiss-Prot annotation was found, the query sequence underwent a BLAST search to determine whether the Swiss-Prot annotation was the best hit among all non-redundant species, not just neighboring species within the cluster. An asterisk in the Cluster column indicates that the sequence had medium or poor general BLAST results to annotated or Swiss-Prot sequences; a tilde indicates that the sequence had good sequence identity to Swiss-Prot sequences from the BLAST results.

#### 2.3.7.1 An example of gene transfer from bacteria to plant and fungi species

The enzyme B8BGK6 from *Oryza sativa subsp. indica*, rice, is the sole Eukaryotic representative in a 3210 member representative node in the 65% RNN for the hotdog-family, a node that is otherwise comprised entirely of bacterial sequences and which presents in a cluster that is otherwise entirely bacterial (Cluster A.20). This outlier enzyme is annotated as a D-tyrosyl-tRNA$^{Tyr}$ deacylase based on automatic annotation from InterPro. The bacterial hotdog members within the same representative node were overwhelmingly annotated as either galactoside O-acetyltransferases or Gcn5-related N-acetyltransferases (GNAT family) (*42*); a BLAST search of B8BGK6 against its node members reveals that the hotdogs only align with ~77% of the eukaryotic enzyme.

A more comprehensive BLAST search against the NCBI non-redundant database reveals that B8BGK6 (gi: 218184431) has only one ortholog with high sequence identity and similar domain organization, an uncharacterized protein from the fungus *Beauveria bassiana D1-5* (gi: 701777303) which does not appear in the hotdog-fold SSN. All subsequent high-scoring BLAST results cover only 77% of the eukaryotic enzyme, corresponding to hotdog-family domains as assigned by the InterPro entry for B8BGK6 (Figure 2.20). These hits have very high sequence identity (as high as 99%) and belong exclusively to bacteria; they are primarily annotated as GNAT family acetyltransferases or YiiD/ galactoside O-acetyltransferases. BLAST hits corresponding to the remaining non-hotdog 22% of B8BGK6 also demonstrated high sequence identity to exclusively bacterial proteins, overwhelmingly annotated as D-tyrosyl-tRNA$^{Tyr}$ deacylase.

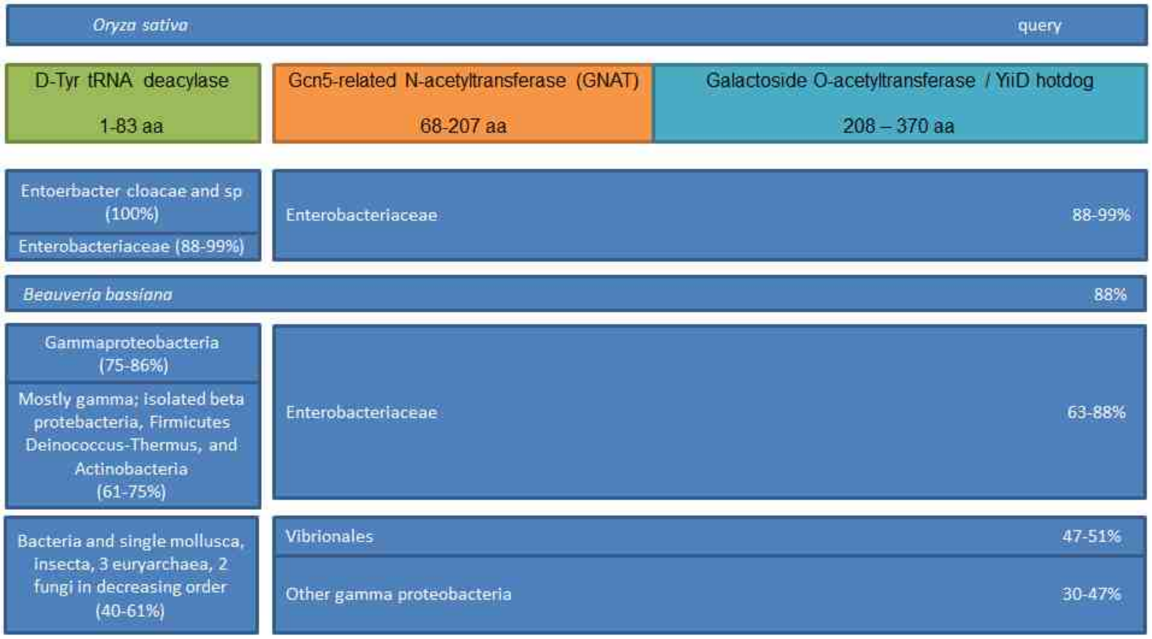

Oryza sativa
query
D-Tyr tRNA deacylase
1-83 aa
Gcn5-related N-acetyltransferase (GNAT)
68-207 aa
Galactoside O-acetyltransferase / YiiD hotdog
208 – 370 aa
Entoerbacter cloacae and sp (100%)
Enterobacteriaceae (88-99%)
Enterobacteriaceae
88-99%
Beauveria bassiana
88%
Gammaproteobacteria (75-86%)
Mostly gamma; isolated beta protebacteria, Firmicutes Deinococcus-Thermus, and Actinobacteria (61-75%)
Enterobacteriaceae
63-88%
Bacteria and single mollusca, insecta, 3 euryarchaea, 2 fungi in decreasing order (40-61%)
Vibrionales
47-51%
Other gamma proteobacteria
30-47%

**Figure 2.20:** Highest matching BLAST hits for B8BGK6 and the taxonomic groups to which hits belonged, arranged in order of sequence identity range for the taxonomic groups. Hits for the hotdog region of B8BGK6, ~77% of the sequence, are results from a BLAST search for the entire sequence. Hits for the deacylase region, ~22% of the sequence, are results from a BLAST search for that particular region.

Notably, both domains were clearly acquired from bacterial sources, as neither has homology to any eukaryotic proteins or domains (Figure 2.20). Furthermore, the hotdog-fold domain has sequence identities of up to 99.3% with hotdog-fold proteins in Enterobacter while the deacylase region has sequence identities up to 100%, also in Enterobacter. It is not unprecedented for fungi and plants to experience gene transfer, although such transfers usually come from bacterial donors (*43*)

D-Tyr-tRNA$^{Tyr}$ deacylases function as checks to recycle mis-aminoacylated D-Tyr-tRNA$^{Tyr}$, as well as other D-aminoacyl tRNAs (*44*). GNAT-family member histone acetyltransferase Hpa3 from yeast has been shown to act in conjunction with such deacylases in D-aminoacyl-tRNA recycling and removal, in order to avoid toxicity (*45*). It is possible that B8BGK6 has combined those activities into a single, bifunctional enzyme, the function of which is to combat D-amino acid toxicity (*46*). tRNA synthetases are also known to acylate coenzyme A and pantethionine arms (*47*), both of which are common substrates of the hotdog family, especially the thieosterases. It is conceivable that this fusion protein is capable of using the hotdog region to cleave the CoA moiety from a thioester, leaving the CoA as a substrate for the deacylase.

A literature search does not detect any precedent for a D-tRNA$^{Tyr}$ deacylase/acetyltransferase bifunctional enzyme, indicating that this enzyme would be of particular interest for further study, both as a suspected instance of gene transfer but also for its domain combination.

## 2.4: Conclusions

Using a sequence similarity network clustering proteins above ~40% sequence identity together, we are able to make reasonable predictions of subfamily membership and/or function for a number of previously unannotated sequences. We have been able to apply characterization annotations from previous publications (e.g., subfamily membership and quaternary structure) to ~143,000 sequences for subfamily annotations and ~63,000 for structure annotations, up from the original 1,100 subfamily annotations and 60 quaternary structure characterizations. We have expanded additional characterization of general subfamily membership (e.g., 4HBT and MaoC-like) to an additional

~61,000 sequences. We have also identified 52 clusters of varying size (8,700 sequences) entirely lacking annotation even on the Pfam level. These clusters are ideal targets for future characterizations.

High-throughput screen results describe a generally promiscuous sequence space. Gene context does not provide sufficient clues to assign function to most of the HTS targets, though a few candidates for assignment have been identified, pending assignment of the context to a function.

We have identified horizontal gene transfer suspects including the transfer of a novel bifunctional deacylase/deacetylase that appears in a single plant and single fungus species despite its overwhelming bacterial lineage. This and the other horizontal gene transfer suspects are intriguing targets for further study.

Ultimately, much of this work has focused on identifying clusters, trends, and discrepancies of interest. In particular, we have found interesting cases of taxonomic boundaries, the lack thereof, and instances of boundary crossing; multi- and varying-domain subsections within larger clusters; multiple clusters lacking any characterization whatsoever; clusters expected to be isofunctional but nonetheless containing sequences with differing quaternary structures, etc.

One could plumb the depths of any one of these concepts and continue to identify additional interesting directions to pursue. This work lays out a map of what is known, what can be inferred, what is not known, and the interface of all three.

## 2.5: References

1. Mitchell, A.; Chang, H. Y.; Daugherty, L.; Fraser, M.; Hunter, S.; Lopez, R.; McAnulla, C.; McMenamin, C.; Nuka, G.; Pesseat, S.; Sangrador-Vegas, A.; Scheremetjew, M.; Rato, C.; Yong, S. Y.; Bateman, A.; Punta, M.; Attwood, T. K.; Sigrist, C. J.; Redaschi, N.; Rivoire, C.; Xenarios, I.; Kahn, D.; Guyot, D.; Bork, P.; Letunic, I.; Gough, J.; Oates, M.; Haft, D.; Huang, H.; Natale, D. A.; Wu, C. H.; Orengo, C.; Sillitoe, I.; Mi, H.; Thomas, P. D.; Finn, R. D., The InterPro protein families database: the classification resource after 15 years. *Nucleic acids research* **2015,** *43*, D213-21.
2. Finn, R. D.; Bateman, A.; Clements, J.; Coggill, P.; Eberhardt, R. Y.; Eddy, S. R.; Heger, A.; Hetherington, K.; Holm, L.; Mistry, J.; Sonnhammer, E. L.; Tate, J.; Punta, M., Pfam: the protein families database. *Nucleic acids research* **2014,** *42*, D222-30.
3. Dillon, S. C.; Bateman, A., The Hotdog fold: wrapping up a superfamily of thioesterases and dehydratases. *BMC bioinformatics* **2004,** *5*, 109.
4. Pidugu, L. S.; Maity, K.; Ramaswamy, K.; Surolia, N.; Suguna, K., Analysis of proteins with the 'hot dog' fold: prediction of function and identification of catalytic residues of hypothetical proteins. *BMC structural biology* **2009,** *9*, 37.
5. Labonte, J. W.; Townsend, C. A., Active site comparisons and catalytic mechanisms of the hot dog superfamily. *Chemical reviews* **2013,** *113*, 2182-204.

6. Heath, R. J.; Rock, C. O., Roles of the FabA and FabZ beta-hydroxyacyl-acyl carrier protein dehydratases in Escherichia coli fatty acid biosynthesis. *J Biol Chem* **1996,** *271*, 27795-801.

7. Park, S. J.; Lee, S. Y., Identification and characterization of a new enoyl coenzyme A hydratase involved in biosynthesis of medium-chain-length polyhydroxyalkanoates in recombinant Escherichia coli. *J Bacteriol* **2003,** *185*, 5391-7.

8. Baev, N.; Schultze, M.; Barlier, I.; Ha, D. C.; Virelizier, H.; Kondorosi, E.; Kondorosi, A., Rhizobium nodM and nodN genes are common nod genes: nodM encodes functions for efficiency of nod signal production and bacteroid maturation. *J Bacteriol* **1992,** *174*, 7555-65.

9. Ponting, C. P.; Aravind, L., START: a lipid-binding domain in StAR, HD-ZIP and signalling proteins. *Trends Biochem Sci* **1999,** *24*, 130-2.

10. Zhuang, Z.; Song, F.; Martin, B. M.; Dunaway-Mariano, D., The YbgC protein encoded by the ybgC gene of the tol-pal gene cluster of Haemophilus influenzae catalyzes acyl-coenzyme A thioester hydrolysis. *FEBS letters* **2002,** *516*, 161-3.

11. Angelini, A.; Cendron, L.; Goncalves, S.; Zanotti, G.; Terradot, L., Structural and enzymatic characterization of HP0496, a YbgC thioesterase from Helicobacter pylori. *Proteins* **2008,** *72*, 1212-21.

12. Gerding, M. A.; Ogata, Y.; Pecora, N. D.; Niki, H.; de Boer, P. A., The trans-envelope Tol-Pal complex is part of the cell division machinery and

required for proper outer-membrane invagination during cell constriction in E. coli. *Mol Microbiol* **2007,** *63*, 1008-25.

13. Ohlrogge, J. B.; Jaworski, J. G., Regulation of Fatty Acid Synthesis. *Annu Rev Plant Physiol Plant Mol Biol* **1997,** *48*, 109-136.

14. Song, F.; Zhuang, Z.; Finci, L.; Dunaway-Mariano, D.; Kniewel, R.; Buglino, J. A.; Solorzano, V.; Wu, J.; Lima, C. D., Structure, function, and mechanism of the phenylacetate pathway hot dog-fold thioesterase PaaI. *J Biol Chem* **2006,** *281*, 11028-38.

15. Noyes, B. E.; Glatthaar, B. E.; Garavelli, J. S.; Bradshaw, R. A., Structural and functional similarities between mitochondrial malate dehydrogenase and L-3-hydroxyacyl CoA dehydrogenase. *Proc Natl Acad Sci U S A* **1974,** *71*, 1334-8.

16. Huang, H.; Pandya, C.; Liu, C.; Al-Obaidi, N. F.; Wang, M.; Zheng, L.; Toews Keating, S.; Aono, M.; Love, J. D.; Evans, B.; Seidel, R. D.; Hillerich, B. S.; Garforth, S. J.; Almo, S. C.; Mariano, P. S.; Dunaway-Mariano, D.; Allen, K. N.; Farelli, J. D., Panoramic view of a superfamily of phosphatases through substrate profiling. *Proc Natl Acad Sci U S A* **2015**.

17. Zhao, S.; Sakai, A.; Zhang, X.; Vetting, M. W.; Kumar, R.; Hillerich, B.; San Francisco, B.; Solbiati, J.; Steves, A.; Brown, S.; Akiva, E.; Barber, A.; Seidel, R. D.; Babbitt, P. C.; Almo, S. C.; Gerlt, J. A.; Jacobson, M. P., Prediction and characterization of enzymatic activities guided by sequence similarity and genome neighborhood networks. *eLife* **2014,** *3*.

18. Uberto, R.; Moomaw, E. W., Protein similarity networks reveal relationships among sequence, structure, and function within the Cupin superfamily. *PLoS One* **2013,** *8*, e74477.

19. Barber, A. E., 2nd; Babbitt, P. C., Pythoscape: a framework for generation of large protein similarity networks. *Bioinformatics* **2012,** *28*, 2845-6.

20. Atkinson, H. J.; Morris, J. H.; Ferrin, T. E.; Babbitt, P. C., Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. *PLoS One* **2009,** *4*, e4345.

21. Brown, S. D.; Babbitt, P. C., Inference of functional properties from large-scale analysis of enzyme superfamilies. *J Biol Chem* **2012,** *287*, 35-42.

22. Mashiyama, S. T.; Malabanan, M. M.; Akiva, E.; Bhosle, R.; Branch, M. C.; Hillerich, B.; Jagessar, K.; Kim, J.; Patskovsky, Y.; Seidel, R. D.; Stead, M.; Toro, R.; Vetting, M. W.; Almo, S. C.; Armstrong, R. N.; Babbitt, P. C., Large-scale determination of sequence, structure, and function relationships in cytosolic glutathione transferases across the biosphere. *PLoS biology* **2014,** *12*, e1001843.

23. Gerlt, J. A.; Babbitt, P. C.; Jacobson, M. P.; Almo, S. C., Divergent evolution in enolase superfamily: strategies for assigning functions. *J Biol Chem* **2012,** *287*, 29-34.

24. Gerlt, J. A.; Bouvier, J. T.; Davidson, D. B.; Imker, H. J.; Sadkhin, B.; Slater, D. R.; Whalen, K. L., Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A web tool for generating protein sequence similarity networks. *BBA - Proteins and Proteomics* **2015**.

25. Smoot, M. E.; Ono, K.; Ruscheinski, J.; Wang, P. L.; Ideker, T., Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **2011,** *27*, 431-2.

26. Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T., Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* **2003,** *13*, 2498-504.

27. Li, W.; Godzik, A., Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **2006,** *22*, 1658-9.

28. Boeckmann, B.; Bairoch, A.; Apweiler, R.; Blatter, M. C.; Estreicher, A.; Gasteiger, E.; Martin, M. J.; Michoud, K.; O'Donovan, C.; Phan, I.; Pilbout, S.; Schneider, M., The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic acids research* **2003,** *31*, 365-70.

29. O'Donovan, C.; Martin, M. J.; Gattiker, A.; Gasteiger, E.; Bairoch, A.; Apweiler, R., High-quality protein knowledge resource: SWISS-PROT and TrEMBL. *Briefings in bioinformatics* **2002,** *3*, 275-84.

30. UniProt, C., UniProt: a hub for protein information. *Nucleic acids research* **2015,** *43*, D204-12.

31. Schnoes, A. M.; Brown, S. D.; Dodevski, I.; Babbitt, P. C., Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS computational biology* **2009,** *5*, e1000605.

32. Consortium, U. Why have some UniProtKB accession numbers been deleted? How can I track them? http://www.uniprot.org/help/deleted_accessions (May 20, 2015),

33. Zimney, L. Case Studies of the Hot Dog-Fold and Acyl-Adenylate-Forming Superfamilies: Characterizing the Importance of Functional Divergence in Cellular Metabolism. University of New Mexico, 2015.

34. Ji, T. Structure and Mechanism to Function: Allosteric Activiation of Phosphomannomutase 1 and Substrate Selectivity in Hotdog-fold Thioesterases. Boston University, 2015.

35. Cock, P. J.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; de Hoon, M. J., Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **2009,** *25*, 1422-3.

36. Blattner, F. R.; Plunkett, G., 3rd; Bloch, C. A.; Perna, N. T.; Burland, V.; Riley, M.; Collado-Vides, J.; Glasner, J. D.; Rode, C. K.; Mayhew, G. F.; Gregor, J.; Davis, N. W.; Kirkpatrick, H. A.; Goeden, M. A.; Rose, D. J.; Mau, B.; Shao, Y., The complete genome sequence of Escherichia coli K-12. *Science* **1997,** *277*, 1453-62.

37. Coordinators, N. R., Database resources of the National Center for Biotechnology Information. *Nucleic acids research* **2015,** *43*, D6-17.

38. Horinouchi, S.; Suzuki, H.; Nishiyama, M.; Beppu, T., Nucleotide sequence and transcriptional analysis of the Streptomyces griseus gene

(afsA) responsible for A-factor biosynthesis. *J Bacteriol* **1989,** *171*, 1206-10.

39. Teufel, R.; Mascaraque, V.; Ismail, W.; Voss, M.; Perera, J.; Eisenreich, W.; Haehnel, W.; Fuchs, G., Bacterial phenylalanine and phenylacetate catabolic pathway revealed. *Proc Natl Acad Sci U S A* **2010**, *107*, 14390-5.

40. Ferrandez, A.; Minambres, B.; Garcia, B.; Olivera, E. R.; Luengo, J. M.; Garcia, J. L.; Diaz, E., Catabolism of phenylacetic acid in Escherichia coli. Characterization of a new aerobic hybrid pathway. *J Biol Chem* **1998**, *273*, 25974-86.

41. Luengo, J. M.; Garcia, J. L.; Olivera, E. R., The phenylacetyl-CoA catabolon: a complex catabolic unit with broad biotechnological applications. *Mol Microbiol* **2001**, *39*, 1434-42.

42. Vetting, M. W.; LP, S. d. C.; Yu, M.; Hegde, S. S.; Magnet, S.; Roderick, S. L.; Blanchard, J. S., Structure and functions of the GNAT superfamily of acetyltransferases. *Arch Biochem Biophys* **2005,** *433*, 212-26.

43. Richardson, A. O.; Palmer, J. D., Horizontal gene transfer in plants. *J Exp Bot* **2007,** *58*, 1-9.

44. Yang, H.; Zheng, G.; Peng, X.; Qiang, B.; Yuan, J., D-Amino acids and D-Tyr-tRNA(Tyr) deacylase: stereospecificity of the translation machine revisited. *FEBS letters* **2003,** *552*, 95-8.

45. Sampath, V.; Liu, B.; Tafrov, S.; Srinivasan, M.; Rieger, R.; Chen, E. I.; Sternglanz, R., Biochemical characterization of Hpa2 and Hpa3, two small

closely related acetyltransferases from Saccharomyces cerevisiae. *J Biol Chem* **2013,** *288*, 21506-13.

46. Soutourina, J.; Plateau, P.; Blanquet, S., Metabolism of D-aminoacyl-tRNAs in Escherichia coli and Saccharomyces cerevisiae cells. *J Biol Chem* **2000,** *275*, 32535-42.

47. Jakubowski, H., Quality control in tRNA charging -- editing of homocysteine. *Acta biochimica Polonica* **2011,** *58*, 149-63.

# CHAPTER 3

# EXPLORATION OF DIVERGENCE OF HADSF PHOSPHATASE SEQUENCE AND FUNCTION WITHIN THE BACTERIAL PHYLUM *FIRMICUTES*

## 3.1: Introduction

### *3.1.1: The HADSF Walkout project*

Closely related species generally have similar genomes and, hence, a similar number and type of proteins. For example, two different strains of *E. coli* have the same number of genes encoding HAD-like proteins, and those genes have very high if not identical sequence similarities: the NCBI record for NagD in E. coli (RefSeq WP_000153129.1) lists a number of strains and species as containing identical sequences (2055 other *E. coli* strains, 3 strains of *Escherichia fergusonii*, 48 different members of *Escherichia sp*, and 117 members of genus *Shigella*). Likewise, the closely related *Yersinia pestis* has 27 non-ATPase HADs, 24 of which are pairwise matches for 24 of the 25 *E. coli* non-ATPase HADs. Conversely, the more distantly related *Bacteroides thetaiotaomicron* shares only one HAD with *E. coli* and has an additional 18 non-ATPase HADs that do not match any *E. coli* HADs (Figure 3.1).

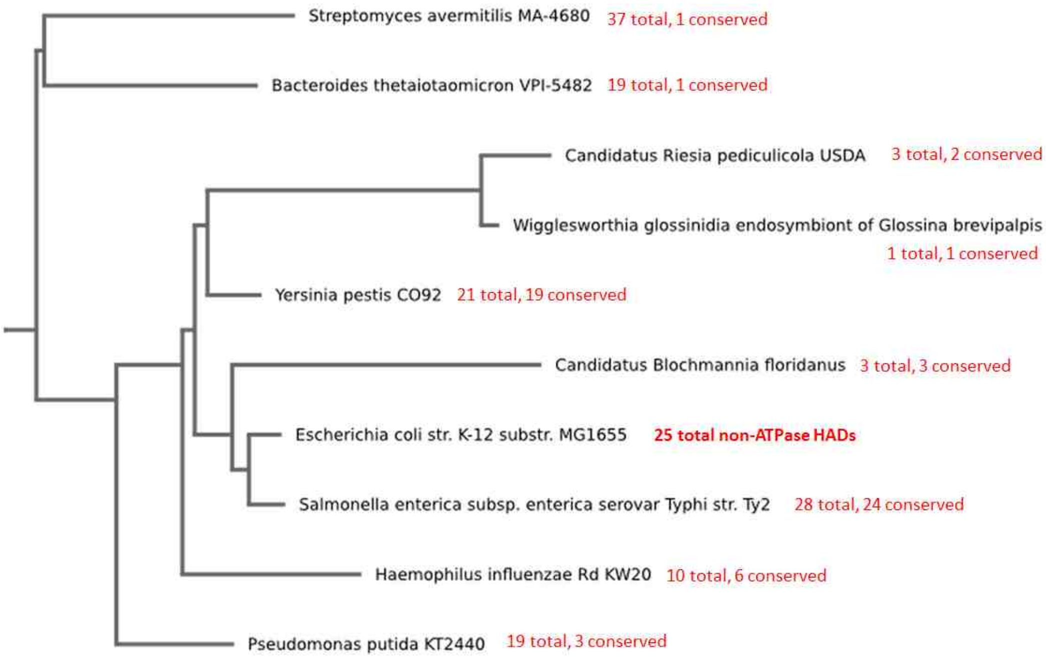

Streptomyces avermitilis MA-4680
37 total, 1 conserved
Bacteroides thetaiotaomicron VPI-5482
19 total, 1 conserved
Candidatus Riesia pediculicola USDA
3 total, 2 conserved
Wigglesworthia glossinidia endosymbiont of Glossina brevipalpis
1 total, 1 conserved
Yersinia pestis CO92
21 total, 19 conserved
Candidatus Blochmannia floridanus
3 total, 3 conserved
Escherichia coli str. K-12 substr. MG1655
25 total non-ATPase HADs
Salmonella enterica subsp. enterica serovar Typhi str. Ty2
28 total, 24 conserved
Haemophilus influenzae Rd KW20
10 total, 6 conserved
Pseudomonas putida KT2440
19 total, 3 conserved

**Figure 3.1:** Phylogenetic tree demonstrating the number of non-ATPase HADSF members in various species, as well as how many HADs are conserved, based on sequence identities compared to *E. coli* HADs. Generated using the Phylogenetic Tree tool at http://supfam.cs.bris.ac.uk/.

We refer to this process—looking at the number of shared/similar vs unshared/new proteins between species X and Y, and between species X and Z—as a "walkout". We start with *E. coli* and "walkout" to increasingly distantly related species, taking count of the number of shared HADs vs new HADs in each new species. This walkout gives a sense of what functions each additional species is capable of, as well as a sense of the structural landscape—different enzymes suggest different structure and, hence, different function. In this study, we conduct a phylum-wide walkout in order to map the HAD-sequence space of Firmicutes, especially as compared to that of *E. coli*.

### *3.1.2: Firmicutes is a key player in the gut microbiome*

Since the first sequencing of the genome, it has become increasingly apparent that we are not products of our genetic codes alone, but a complex interplay among many factors, such as gene expression and epigenetics, some of which we are just discovering. Scientific interest, especially popular science, has recently focused on the microbiome of the gut, particularly its role in human metabolism (*1*-*3*) and the implications changes in gut flora can have, especially with regards to disease and obesity (*4*-*9*). Approximately seven phyla of bacteria colonize the human and mouse gut, the most predominant of which are Firmicutes and Bacteroidetes (Table 3.1) and the relative abundance of which have been shown to correlate with obesity in mice (*7, 10*).

| | **Firmicutes** | **Bacteroidetes** | **Other** |
|---|---|---|---|
| **Humans (*10*)** | 51% | 48% | <1% |
| **Mice (*7*)** | 60-80% | 20-40% | <1% |

**Table 3.1:** Relative distribution of bacterial phyla in the colons of mice and humans. 'Other' encompasses Proteobacteria, Actinobacteria, Fusobacteria, Cyanobacteria, and Verrucobacteria, each of which represented <1% of bacterial sequences in the studies.

Within Firmicutes, we limited our scope to those members belonging to the family divisions noted in the online SUPERFAMILY database (Figure 3.2).

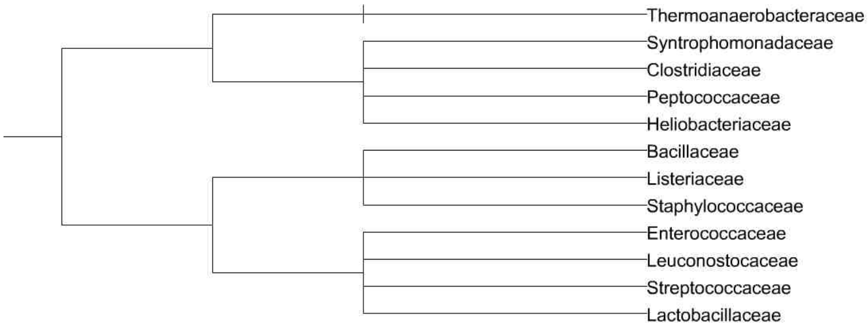


**Figure 3.2:** Phylogenetic tree of Firmicutes families represented in this study, generated using the NCBI taxonomy database, phyloT (http://phylot.biobyte.de/index.html**)**, and the Interactive Tree of Life (IToL).

### *3.1.3: Goals*

For this study, we were interested in comparative analysis of HAD phosphatases between representatives in Firmicutes to determine the overlap of HAD representation in the phylum. Specifically, we tracked the degree of sequence conservation, the number of HAD phosphatases across and within taxonomic representatives, gene context, and whether the HAD domain was part of a two-or-more-domain protein, which could indicate novel biosynthetic pathways. We also probed gene contexts and biological ranges of enzymes with potentially new function identified through HTS results via the EFI.

## 3.2: Methods

### *3.2.1: Manual bioinformatics analysis—gene contexts of Firmicutes HAD members*

We used the SUPERFAMILY database's taxonomic visualization tool to collect a list of HAD members within the Firmicutes phylum (*11*). After removing the multi-domain meta-ATPases, we manually ran each Firmicutes HAD sequence through the STRING protein-protein interaction database (*12*). We identified gene context by visually inspecting the “Neighborhood view” for recurrence of neighborhood proteins either globally or conserved within a taxonomic grouping.

We noted occurrence of fusion proteins when the gene of interest was shown combined with another gene that recurred sufficiently frequently to be considered conserved context; for example, trehalose-6-phosphate phosphatase is frequently fused to trehalose-6-phosphate synthase (Figure 3.3). Additionally, we accessed the SUPERFAMILY profile for each protein to determine whether the query protein is a multi-domain protein; we noted proteins as multi-domain proteins if the protein’s profile showed another domain present.

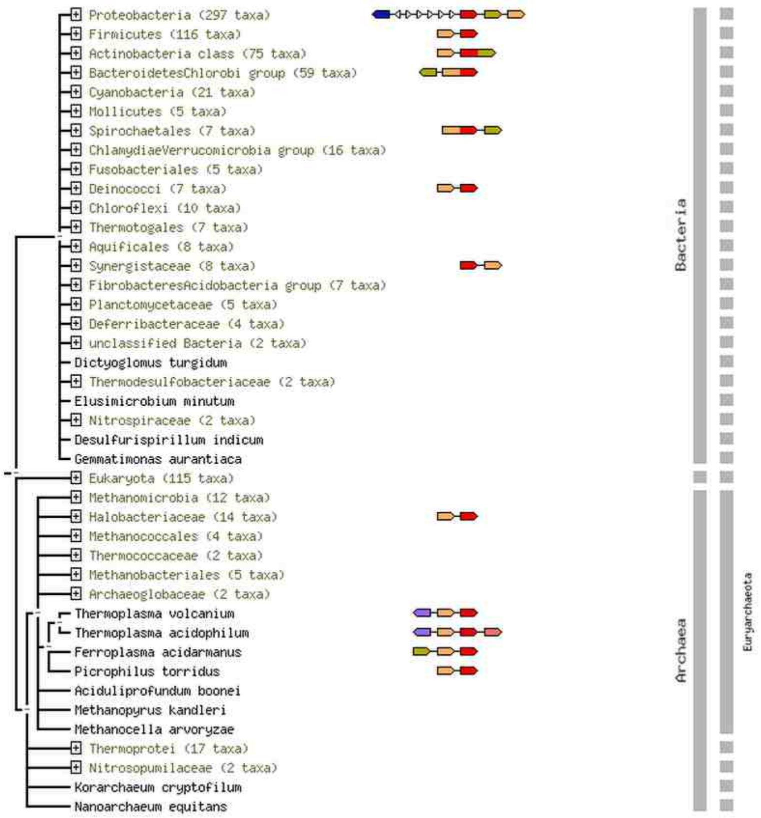


**Figure 3.3:** STRING database result for trehalose-6-phosphate phosphatase (Ta1209) showing that its orthologs are known to fuse with another domain. Protein domains shown are Ta1209 (red), trehalose-6-phosphate synthase (orange), glycosyl hydrolase (olive), tryptophanyl-tRNA synthetase (purple), glutamine synthetase (blue).

### *3.2.2: Generation of taxonomic lineages*

We manually compiled a taxonomy database in Excel by copying the taxonomic lineages for each species of interest from UniProt (*13*). After initial compilation, we used Excel to compare the genera of query species against the existing database; if a matching genus was found, its taxonomic data was applied to the new species and, if not, the species taxonomic data was copied from UniProt and added to the local taxonomy database.

### *3.2.3: Manual biological range of selected sequences from EFI HTS results*

Putative orthologs for each query were identified using BLAST searches of the NCBI non-redundant protein sequences database. Results were cut off at >80% query coverage and >40% sequence identity. The NCBI COBALT tool (*14*) was used to make multiple sequence alignments (MSAs) for the putative orthologs, which were then inspected for inclusion of the HAD superfamily catalytic motif DxD; any lacking the motif were removed. Taxonomic lineages to were assigned to each species using the taxonomic lineage database described in Section 3.2.2.

For visualization purposes, we pared down the resulting ortholog lists to remove species with multiple strains: we chose the strain with the highest hit sequence identity as the representative strain for that species and eliminated the others. We generated another phylogenetic tree and MSA using COBALT and default parameters. We modified and annotated the MSAs in FigTree (http://tree.bio.ed.ac.uk/software/figtree/) and visualized the final MSAs in a circular format using IToL.

### *3.2.4: Macro-assisted gene context of proteins from EFI HTS results*

For each query protein, we compiled a list of hits with >35% sequence identity and >80% query coverage as described in Section 3.2.3. We defined the gene neighborhood as the proteins having accession numbers within 10 numbers of the query protein; for example, a query protein with accession number

WC_00015 would have neighbors with accession numbers WC00005 through WC00014 and WC00016 through WC00025. Using a macro recorder (https://www.jitbit.com/macro-recorder/) and default parameters of the web-based BLAST program, we ran a species-specific BLAST for each neighbor in each species with a query result, recording the first hit in that species even if it fell below the >35% sequence identity and >80% query coverage parameters. Because such non-meaningful hits were automatically included, we later manually curated all results to remove them. We calculated neighbor distances from each query result in each species by subtracting the neighbor result accession number from the query result accession number, resulting in a neighbor value between +/- 10. An illustration of this process, which is somewhat different from the process described in Chapter 2, is shown in Figure 3.4. In cases with potential conserved gene context, we assigned gene function based on top hits/consensus (*15*).

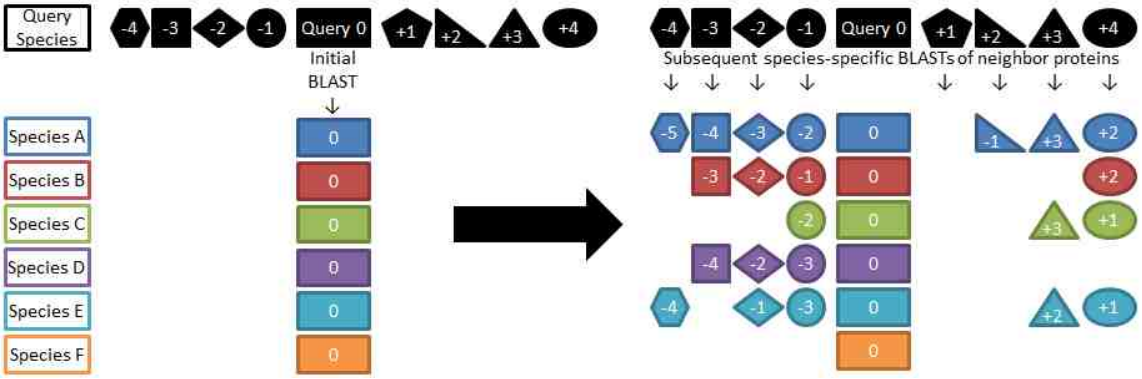


**Figure 3.4:** Initial BLAST results of a query protein result in a list of species containing putative orthologs. Each neighbor to the original query undergoes a species-specific BLAST search for each query ortholog species. If a neighbor has an ortholog in a given species, it is compared to the query ortholog in that species to determine whether the two orthologs are still neighbors.

### *3.2.5: Sequence similarity network generation*

As described in Chapter 2, Sequence Similarity Networks (SSNs) can be used to view relationships among large numbers of sequences based on the results of all-by-all BLAST searches. We generated the SSN for the Firmicutes HAD Walkout by using blast+, the standalone NCBI C++-based BLAST program (*16*), and a user-generated database created from the list of the 2299 Firmicutes HAD-like proteins and the 25 non-ATPase HAD-like proteins in *E. coli*, as collected in the SUPERFAMILY database. We used the UniProt ID mapping function (*13*) to acquire a multi-FASTA file for the HAD-like proteins; 40 were irretrievable and thus excluded from the SSN (Table 3.2). We visualized and edited networks in in both Cytoscape 2.8 and 3.1 (*17, 18*).

| GI number | Species | Protein family |
|---|---|---|
| 126698043 | Clostridium difficile 630 | Predicted hydrolases Cof |
| 126700721 | Clostridium difficile 630 | Predicted hydrolases Cof |
| 28376998 | Lactobacillus plantarum | Predicted hydrolases Cof |
| 28376999 | Lactobacillus plantarum | ß-phosphoglucomutase-like |
| 28377028 | Lactobacillus plantarum | ß-phosphoglucomutase-like |
| 28377072 | Lactobacillus plantarum | Meta-cation ATPase, catalytic domain P |
| 28377304 | Lactobacillus plantarum | Predicted hydrolases Cof |
| 28377306 | Lactobacillus plantarum | Predicted hydrolases Cof |
| 28377370 | Lactobacillus plantarum | Predicted hydrolases Cof |
| 28377449 | Lactobacillus plantarum | Meta-cation ATPase, catalytic domain P |
| 28377578 | Lactobacillus plantarum | Phosphonoacetaldehyde hydrolase-like |
| 28377656 | Lactobacillus plantarum | Predicted hydrolases Cof |
| 28377673 | Lactobacillus plantarum | Predicted hydrolases Cof |
| 28377710 | Lactobacillus plantarum | ß-phosphoglucomutase-like |
| 28377931 | Lactobacillus plantarum | Meta-cation ATPase, catalytic domain P |
| 28378095 | Lactobacillus plantarum | Predicted hydrolases Cof |
| 28378239 | Lactobacillus plantarum | NagD-like |

| | | |
|---|---|---|
| 28378407 | Lactobacillus plantarum | Predicted hydrolases Cof |
| 28378530 | Lactobacillus plantarum | Predicted hydrolases Cof |
| 28378567 | Lactobacillus plantarum | Meta-cation ATPase, catalytic domain P |
| 28378579 | Lactobacillus plantarum | ß-phosphoglucomutase-like |
| 28378626 | Lactobacillus plantarum | Phosphonoacetaldehyde hydrolase-like |
| 28378687 | Lactobacillus plantarum | ß-phosphoglucomutase-like |
| 28378834 | Lactobacillus plantarum | NagD-like |
| 28379128 | Lactobacillus plantarum | Predicted hydrolases Cof |
| 28379247 | Lactobacillus plantarum | Predicted hydrolases Cof |
| 28379271 | Lactobacillus plantarum | Predicted hydrolases Cof |
| 28379308 | Lactobacillus plantarum | HAD-related |
| 28379344 | Lactobacillus plantarum | ß-phosphoglucomutase-like |
| 28379365 | Lactobacillus plantarum | Predicted hydrolases Cof |
| 28379476 | Lactobacillus plantarum | Meta-cation ATPase, catalytic domain P |
| 28379494 | Lactobacillus plantarum | ß-phosphoglucomutase-like |
| 28379614 | Lactobacillus plantarum | Predicted hydrolases Cof |
| 28379680 | Lactobacillus plantarum | Meta-cation ATPase, catalytic domain P |
| 28379707 | Lactobacillus plantarum | Meta-cation ATPase, catalytic domain P |
| 28379733 | Lactobacillus plantarum | Meta-cation ATPase, catalytic domain P |
| 28379763 | Lactobacillus plantarum | Meta-cation ATPase, catalytic domain P |
| 28379848 | Lactobacillus plantarum | Predicted hydrolases Cof |
| 28379855 | Lactobacillus plantarum | ß-phosphoglucomutase-like |
| 23100582 | Oceanobacillus iheyensis HTE831 | Predicted hydrolases Cof |

**Table 3.2:** Proteins that were not retrievable using the UniProt ID mapping tool.

We used the blastp function in the command line to BLAST all proteins in the user-created database against each other using default parameters, retaining only results with E-values better (smaller) than $10^{-10}$. We manually removed self-paired hits from the BLAST results and used Excel to match SUPERFAMILY attributes to the UniProt ID mapping results. Because there were multiple results for some protein pairs, we wrote a Python program, ParseBLAST, to retain only the best E-value result for each protein pair (Appendix 2.1). We imported the annotated BLAST results to Cytoscape as a table using E-value as the interaction between hit and query.

### *3.2.6: Annotation of sequence similarity networks*

Much of the utility of sequence similarity networks and their representative node variants is in the simplicity and speed with which complicated relationships can be visually inspected. This is enhanced by the ability to colorfully annotate these networks with myriad different types of data.

The manually generated SSN for the Firmicutes HAD walkout was small enough for annotation data to be managed in Excel. We annotated the network based on protein family and taxonomic lineage information acquired from SUPERFAMILY, length acquired from UniProt, domain information from SUPERFAMILY and STRING, and gene context when applicable. Because ATPases tend to be complex multi-domain proteins and clustered together, we removed proteins assigned to the “Meta-cation ATPase, catalytic domain P” family from the SSN (Figure 3.5). We further refined the network by filtering it to a stringency of 40% sequence identity which, in the HADs, roughly corresponds to an E-value of $10^{-20}$ or better (*19*).

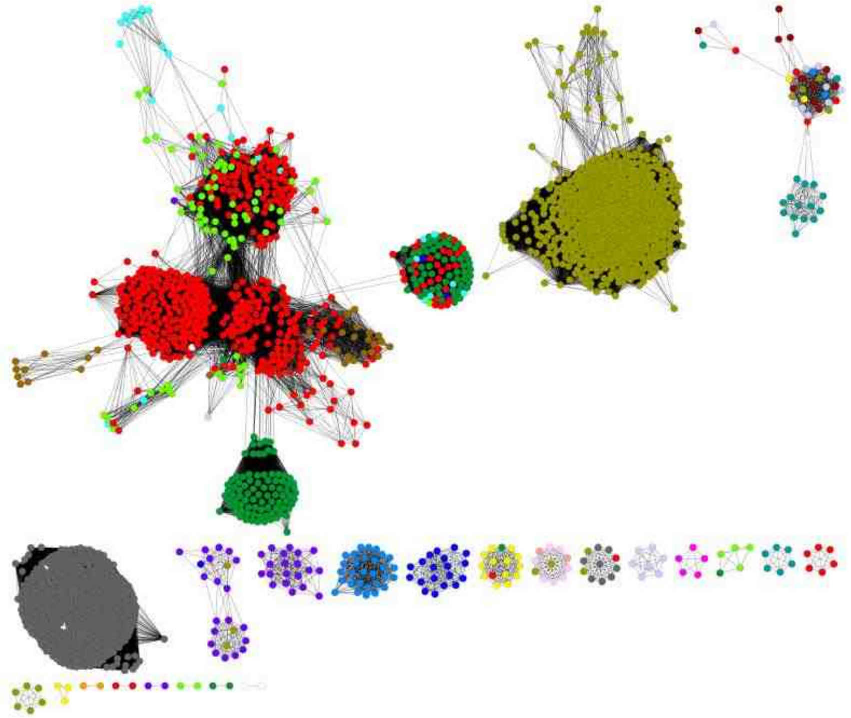


**Figure 3.5:** Original network generated for Firmicutes HAD-like members of SUPERFAMILY, showing all BLAST results with E-values of $10^{-10}$ or better. The network is painted according to protein family, as annotated in SUPERFAMILY, and colored thusly: 5'(3')-deoxyribonucleotidase dNT-2 (turquoise), ß-phosphoglucomutase-like (red), BT0820-like (dark red), Class B acid phosphatase AphA (yellow), enolase-phosphatease E1 (peach), HAD-related (lime green), histidinol phosphatase-like (blue), hypothetical protein (lavender), Magnesium-dependent phosphatase-1 Mdp1 (pink), ATPases (grey), MtnX-like (light blue), NagD-like (dark green), phosphonoacetaldehyde hydrolase-like (brown), phosphoserine phosphatase (purple), predicited hydrolase Cof (olive), trehalose-phosphatase (magenta), YihX-like (cyan), phosphatase domain of polynucleotide kinase (orange). ATPases are discarded in refined versions of this network shown below.

Due to the very large size of the HADSF (>370,00 UniProtKB sequences), we used a representative node network downloaded from the Structure Function Linkage Database (SFLD) to illustrate the entire network (*20*). Because the size of this RNN exceeded the data capacities of Excel, we used our Python program, AssignAttributes, to map user-generated annotations to the raw data and keyIDs from a network (Appendix 2.4). We used this mapped annotation data to paint and/or filter the network according to different attributes.

## 3.3: Results and discussion

### *3.3.1: Diversity of Firmicutes HAD-members painted on the sequence similarity network*

Within Firmicutes, there is great diversity in the number of HAD members, both across families and within genera. Table 3.3 illustrates the number of HAD-like proteins, excluding ATPases, across these classifications. The most dramatic disparities are in Bacillaceae (10-29 HADs) and Lactobacillaceae (10-28 HADs); the latter is more striking, as the species containing the most and fewest HADs are not only in the same family, but the same genus.

| Family | Species | non-ATPases |
|---|---|---|
| Bacillaceae | Bacillus megaterium QM B1551 | 29 |
| Bacillaceae | Bacillus thuringiensis str. Al Hakam | 29 |
| Bacillaceae | Bacillus cereus ATCC 14579 | 28 |
| Bacillaceae | Bacillus weihenstephanensis | 25 |
| Bacillaceae | Lysinibacillus sphaericus | 21 |
| Bacillaceae | Bacillus amyloliquefaciens FZB42 | 21 |
| Bacillaceae | Bacillus clausii KSM-K16 | 21 |
| Bacillaceae | Bacillus subtilis | 19 |
| Bacillaceae | Bacillus halodurans | 18 |
| Bacillaceae | Bacillus pumilus | 18 |
| Bacillaceae | Bacillus licheniformis DSM 13 = ATCC 14580 | 18 |
| Bacillaceae | Oceanobacillus iheyensis HTE831 | 16 |
| Bacillaceae | Bacillus pseudofirmus | 14 |
| Bacillaceae | Geobacilus sp WCH70 | 14 |
| Bacillaceae | Geobacillus thermodenitrificans NG80-2 | 12 |

| | | |
|---|---|---|
| Bacillaceae | Anoxybacillus flavithermus WK1 | 12 |
| Bacillaceae | Geobacillus kaustophilus HTA426 | 10 |
| Clostridiaceae | Clostridium beijerinckii | 31 |
| Clostridiaceae | Clostridium acetobutylicum 824 | 27 |
| Clostridiaceae | Clostridium saccharolyticum WM1 | 25 |
| Clostridiaceae | Clostridium cellulovorans 743B | 22 |
| Clostridiaceae | clostridium difficile 630 | 20 |
| Clostridiaceae | Clostridium phytofermentans | 18 |
| Clostridiaceae | Clostridium perfringens 13 | 16 |
| Clostridiaceae | Clostridium botulinum | 14 |
| Clostridiaceae | Clostridium ljungdahlii DSM 13528 | 13 |
| Clostridiaceae | Alkaliphilus metalliredigens | 12 |
| Clostridiaceae | Clostridum cellulolyticum H10 | 12 |
| Clostridiaceae | Clostridium kluyveri | 10 |
| Clostridiaceae | Clostridium novyi NT | 10 |
| Clostridiaceae | Clostridium thermocellum | 9 |
| Clostridiaceae | Clostridium tetani | 9 |
| Clostridiaceae | Alkaliphilus oremlandii | 9 |
| Enterococcaceae | Enterococcus faecium Aus0004 | 23 |
| Enterococcaceae | Tetragenococcus halophilus NBRC 12172 | 23 |
| Enterococcaceae | Enterococcus faecalis 62 | 20 |
| Enterococcaceae | Enterococcus hirae ATCC 9790 | 16 |
| Enterococcaceae | Enterococcus sp. 7L76 | 13 |
| Enterococcaceae | Melissococcus plutonius DAT561 | 11 |
| Heliobacteriaceae | Heliobacterium modesticaldum Ice1 | 7 |
| Lactobacillaceae | Lactobacillus plantarum | 28 |
| Lactobacillaceae | Lactobacillus crispatus ST1 | 20 |
| Lactobacillaceae | Lactobacillus casei | 20 |
| Lactobacillaceae | Lactobacillus casei str Zhang | 20 |
| Lactobacillaceae | Lactobacillus acidophilus | 19 |
| Lactobacillaceae | Lactobacillus brevis | 18 |
| Lactobacillaceae | Lactobacillus sakei | 18 |
| Lactobacillaceae | Lactobacillus gasseri | 17 |
| Lactobacillaceae | Lactobacillus rhamnosus | 17 |
| Lactobacillaceae | Pediococcus pentosaceus | 16 |

| | | |
|---|---|---|
| Lactobacillaceae | Lactobacillus johnsonii | 16 |
| Lactobacillaceae | Lactobacillus salivarius | 13 |
| Lactobacillaceae | Lactobacillus helveticus | 12 |
| Lactobacillaceae | Lactobacillus fermentum | 11 |
| Lactobacillaceae | Lactobacillus delbrueckii | 10 |
| Leuconostocaceae | Leuconostoc gasicomitatum LMG 18811 | 15 |
| Leuconostocaceae | Leuconostoc kimchii IMSNU 11154 | 13 |
| Leuconostocaceae | Leuconostoc sp. C2 | 13 |
| Leuconostocaceae | Leuconostoc citreum KM20 | 12 |
| Leuconostocaceae | Leuconostoc mesenteroides ssp. mesenteroides ATCC 8293 | 12 |
| Leuconostocaceae | Oenococcus oeni PSU-1 | 10 |
| Leuconostocaceae | Weissella koreensis KACC 15510 | 10 |
| Listeriaceae | Listeria seeligeri serovar 1/2b str. SLCC3954 | 21 |
| Listeriaceae | Listeria monocytogenes serotype 4b str. CLIP 80459 | 21 |
| Listeriaceae | Listeria innocua Clip11262 | 21 |
| Listeriaceae | Listeria welshimeri serovar 6b str. SLCC5334 | 20 |
| Listeriaceae | Listeria ivanovii subsp. ivanovii PAM 55 | 17 |
| Peptococcaceae | Desulforudis audaxviator | 8 |
| Peptococcaceae | Desulfotomaculum acetoxidans | 7 |
| Peptococcaceae | Desulfotomaculum reducens | 7 |
| Peptococcaceae | Pelotomaculum thermopropionicum | 7 |
| Peptococcaceae | Desulfitobacterium hafniense Y51 | 7 |
| Staphylococcaceae | Staphylococcus carnosus ssp. carnosus TM300 | 18 |
| Staphylococcaceae | Staphylococcus saprophyticus ssp. saprophyticus ATCC 15305 | 18 |
| Staphylococcaceae | Staphylococcus aureus ssp. aureus NCTC 8325 | 17 |
| Staphylococcaceae | Staphylococcus haemolyticus JCSC1435 | 16 |
| Staphylococcaceae | Staphylococcus epidermidis RP62A | 15 |
| Staphylococcaceae | Staphylococcus lugdunensis HKU09-01 | 15 |
| Staphylococcaceae | Staphylococcus pseudintermedius HKU10-03 | 15 |
| Staphylococcaceae | Macrococcus caseolyticus JCSC5402 | 12 |
| Streptococcaceae | Streptococcus gallolyticus UCN34 | 27 |
| Streptococcaceae | Streptococcus uberis | 26 |
| Streptococcaceae | Streptococcus dysgalactiae | 26 |

| | | |
|---|---|---|
| Streptococcaceae | Streptococcus thermophilus | 22 |
| Streptococcaceae | Streptococcus agalactiae | 22 |
| Streptococcaceae | Streptococcus gordonii | 21 |
| Streptococcaceae | Streptococcus suis | 21 |
| Streptococcaceae | Streptococcus mutans | 21 |
| Streptococcaceae | Streptococcus sanguinis | 20 |
| Streptococcaceae | Lactococcus lactis | 20 |
| Streptococcaceae | Streptococcus mitis B6 | 17 |
| Streptococcaceae | Streptococcus pneumoniae TCH8431/19A | 17 |
| Streptococcaceae | Streptococcus pyogenes | 14 |
| Streptococcaceae | Streptococcus pneumoniae | 14 |
| Syntrophomonadaceae | Syntrophomonas wolfei subsp. wolfei str. Goettingen | 7 |
| Syntrophomonadaceae | Syntrophothermus lipocalidus DSM 12680 | 6 |
| Thermoanaerobacteraceae | Thermoanaerobacter sp. X514 | 17 |
| Thermoanaerobacteraceae | Thermaoanaerobacter tengongensis | 11 |
| Thermoanaerobacteraceae | Thermanaerobacter italicus | 9 |
| Thermoanaerobacteraceae | Thermoanaerobacter mathranii | 8 |
| Thermoanaerobacteraceae | Moorella thermacetica | 7 |
| Thermoanaerobacteraceae | Ammonifex degensii | 6 |
| Thermoanaerobacteraceae | Carboxydothermus hydrogenoformans | 6 |

**Table 3.3:** Number of non-ATPase HADs in the species of Firmicutes included in the SUPERFAMILY database, arranged by family and in descending order of number of HADs.

When families are mapped to the SSN, we see that larger clusters contain all or most Firmicutes families, whereas smaller clusters are family-specific; in mid-sized clusters, families segregate into individual regions (Figure 3.6). The latter clustering pattern is consistent with families diverging from a common ancestor at some point and subsequently evolving independently.

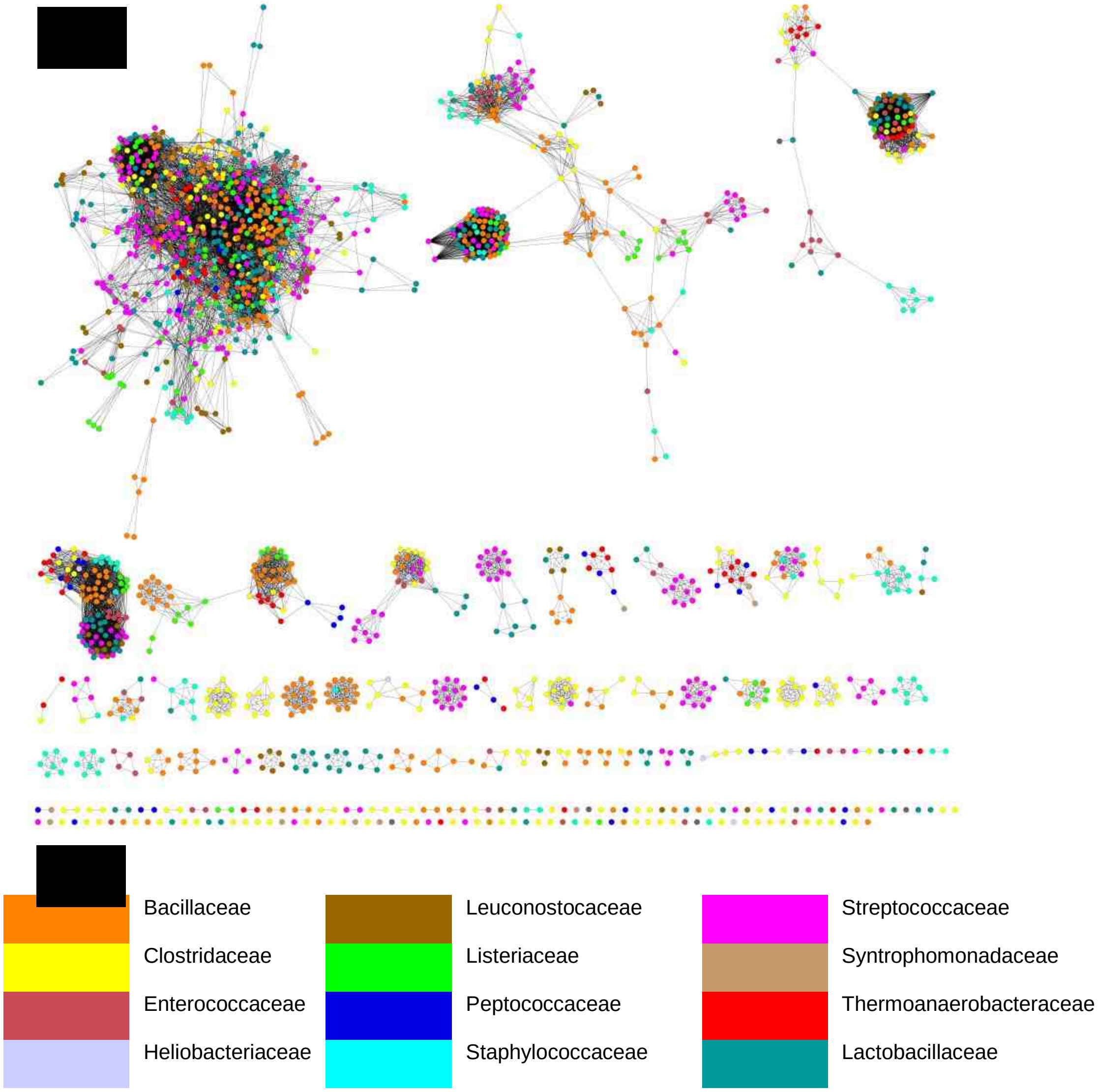


**Figure 3.6:** (A) Firmicutes SSN with a 40% SI threshold, excluding ATPases. Nodes are colored according to phylum in the key (B).

### *3.3.2: Family-level sequence similarity networks of Firmicutes HADSF members*

The Firmicutes SSN painted according to species distribution on the Family level reveals that several families contain connected pairs of sequences within the same isofunctional cluster. These connected pairs have >40% sequence identity to one another, suggesting that they are likely to be isofunctional and/or isostructural; that they are present in the same species suggests multiple gene duplication events or, more likely, a single gene duplication event in an ancestor shared by the species exhibiting these >40% SI sequence pairs. In the case of Lactobacillaceae, such pairs occur 8 times in one isofunctional cluster assigned to the Cof hydrolase subfamily and once in a ß-PGM cluster Figure 3.7. ß-PGM and Cof hydrolase subfamily clusters in Clostridiaceae, Enterococcaceae, and Bacillaceae also contain same-species >40% SI pairs while Leuconostocaceae and Listeriaceae contain such pairs only in their ß-PGM and Cof clusters, respectively.

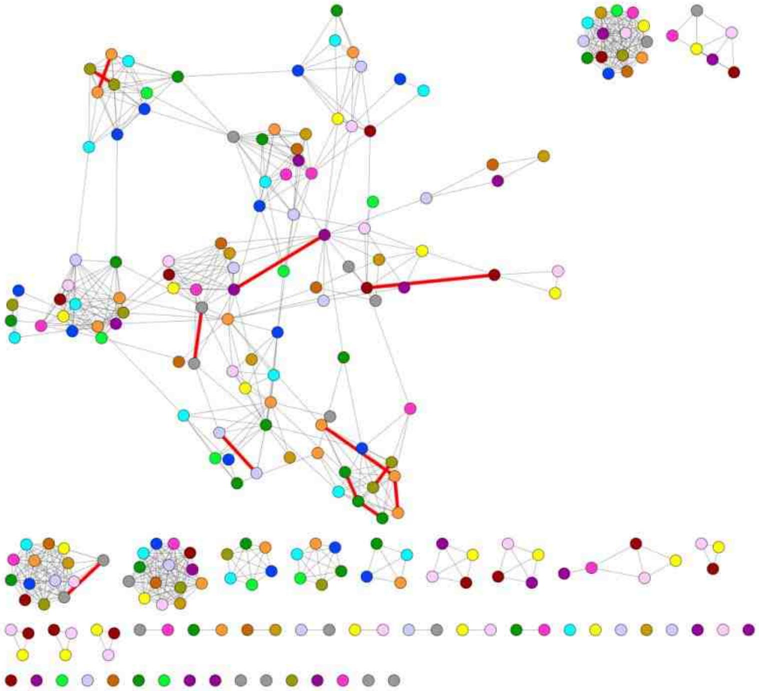


**Figure 3.7:** 40% SI family-level SSN of *Lactobacillaceae* with edges between same-species HADs colored red. Nodes are colored according to species: *Lactobacillus acidophilus* (orange), *L. brevis (*grey), *L. casei* (pink), *L. casei str Zhang* (yellow), *L. crispatus ST1* (dark green), *L. delbrueckii* (lime green), *L. fermentum* (brown), *L. gasseri* (cyan), *L. helveticus* (olive), *L. johnsonii*

(blue), *L. reuter* (tan), *L. rhamnosus* (dark red), *L. sakei* (purple), *L. salivarius* (magenta), *Pediococcus pentosaceus* (lavender).

Additionally, these family-specific SSNs show that while there are isofunctional clusters conserved throughout the family and containing representatives from every member species, there are also orphan sequences: single, isolated sequences or two-sequence clusters in families with >5 species (Figure 3.7). When mapped to the phylum-level SSN, many of these orphan proteins associate with clusters of enzymes in different families and some associate with other orphan proteins from different families (Figure 3.8). That some of these orphans have higher identity to sequences from other taxonomic groups suggests that gene transfer may have occurred across taxa. Conversely, those that do not exhibit high identity to other Firmicutes sequences are more divergent; if sufficiently so, they may represent potential new functions or structures.

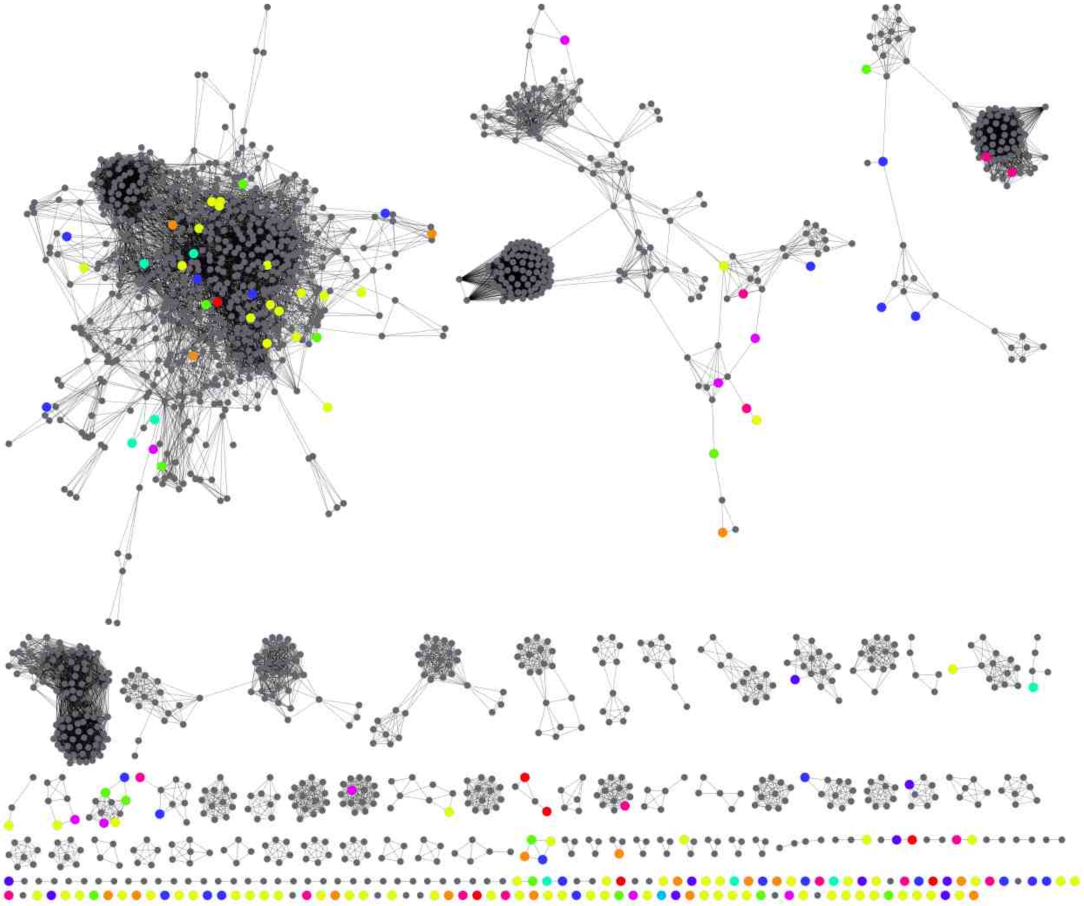


**Figure 3.8:** Orphan members of individual families in Firmicutes, mapped on the phylum-level SNN. Heliobacteraceae and Syntrophomonadaceae were excluded, as they contain only 1 and 2 members, respectively. Orphans belong to the following families: Bacillaceae (orange), Clostridiaceae (yellow), Enterococcaceae (green), Lactobacillaceae (cyan), Leuconostocaceae (blue), Peptococcaceae (purple), Staphylococcaceae (magenta), Streptococcaceae (pink), Thermoanaerobacteraceae (red).

### *3.3.3: Clustering of sequences along subfamily divisions*

The phylum-level sequence similarity network pruned of ATPases and constructed at the 40% sequence identity level, shows that the SUPERFAMILYascribed protein function families generally cluster together, as one would expect given that protein families are assigned computationally (Figure 3.9).

Notable exceptions occur in clusters 2, 4, 11, 21, 23, 36, in which subgroups intermingle, and clusters 7, 12, 33, and part of 2, in which subgroups co-occur but remain somewhat segregated internally. Other clusters contain multiple subgroups but the aforementioned are the largest. The most commonly co-occurring subgroups are the ß-PGM and HAD-related groups, co-occurring in five clusters; however, given that the HAD-related group is a catch-all for sequences with minimal characterization, it is more likely that such groups contain solely ß-PGM sequences. Cluster 4 is particularly interesting for being the largest cluster to contain so many different subgroups. Cluster 2 is notable for multiple subgroups that are very distinctly sub-clustered—it may be thought of as containing NagD and ß-PGM sequences, as the HAD-like nodes are likely to actually be ß-PGM sequences that are under-characterized.

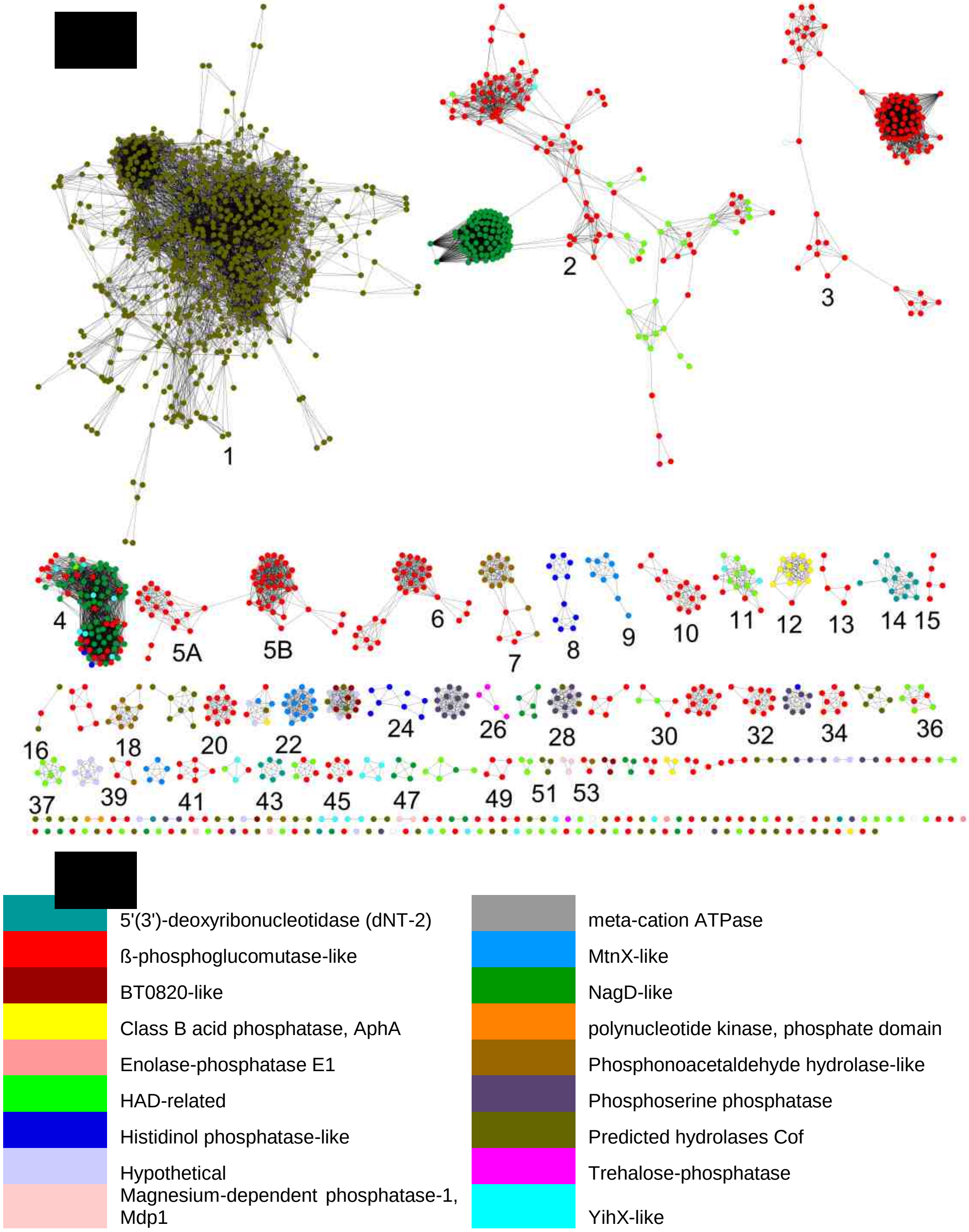

[redacted]
1
2
3
4
5A
5B
6
7
8
9
10
11
12
13
14
15
16
18
20
22
24
26
28
30
32
34
36
37
39
41
43
45
47
49
51
53
[redacted]
5'(3')-deoxyribonucleotidase (dNT-2)
ß-phosphoglucomutase-like
BT0820-like
Class B acid phosphatase, AphA
Enolase-phosphatase E1
HAD-related
Histidinol phosphatase-like
Hypothetical
Magnesium-dependent phosphatase-1, Mdp1
meta-cation ATPase
MtnX-like
NagD-like
polynucleotide kinase, phosphate domain
Phosphonoacetaldehyde hydrolase-like
Phosphoserine phosphatase
Predicted hydrolases Cof
Trehalose-phosphatase
YihX-like

**Figure 3.9:** (A) Firmicutes SSN with a 40% SI threshold, excluding ATPases. (B) Nodes are colored according to protein family, as annotated in SUPERFAMILY**.**

Notably, as in the family-level SSNs, there are many very small clusters with fewer than five members; these ‘orphan’ proteins have less than 40% sequence identity to the vast majority of the other Firmicutes HADs, indicating that many HADs in Firmicutes are not internally orthologogous on the phylum level. This is particularly interesting when these same proteins are mapped onto a larger HAD network, shown below (

Figure 3.10).

Any of the clusters containing Firmicutes orphans as an outlier in other phyla would be interesting for further study the circumstances under which Firmicutes HADs were more closely related to non-Firmicutes orthologs. The smaller clusters are particularly interesting as these represent orthologs and, hence, functionalities that are limited even among other phyla represented in the HADSF.

For example, one trio of orphans from *Streptococcus* (second from the left, second from bottom row in

Figure 3.10) has ~40% sequence identity to AphA (*21*) in *E. coli* but no other members of Firmicutes. On the full HADSF network, this trio clusters with four other meta-nodes containing dominant species of: *Haemophilus influenzae*, *Salmonella enterica*, *Photobacterium sp*, and *Edwardsiella tarda*, all of which are pathogenic. In *Salmonella*, AphA has been shown to be necessary for assimilation of nicotinamide mononucleotide (*22*). Its role in Firmicutes remains to be seen.

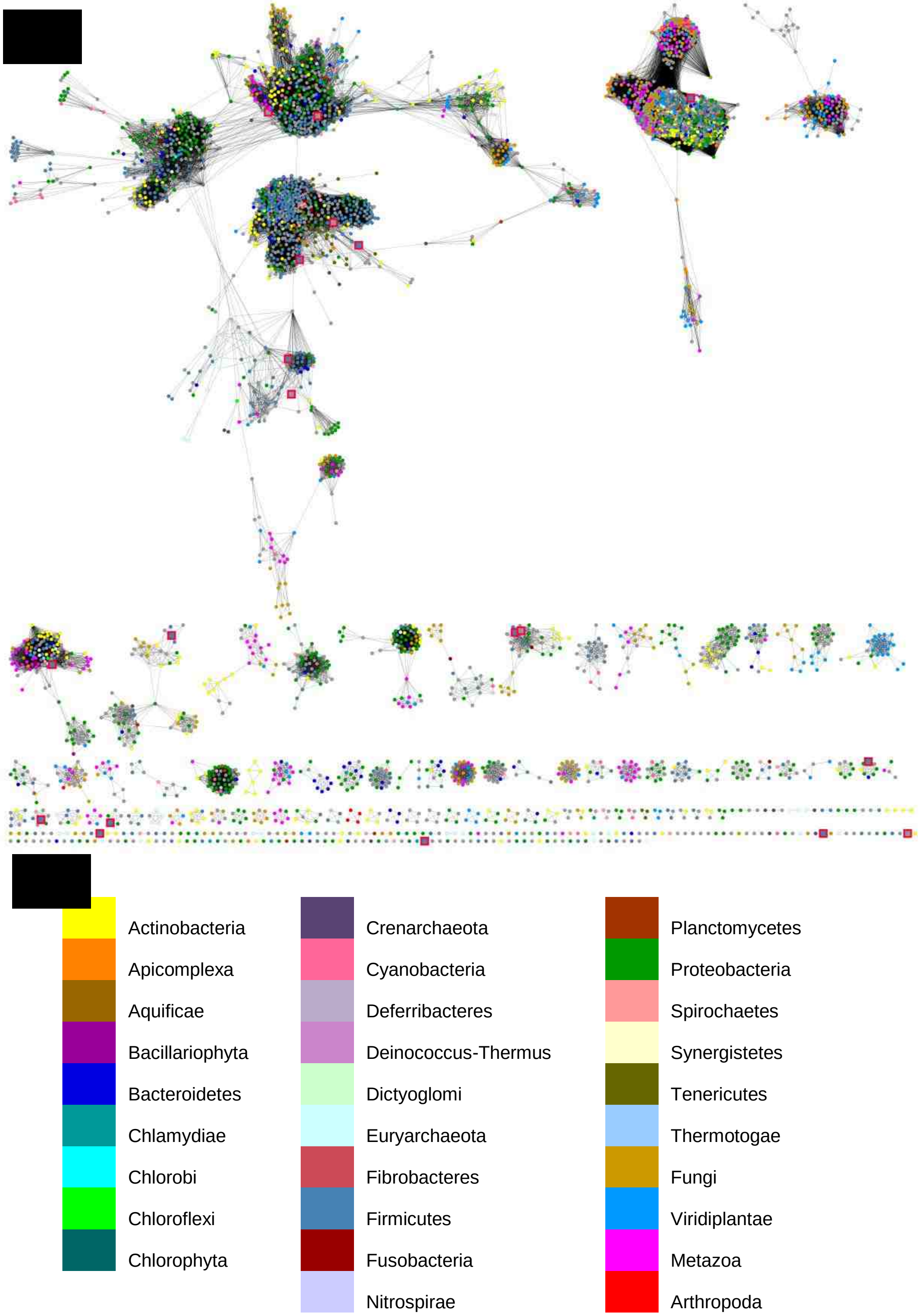

[redacted]
[redacted]

**Figure 3.10:** (A) 50% RNN of the HADSF downloaded from SFLD (last generated April 11 2014) with an edges E-value cutoff of $10^{-20}$. Orphans from the phylum-level Firmicutes walkout SSN are enlarged as squares with red borders. (B) Nodes are colored according to phylum/kingdom.

### *3.3.4: Function inference from Swiss-Prot annotations*

As shown extensively in Chapter 2, pre-existing data may be used to infer function or subfamily membership; this concept applies to the Firmicutes HADs as well. Based on the E-value cutoff for network generation, clusters are expected to be isofunctional, particularly if the cluster exhibits no extreme sub-clustering: Cluster 3 exhibits sub-clustering too extreme to apply annotations beyond a sub-cluster whereas Cluster 14 does not. Though there are not as many sequences involved in the Firmicutes HAD SSN, and hence there are not as many unannotated sequences of concern, it is nonetheless significant to be able to assign function to previously unassigned sequences based on cluster membership with sequences of known function. Records from the manually curated Swiss-Prot database were used to paint a SSN according to function (Figure 3.11) and to assign functional annotations across the clusters in Table 3.1.

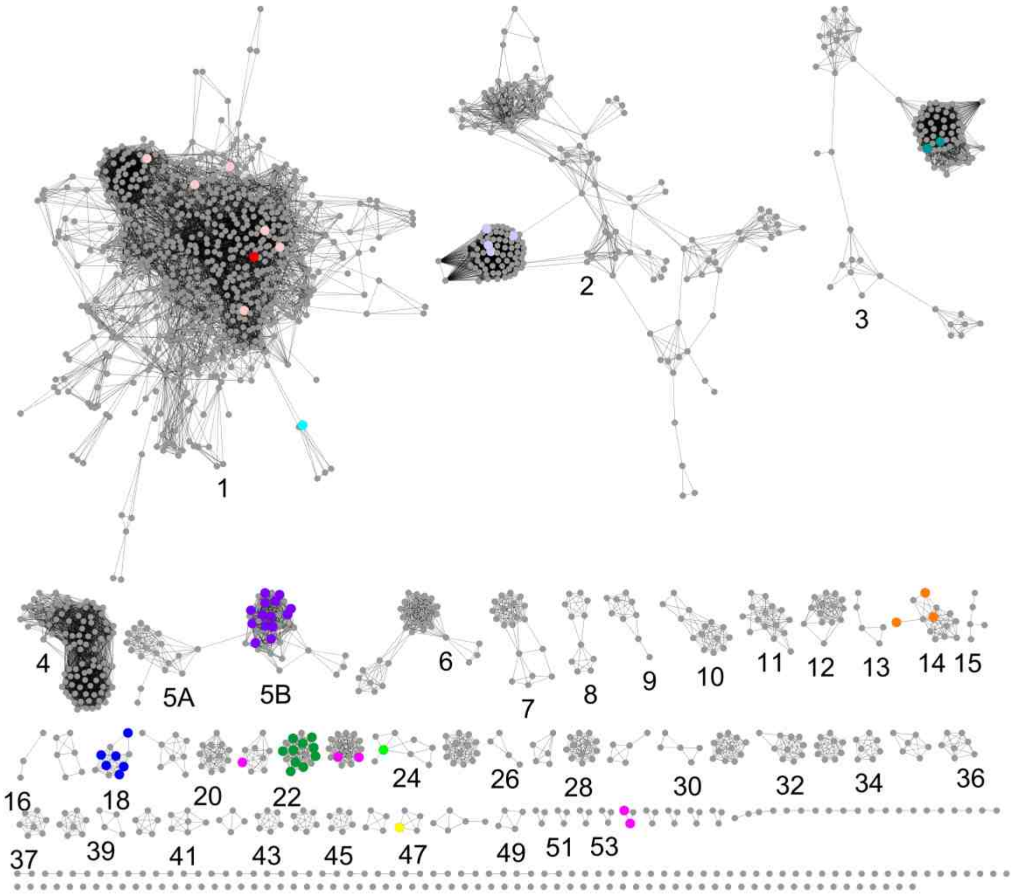


**Figure 3.11:** Firmicutes SSN painted according to Swiss-Prot annotations. Nodes with annotations are enlarged and colored as follows: NagD (lilac), ß-PGM (turquoise), PpaX (purple), 5'(3')-deoxyribonucleotidase (orange), Phosphonoacetaldehyde phosphonohydrolase (blue), Putative nucleotidase (magenta), HK-MTPenyl-1-P phosphatase (green), HBP phosphatase (lime green), AraL (yellow), Stress response protein YhaX (red), Kanosamine-6-phosphate phosphatase (cyan), putative phosphatase (peach). Those with function annotations applied to members of a cluster are again referenced in Table 3.4.

| Cluster | # members | | Function annotation from Swiss-Prot |
|---|---|---|---|
| 2 (subcluster) | 72 | ⬜ | NagD |
| 3 (subcluster) | 60 | 🟦 | β-PGM |
| 5B | 36 | 🟪 | PpaX |
| 14 | 13 | 🟧 | 5'(3')-deoxyribonucleotidase |
| 18 | 10 | 🟦 | Phosphonoacetaldehyde phosphonohydrolase |
| 21 | 8 | 🟪 | Putative nucleotidase |
| 22 | 15 | 🟩 | HK-MTPenyl-1-P phosphatase |
| 23 | 15 | 🟪 | Putative nucleotidase |
| 24 | 8 | 🟩 | HBP phosphatase |
| 47 | 4 | 🟨 | AraL |
| 54 | 3 | 🟪 | Putative nucleotidase |

**Table 3.4:** Clusters annotated according to Swiss-Prot annotations of member nodes, as well as the number of members in the cluster. Abbreviations are: ß-PGM (ß-phosphoglucomutase), HK-MTPenyl-1-P phosphatase (2-hydroxy-3-keto-5-methylthiopentenyl-1-phosphate phosphatase), HBP phosphatase (D,D-heptose 1,7-bisphosphate phosphatase).

### *3.3.5: Multi-domain and fusion sequences in the SSN*

Approximately one third of the sequences identified as being multi-domain or fusion sequences through SUPERFAMILY and STRING cluster together in the central area of Cluster 1, some of which correspond to a small cluster of longer sequences (400-500 a.a.) in the same region (Figure 3.12). The multi-domain or fusion sequences fall into two categories: with additional HAD domains or associating with cyclophilin-like domains. However, beyond both belonging to Cluster 1, which contains exclusively Cof hydrolase subfamily members, there is no discernible clustering pattern (length, to the HAD vs cyclophilin multi-domain sequences, etc.). Unfortunately, the Cof hydrolase family is very large and minimally characterized, so little can be inferred from the Cluster 1 multi-domain proteins, except that they are intriguing targets for further investigation.

A number of sequences noted as multi-domain sequences are situated as unusually lengthy outliers in clusters predominantly comprised of shorter sequences. These individual sequences and the clusters they inhabit would also be particularly attractive targets.

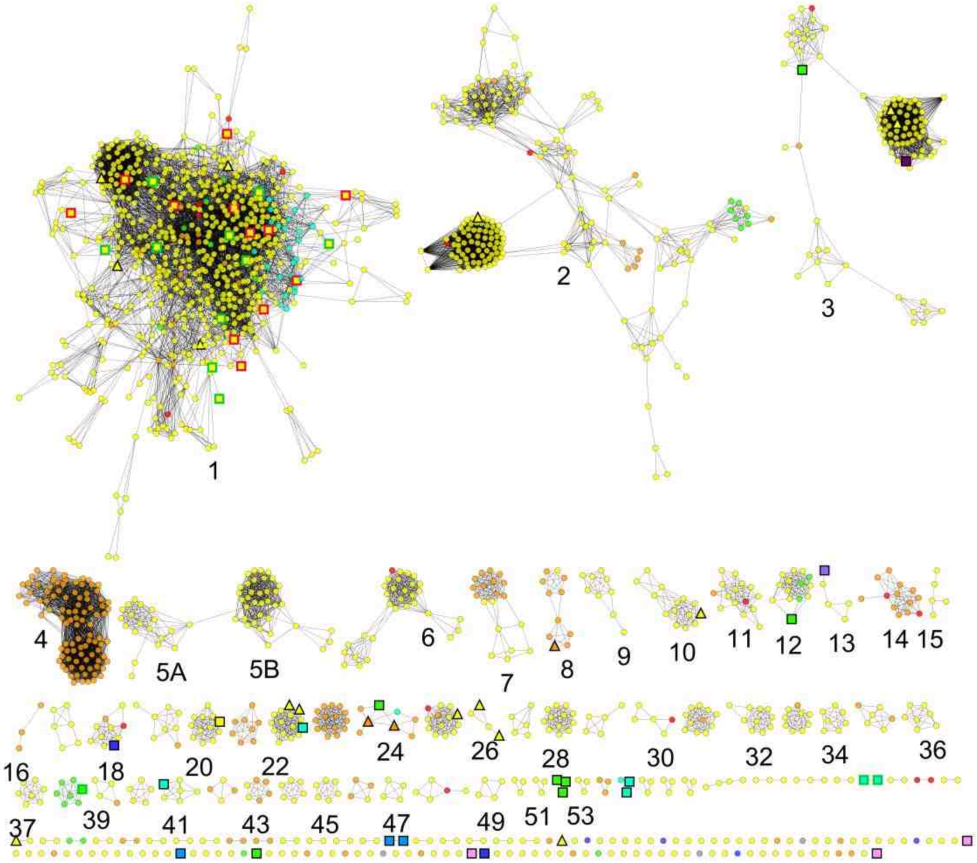


**Figure 3.12:** SSN painted according to length and noted domain information. Sequences that are themselves multi-domain proteins (square) and sequences demonstrated to have orthologs involved in fusion proteins (triangle) are enlarged with thick borders. In Cluster 1, border color represents sequences with two or more HAD domains (green) vs. cyclophilin-like domains (red). Nodes are colored according to residue length: <100 (red), 100-199 (orange), 200-299 (yellow),

300-399 (green), 400-499 (cyan), 500-599 (light blue), 600-699 (dark blue), 700-799 (purple), 800-899 (magenta), 900-999 (dark purple), >1000 (pink).

Cluster 22 is of interest because it contains one multi-domain protein and two sequences noted as having fusion protein orthologs, while the rest of the cluster sequences are of typical lengths for HADSF members. The multi-domain sequence (UniProt: Q65KJ7) is described in Swiss-Prot as having bifunctional HK-MTPenyl-1-P phosphatases (MtnX) methylthioribulose-1-phosphate dehydratase (MtnB) activity, which is the form fusion protein orthologs take for the other two noted sequences. Thus, this cluster is linked to two of the steps in the methionine salvage pathway: the dehydratase (MtnB) and phosphatase (MtnX) steps in converting 5-methylthioribulose-1-phosphate to 2-keto-4-methylthiobutyrate (*23*). A brief investigation of other members of this cluster reveals that those with entries in the NCBI Gene database (UniProt ID/genomic sequence: Q819E7/NC_004722.1 and Q5L1E1/NC_006510.1) contain an MtnB gene adjacent to the query sequence. The same immediate gene context and functionality can thus be expected for all members of this cluster.

Cluster 24 is interesting for similar reasons: it contains a multi-domain sequence and two sequences with fusion protein orthologs as well as a sequence annotated as an HBP phosphatase, also called GmhB. The additional domains for the multi-domain sequence and one of the fusion protein orthologs are a sugar transferase and isomerase, respectively; the other fusion protein ortholog is combined with mannose-1-phosphate guanylyltransferase or adjacent to UDP-glucose-4-epimerase. Together, this suggests that members of this cluster are ancestors or otherwise close relatives of bifunctional phosphatase/epimerase GmhB/A (*24*).

Three clusters contain one or two multi-domain sequences among sequences lacking multi-domain annotations while retaining similar lengths as the multi-domain sequence (clusters 20, 38, 55). Because the sequence lengths are so similar in these cases, it is likely that the “single” domain sequences are actually multi-domain sequences that were overlooked by the automatic domain detection used by SUPERFAMILY. The second domain and its annotation should be applied to such sequences.

### *3.3.6: Biological range(s) and gene context(s) of HTS-identified proteins*

HTS screens of prokaryotic HADSF members chosen as comprehensive samples based on diverse structure and function (*19*) revealed several proteins with potentially novel function, both within the Firmicutes phylum and without. We generated biological ranges and inspected gene contexts for these targets. Gene context findings are summarized in Table 3.5 and phylogenetic tree representations of biological range are found in Appendix A.1.5.

| EFI ID (UniProt) | Species | HTS activity (*19*) | # | Context (% orthologs exhibiting context) |
|---|---|---|---|---|
| **501036 (Q88RS0)** | *Pseudomonas putida KT2440* | D,D-heptose-1,7-bisphosphate, fructose-1,6-bisphosphate | 418 | Conserved at 60% across classes, 100% within Pseudomondales (phospholipid/glycerol acyltransferase, glycyl-tRNA synthetase subunit beta, glycyl-tRNA synthetase subunit alpha) and conserved at 60% in Pseudomondales only (Potassium uptake transporter, ribosomal RNA small subunit methyltransferase RsmB, methionyl-tRNA formyltransferase, peptide deformylase). |
| ***501163 (Q926W0)** | *Listeria innocua Clip11262* | 5-6 carbon alcohol sugars | 31 | Conserved at 90-100% within Listeria: phosphotransferase system mannitol-specific enzyme IIA, ROK family protein, PTS system, mannitol-specific IIBC subunit, oxidoreductase, Gfo/Idh/MocA family. |
| **501172 (P77366)** | *Escherichia coli str. K-12 substr. MG1655* | BPGM | 506 | Across the entire range: kojibiose or maltose glycosyl hydrolase (80%), ABC transporter (32%) |
| **501236 (Q8A5Q8)** | *Bacteroides thetaiotaomicron VPI-5482* | AMP, GMP | 240 | Maximum context conservation is on the genus level, with 30-45% neighbor conservation.  Due to the minimal conservation, neighbors are not reported |
| **501272 (Q81IN0)** | *Bacillus cereus ATCC 14579* | Phosphocholine | 71 | 4-aminobutyrate aminotransferase (94%), PAS domain S-box protein (93%), succinate-semialdehyde dehydrogenase (93%), Lipid kinase (92%), aspartyl/glutamyl-tRNA amidotransferase subunits A (92%), subunit B (92%), subunit C (93%), aminopeptidase 2 (92%), polysaccharide deacetylase (92%), nucleoside permease nupC (91%), hypothetical two domain protein (87%), multidrug resistance protein smr (87%), ABC transporter substrate-binding protein (70%), |
| **501279 (P32662)** | *Escherichia coli str. K-12 substr. MG1655* | 5-carbon acid sugars | 757 | tryptophanyl-tRNA synthetase (60%), ribulose-phosphate 3-epimerase (77%), DNA adenine methylase (47%), sporulation and cell division repeat protein (33%), 3-dehydroquinate synthase (48%), Shikimate kinase (~50%), type IV pilus secretin PilQ (45%) |

| | | | | |
|---|---|---|---|---|
| **501365 (Q9K8N3)** | *Bacillus halodurans C-125* | phoshposerine, phosphothreonine | 9 | Difficult to say due to small sample number. Highest conserved neighbor is alkylphosphonate ABC transporter ATP-binding protein (8/9), followed by components of alkylphosphonate ABC transporter permease (5/9) |
| **502337 (Q11S56)** | *Cytophaga hutchinsonii strain ATCC 33406* | sorbitol-1-phosphate, 3'-deoxy-sorbitol-6-phosphate | 9 | Conserved in Bacteroidetes: hypothetical protein (100%), L-glutamine synthetase or ligase (77%) |
| **900338 (Q4K5L5)** | *Pseudomonas fluorescens Pf-5* | pNPP, pyridoxal-5P | 160 | The following genes are conserved in Pseudomonas: acyl-CoA thieosterase (90%), GNAT family acetyl transferase (90%), histone deacetylase (85%), flavoprotein (90%), DNA/RNA helicase (80%) |

**Table 3.5:** Gene context summary of HADSF members identified via HTS screens from the EFI. EFI ID and UniprotKB ID are given in the first column, followed by species and highest-scoring HTS results from the screening (*19*). The # column indicates how many orthologs above 40% SI and 80% query coverage were found; the biological range of these orthologs is presented I the appendix. Gene context is presented as the conserved sequence annotation with the percentage of orthologs in which it was conserved in parentheses. * Indicates a record that is discussed further below.

Of the HTS-identified queries, an enzyme annotated “hypothetical protein” from *Listeria innocua* (Uniprot: Q926W0 EFI: 501163) was revealed to have high activity with five- or six-carbon alcohol sugars. The enzyme is limited to Firmicutes, with orthologs appearing primarily in the *Listeria* and *Bacillus* families (Figure 3.13).

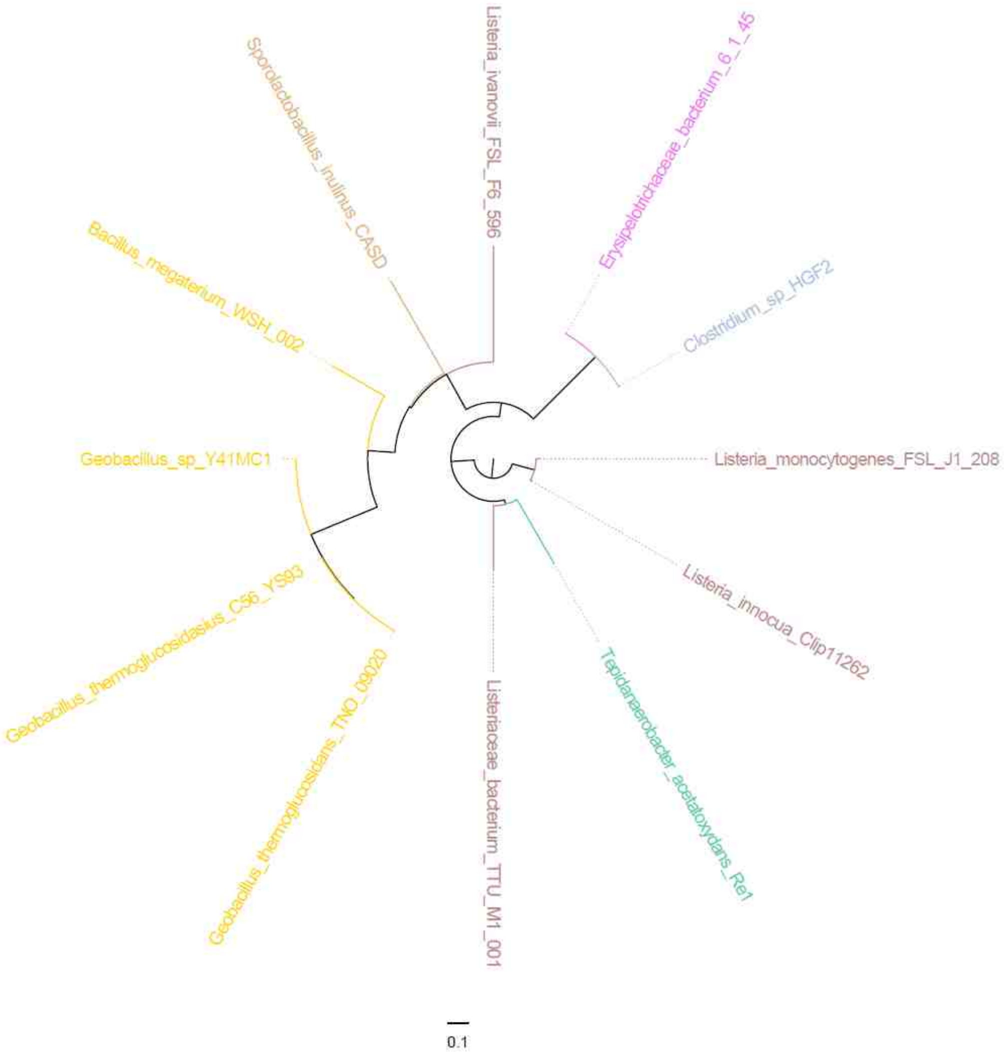

Listeria_ivanovii_FSL_F6_596
Erysipelotrichaceae_bacterium_6_1_45
Clostridium_sp_HGF2
Listeria_monocytogenes_FSL_J1_208
Listeria_innocua_Clip11262
Tepidanaerobacter_acetatoxydans_Re1
Listeriaceae_bacterium_TTU_M1_001
Geobacillus_thermoglucosidans_TNO_09020
Geobacillus_thermoglucosidasius_C56_YS93
Geobacillus_sp_Y41MC1
Bacillus_megaterium_WSH_002
Sporolactobacillus_inulinus_CASD
0.1

**Figure 3.13:** Phylogenetic tree representing the biological range of Q926W0 from *Listeria innocua* limited to highest % SI orthologs in each species and colored according to Family. *Listeriaceae* (brown), *Thermoanaerobacteriaceae* (teal), *Clostridiaceae* (steel), *Sporolactobacillaceae* (tan), *Erysipelotrichaceae* (magenta), and *Bacillaceae* (gold).

Gene context for these orthologs is conserved primarily within the genus *Listeria* and reveals that the encoding gene is adjacent to members of the phophoenolpyruvate:carbohydrate phosphotransferases (PTS) system (Table 3.5). The PTS system transports and phosphorylates carbohydrates and is involved in both catalysis and regulation, including regulation of carbohydrate flux within the cell (*25-27*). In particular, it appears in the neighborhood of mannitol-specific enzymes IIA and IIBC, a ROK transcription factor, an oxidoreductase, and a Gfo/Idh/MocA family/MviM sugar dehydrogenase. Based on the HTS and gene context results, it is now hypothesized that this HADSF member dephosphorylates a sugar, likely xylitol or mannitol, for usage in the PTS system (*19*).

|  | Q926W0 ortholog | PTS system, mannitol-specific, IIA component |  | ROK family protein |  | RpiR transcription regulator |  | PTS system, mannitol-specific, IIBC component |  | Gfo/Idh/MocA family oxidoreductase |  |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Species | %SI | %SI | distance | %SI | distance | %SI | distance | %SI | distance | %SI | distance |
| **Listeria innocua Clip11262** | 100% | 100% | 1 | 100% | 2 | 100% | 3 | 100% | -1 | 100% | -2 |
| Listeria innocua ATCC 33091 | 99% | 99% | -1 | 99% | -2 | 100% | -3 | 99% | 1 | 99% | 2 |
| Listeria innocua FSL J1-023 | 98% | 97% | 1 | 96% | 2 | 100% | 3 | 100% | -1 | 99% | -2 |
| Listeria monocytogenes FSL J1-208 | 97% | 93% | 1 | 91% | 2 | 99% | 3 | 99% | -1 | 99% | -2 |
| Listeria monocytogenes FSL J2-064 | 95% | 94% | -1 | 91% | -2 | 99% | -3 | 99% | 1725 | 99% | 1724 |
| Listeria monocytogenes str. 1/2a F6854 | 95% | 93% | 1 | 91% | 2 | 99% | 3 | 99% | -1 | 99% | -2 |
| Listeria monocytogenes serotype 4b str. CLIP 80459 | 94% | 95% | 1 | 91% | 2 | 99% | 3 | 99% | -1 | 99% | -2 |
| Listeria monocytogenes serotype 4b str. F2365 | 94% | 95% | 1 | 90% | 2 | 99% | 3 | 99% | -1 | 99% | -2 |
| Listeria monocytogenes 08-5578 | 94% | 93% | -1 | 91% | -2 | 99% | -3 | 99% | 1 | 99% | 2 |
| Listeria monocytogenes EGD-e | 94% | 93% | 1 | 91% | 2 | 99% | 3 | 99% | -1 | 99% | -2 |
| Listeria monocytogenes Finland 1998 | 94% | 93% | 1 | 91% | 2 | 99% | 3 | 99% | -1 | 99% | -2 |
| Listeria monocytogenes 10403S | 94% | 93% | 1 | 91% | 2 | 99% | 3 | 99% | -1 | 99% | -2 |
| Listeria monocytogenes FSL J2-071 | 94% | 94% | 1 | 89% | 2 | 99% | 3 | 98% | -1 | 99% | -2 |
| Listeria monocytogenes FSL J2-003 | 94% | 91% | 317 | 91% | 316 | 99% | 315 | 99% | -1 | 99% | -2 |
| Listeria monocytogenes FSL J1-194 | 94% | 95% | -1 | 90% | -2 | 99% | -3 | 99% | 1 | 99% | 2 |
| Listeria monocytogenes str. 4b H7858 | 94% | 95% | 1 | 90% | 2 | 99% | 3 | 99% | -1 | 99% | -2 |
| Listeria monocytogenes FSL F2-515 | 93% | 93% | -1 | 27% | -435 | 99% | 506 | 99% | -890 | 98% | 782 |
| Listeriaceae bacterium TTU M1-001 | 58% | 62% | 1 | 31% | -2113 | 64% | 2 | 85% | -1 | 86% | -2 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Tepidanaerobacter acetatoxydans Re1 | 54% | 53% | 1 | 36% | 3 | 34% | -39 | 68% | -1 | 77% | -2 |
| Clostridium sp. HGF2 | 49% | 50% | -109 | 39% | -30 | 28% | -450 | 58% | -90 | 80% | -111 |
| Erysipelotrichaceae bacterium 6_1_45 | 49% | 50% | -1 | 39% | -2 | 29% | 1635 | 58% | 1 | 80% | 2 |
| Erysipelotrichaceae bacterium 3_1_53 | 49% | 55% | 1 | 40% | 2 | 31% | -2993 | 59% | -1 | 79% | -2 |
| Erysipelotrichaceae bacterium 2_2_44A | 48% | 50% | -1 | 39% | -2 | 31% | -3557 | 58% | 1 | 81% | 2 |
| Sporolactobacillus inulinus CASD | 41% | 29% | -486 | 28% | -378 | 24% | -1783 | 51% | -486 | 19% | -1461 |
| Bacillus megaterium WSH-002 | 40% | 45% | 155 | 32% | 851 | 24% | -124 | 51% | -655 | 28% | 1273 |
| Bacillus megaterium QM B1551 | 40% | 44% | -155 | 26% | -3611 | 26% | -15 | 51% | 596 | 28% | 354 |
| Geobacillus thermoglucosidasius C56-YS93 | 40% | 30% | -2007 | 25% | -986 | 26% | -2026 | 52% | -2009 | 22% | -1779 |
| Geobacillus sp. Y4.1MC1 | 40% | 40% | -2066 | 25% | -916 | 26% | -1969 | 53% | -1955 | 27% | -1898 |
| Geobacillus thermoglucosidans TNO-09.020 | 40% | 41% | -1934 | 25% | -829.1 | 26% | -1841 | 53% | -1825 | 27% | -1768 |
| Listeria ivanovii FSL F6-596 | 40% | 30% | 1552 | 27% | 1695 | 28% | 667 | 28% | 1645 | 25% | 97 |

**Table caption:** Relevant gene context for Q926W0, illustrating the biological range and sequence identity of orthologs as well as the sequence identity of any orthologs to Q926W0's neighbors and, if present, their distance from the matching Q926W0 ortholog.  Species are colored according to family, except for the query protein (highlighted in green): Listeriaceae (brown), Thermoanaerobacteraceae (teal), Clostridiaceae (steel), Sporolactobacillaceae (tan), Erysipelotrichaceae (magenta), Bacillaceae (gold).

## 3.4: Conclusions

Exploration of the Firmicutes HADSF sequence space was facilitated by the creation of a sequence similarity network of representative members. This network demonstrates the great diversity across Firmicutes families and within genera, ranging from 10-30 members in the widest case. Several HADSF members are revealed to be identifiable copies from gene duplication event(s); these copies are members of the same species with >40% sequence identity, suggesting isofunctionality and gene duplication. Family-level sequence similarity networks also reveal orphan sequences that are highly divergent from their native family. Some of these may be the result of gene transfer events from other Firmicutes families or from separate taxa entirely—those orphans associating with other taxa on the SSN containing all known HADs are particularly likely candidates for gene transfer and should be further pursued.

Protein family-level sequence similarity networks reveal the relatively close sequence identity relationship between NagD and ß-PGM HAD sequences as well as the general sub-clustering behavior of the other HADSF members. Based on co-clustering, several "HAD-like" may be classified as ß-PGM and unclassified members of annotated clusters may also be annotated.

Several fusion proteins and fusion protein orthologs were identified, particularly MtnX/MtnB fusion proteins and yet-unfused orthologs in cluster 22, as well as GmhB/GmhA in cluster 24. Additional clusters 20, 38, and 55 were identified as probable multi-domain clusters that should be further investigated.

The co-clustering of single and multiple domain proteins provides insight into when gene fusion events may have occurred.

Biological range and gene context were used to identify *Listeria innocua* protein Q926W0 as a member of the PTS system, likely acting on xylitol or mannitol. Gene contexts for other sequences were also identified and may be used in the future for function inference or guides for function discovery.

### 3.5: References

1. Tremaroli, V.; Backhed, F., Functional interactions between the gut microbiota and host metabolism. *Nature* **2012,** *489*, 242-9.
2. Chow, J.; Lee, S. M.; Shen, Y.; Khosravi, A.; Mazmanian, S. K., Host-bacterial symbiosis in health and disease. *Advances in immunology* **2010,** *107*, 243-74.
3. Bik, E. M., Composition and function of the human-associated microbiota. *Nutrition reviews* **2009,** *67 Suppl 2*, S164-71.
4. Schippa, S.; Conte, M. P., Dysbiotic events in gut microbiota: impact on human health. *Nutrients* **2014,** *6*, 5786-805.
5. Frank, D. N.; Zhu, W.; Sartor, R. B.; Li, E., Investigating the biological and clinical significance of human dysbioses. *Trends in microbiology* **2011,** *19*, 427-34.
6. Turnbaugh, P. J.; Gordon, J. I., The core gut microbiome, energy balance and obesity. *The Journal of physiology* **2009,** *587*, 4153-8.

7. Ley, R. E.; Backhed, F.; Turnbaugh, P.; Lozupone, C. A.; Knight, R. D.; Gordon, J. I., Obesity alters gut microbial ecology. *Proc Natl Acad Sci U S A* **2005,** *102*, 11070-5.

8. Hawrelak, J. A.; Myers, S. P., The causes of intestinal dysbiosis: a review. *Alternative medicine review : a journal of clinical therapeutic* **2004,** *9*, 180-97.

9. Sekirov, I.; Russell, S. L.; Antunes, L. C.; Finlay, B. B., Gut microbiota in health and disease. *Physiological reviews* **2010,** *90*, 859-904.

10. Eckburg, P. B.; Bik, E. M.; Bernstein, C. N.; Purdom, E.; Dethlefsen, L.; Sargent, M.; Gill, S. R.; Nelson, K. E.; Relman, D. A., Diversity of the human intestinal microbial flora. *Science* **2005,** *308*, 1635-8.

11. Gough, J.; Karplus, K.; Hughey, R.; Chothia, C., Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J Mol Biol* **2001,** *313*, 903-19.

12. Szklarczyk, D.; Franceschini, A.; Kuhn, M.; Simonovic, M.; Roth, A.; Minguez, P.; Doerks, T.; Stark, M.; Muller, J.; Bork, P.; Jensen, L. J.; von Mering, C., The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research* **2011,** *39*, D561-8.

13. UniProt, C., UniProt: a hub for protein information. *Nucleic acids research* **2015,** *43*, D204-12.

14. Papadopoulos, J. S.; Agarwala, R., COBALT: constraint-based alignment tool for multiple protein sequences. *Bioinformatics* **2007,** *23*, 1073-9.

15. Blattner, F. R.; Plunkett, G., 3rd; Bloch, C. A.; Perna, N. T.; Burland, V.; Riley, M.; Collado-Vides, J.; Glasner, J. D.; Rode, C. K.; Mayhew, G. F.; Gregor, J.; Davis, N. W.; Kirkpatrick, H. A.; Goeden, M. A.; Rose, D. J.; Mau, B.; Shao, Y., The complete genome sequence of Escherichia coli K-12. *Science* **1997,** *277*, 1453-62.

16. The NCBI C++ Toolkit Book [Internet]. http://www.ncbi.nlm.nih.gov/toolkit/doc/book/toolkit.fm

17. Smoot, M. E.; Ono, K.; Ruscheinski, J.; Wang, P. L.; Ideker, T., Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **2011,** *27*, 431-2.

18. Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T., Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* **2003,** *13*, 2498-504.

19. Huang, H.; Pandya, C.; Liu, C.; Al-Obaidi, N. F.; Wang, M.; Zheng, L.; Toews Keating, S.; Aono, M.; Love, J. D.; Evans, B.; Seidel, R. D.; Hillerich, B. S.; Garforth, S. J.; Almo, S. C.; Mariano, P. S.; Dunaway-Mariano, D.; Allen, K. N.; Farelli, J. D., Panoramic view of a superfamily of phosphatases through substrate profiling. *Proc Natl Acad Sci U S A* **2015**.

20. Akiva, E.; Brown, S.; Almonacid, D. E.; Barber, A. E., 2nd; Custer, A. F.; Hicks, M. A.; Huang, C. C.; Lauck, F.; Mashiyama, S. T.; Meng, E. C.; Mischel, D.; Morris, J. H.; Ojha, S.; Schnoes, A. M.; Stryke, D.; Yunes, J.

M.; Ferrin, T. E.; Holliday, G. L.; Babbitt, P. C., The Structure-Function Linkage Database. *Nucleic acids research* **2014,** *42*, D521-30.

21. Thaller, M. C.; Schippa, S.; Bonci, A.; Cresti, S.; Rossolini, G. M., Identification of the gene (aphA) encoding the class B acid phosphatase/phosphotransferase of Escherichia coli MG1655 and characterization of its product. *FEMS microbiology letters* **1997,** *146*, 191-8.

22. Grose, J. H.; Bergthorsson, U.; Xu, Y.; Sterneckert, J.; Khodaverdian, B.; Roth, J. R., Assimilation of nicotinamide mononucleotide requires periplasmic AphA phosphatase in Salmonella enterica. *J Bacteriol* **2005,** *187*, 4521-30.

23. Nakano, T.; Ohki, I.; Yokota, A.; Ashida, H., MtnBD is a multifunctional fusion enzyme in the methionine salvage pathway of Tetrahymena thermophila. *PLoS One* **2013,** *8*, e67385.

24. Valvano, M. A.; Messner, P.; Kosma, P., Novel pathways for biosynthesis of nucleotide-activated glycero-manno-heptose precursors of bacterial glycoproteins and cell surface polysaccharides. *Microbiology* **2002,** *148*, 1979-89.

25. Postma, P. W.; Lengeler, J. W.; Jacobson, G. R., Phosphoenolpyruvate:carbohydrate phosphotransferase systems of bacteria. *Microbiological reviews* **1993,** *57*, 543-94.

26. Deutscher, J.; Ake, F. M.; Derkaoui, M.; Zebre, A. C.; Cao, T. N.; Bouraoui, H.; Kentache, T.; Mokhtari, A.; Milohanic, E.; Joyet, P., The

bacterial phosphoenolpyruvate:carbohydrate phosphotransferase system: regulation by protein phosphorylation and phosphorylation-dependent protein-protein interactions. *Microbiology and molecular biology reviews : MMBR* **2014,** *78*, 231-56.

27. Gabor, E.; Gohler, A. K.; Kosfeld, A.; Staab, A.; Kremling, A.; Jahreis, K., The phosphoenolpyruvate-dependent glucose-phosphotransferase system from Escherichia coli K-12 as the center of a network regulating carbohydrate flux in the cell. *European journal of cell biology* **2011,** *90*, 711-20.

# CHAPTER 4

# CONVERGENT AND DIVERGENT EVOLUTION IN HADSF PHOSPHATASES FROM *E. COLI* AND *BACTERIODES THETAIOTAOMICRON*

## 4.1: Introduction

### 4.1.1: *Flavin mononucleotide synthesis in* E. coli *and* Bacteroides thetaiotaomicron

Riboflavin, vitamin $B_2$, is converted to its active forms of flavin mononucleotide (FMN) and flavin adenine dinucleotide (FAD), both of which are important to health and are used as electron carriers and flavoprotein cofactors (*1*). Riboflavin kinase phosphorylates riboflavin to form FMN, which is further converted to FAD by FAD synthetase; riboflavin kinase and FAD synthetase are known to co-exist in bifunctional enzymes in bacteria, shown in Figure 4.1 (*2-4*). In *E. coli*, they are encoded by the gene ribF (UniProtKB: P0AG40). The conversion of riboflavin and ATP to FMN and ADP is catalyzed by riboflavin kinase while the reverse reaction is catalyzed by FMN phospahtases/hydrolases (*5*). The two functionalities have been reported to co-exist in bifunctional enzymes (*6*).

Two HADSF proteins, yigB in *E. coli* (UniProt: P0ADP0, EFI: 501262) and BT2542 in *B. thetaiotaomicron* (UniProt: Q8A4Q5, EFI: 501088), have been identified as probable flavin mononucleotide phosphatases, which has been supported by recent publications (*7, 8*) and previous work in this lab (*9*).

However, the two proteins have poor sequence identity to one another, do not share gene context and, despite both being members of the HADSF, do not share the typical DxD motif considered a defining feature of the HADSF as described in Section 1.5.  YigB contains the motif but BT2542 is missing the general acid/base Asp residue, instead containing a DxG motif (Figure 4.2).

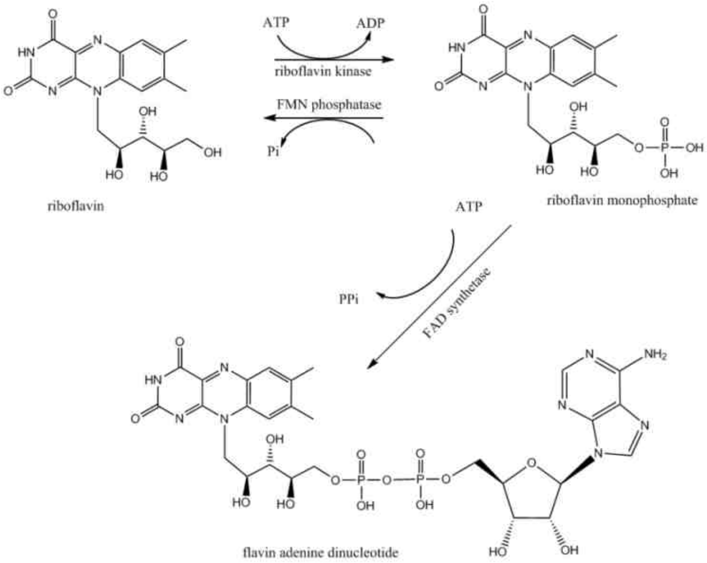


**Figure 4.1**: Conversion of riboflavin to FMN and FAD, displayed in ChemDraw15.

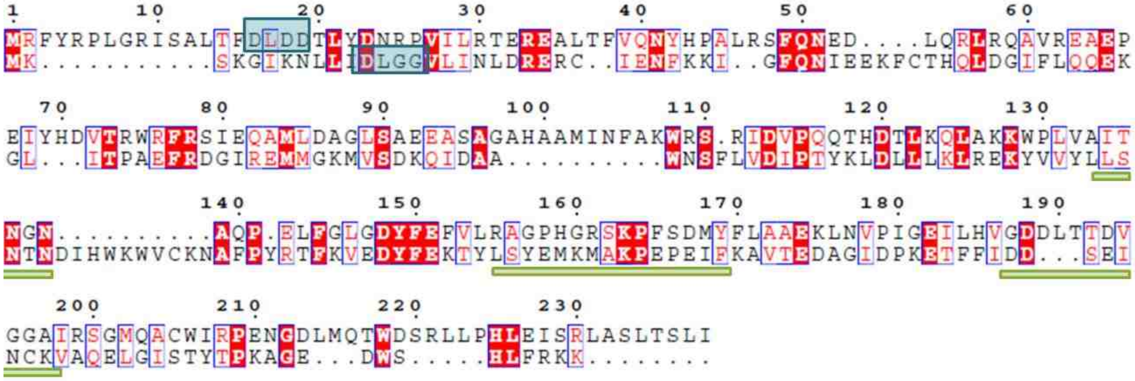

1 10 20 30 40 50 60
MRFYRPLGRISALTFDLDDTLYDNRPVILRTEREALTFVQNYHPALRSFQNED....LQRLRQAVREAEP
MK...........SKGIKNLLIDLGGVLINLDRERC..IENFKKI..GFQNIEEKFCTHQLDGIFLQQEK
70 80 90 100 110 120 130
EIYHDVTRWRFRSIEQAMLDAGLSAEEASAGAHAAMINFAKWRS.RIDVPQQTHDTLKQLAKKWPLVAIT
GL...ITPAEFRDGIREMMGKMVSDKQIDAA..........WNSFLVDIPTYKLDLLLKLREKYVVYLLS
140 150 160 170 180 190
NGN..........AQP.ELFGLGDYFEFVLRAGPHGRSKPFSDMYFLAAEKLNVPIGEILHVGDDLTTDV
NTNDIHWKWVCKNAFPYRTFKVEDYFEKTYLSYEMKMAKPEPEIFKAVTEDAGIDPKETFFIDD...SEI
200 210 220 230
GGAIRSGMQACWIRPENGDLMQTWDSRLLPHLEISRLASLTSLI
NCKVAQELGISTYTPKAGE...DWS.....HLFRKK........

**Figure 4.2:** Alignment of yigB (top) with BT2542 (bottom) visualized by ESPript: http://espript.ibcp.fr (*10*).  The canonical DxD motif region is boxed in blue, the three other HADSF motifs are underlined in green, conserved residues are red, and conserved residue types are boxed.  The two sequences share only 15.6% sequence identity.

In addition to these two proteins, HTS results from the EFI (*8*), previous work done in this lab (*9*), and the literature (*7*) indicate that three additional HADSF members have specific activity with FMN— ybjI in *E. coli* (UniProt: P75809, EFI: 501335)— or promiscuous activities including FMN—yigL in *E. coli* (UniProt: P27848, EFI: 501312) and Q83SV5 in *Salmonella enterica serovar Typhi* (EFI: 501310). We tracked the biological ranges and gene contexts of these five putative FMN hydrolases, seeking evidence supporting their function assignment as well as their evolutionary relationship (Figure 4.3).

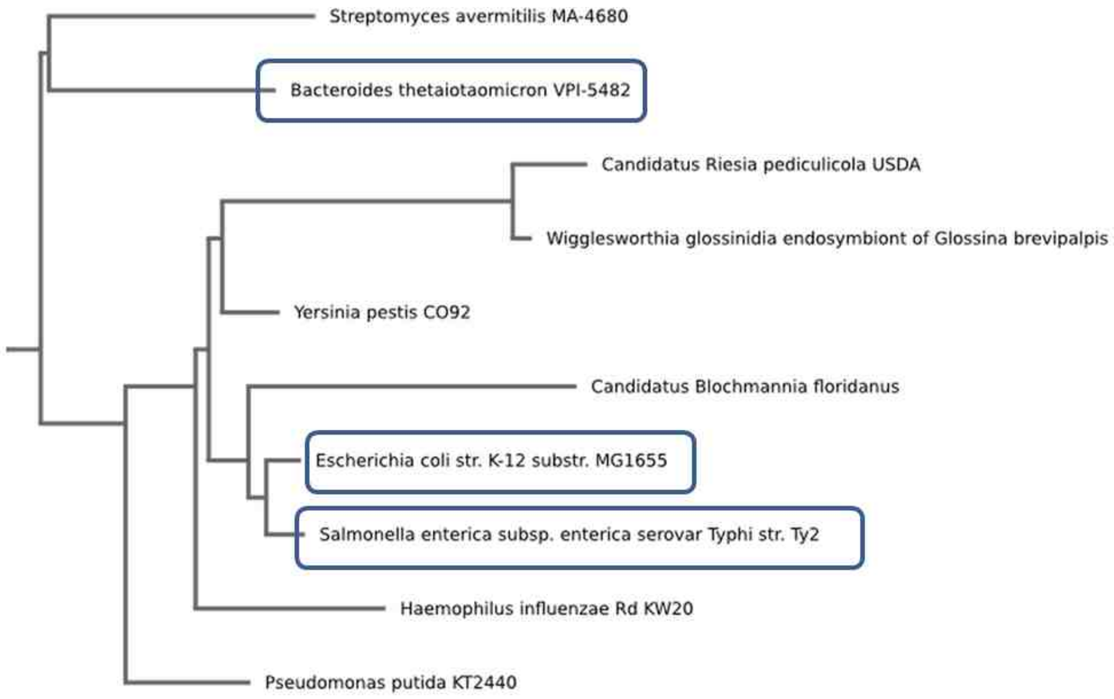

Streptomyces avermitilis MA-4680
Bacteroides thetaiotaomicron VPI-5482
Candidatus Riesia pediculicola USDA
Wigglesworthia glossinidia endosymbiont of Glossina brevipalpis
Yersinia pestis CO92
Candidatus Blochmannia floridanus
Escherichia coli str. K-12 substr. MG1655
Salmonella enterica subsp. enterica serovar Typhi str. Ty2
Haemophilus influenzae Rd KW20
Pseudomonas putida KT2440

**Figure 4.3:** Phylogenetic tree relating *E. coli, S. enterica,* and *B. thetaiotaomicron* (boxed), among other species. Generated using the Phylogenetic Tree tool at http://supfam.cs.bris.ac.uk/.

### *4.1.2: Comparable HADSF members in* E. coli *and* Bacteroides thetaiotaomicron

Previous work from the Dunaway-Mariano lab identified two other protein pairs with an interesting possible evolutionary linkage—yidA in *E. coli* and

BT3352 in *B. thetaiotaomicron*, both of which have high activity towards erythose 4-phosphate (*9, 11*). Substrate specificity profiles and query species gene contexts that were previously determined in the lab suggested that yidA's physiological substrate could be 2-keto-3-deoxy-6-phosphogalactonate (KDPG), a substrate with which BT3352 is not active. KDPG is an intermediate in the galactonate degradation pathway, the buildup of which is toxic *(12, 13).*

Given that the host organisms for yidA and BT3352 share an environment in the human gut, they may be subject to similar selective pressures or be subjects of gene transfer events (*14-16*). Either or both of these approaches may explain why, despite their taxonomic distance (Figure 4.3), they share such remarkable structure and activity similarities. This study explores the biological range and sequence identity-based relationship between the two proteins to elucidate the nature of their relationship.

## 4.2: Materials and methods

### *4.2.1: Generating taxonomic lineages*

We manually compiled a taxonomy database in Excel by copying the taxonomic lineages for each species of interest from the UniProtKB (*17*). After initial compilation, we used Excel to compare the genera of query species against the existing database. If a matching genus was found, its taxonomic data was applied to the new species; if not, the species taxonomic data was copied from the UniprotKB and added to the local taxonomy database.

### *4.2.2: Manual biological ranges for* E. coli *and* Salmonella *proteins*

Each query protein underwent a BLAST search (web interface, default parameters), retaining all hits with >80% query coverage and >35% sequence identity. Hits with sequence identities between 35% and 40% were retained only if they displayed gene context similar to those with higher sequence identities. The online NCBI tool COBALT (*18*) was used to make a multiple sequence alignment (MSA) of these retained hits. Each alignment was visually inspected and any sequences not adhering to the canonical DxD catalytic domain motif were removed (except in the case of BT2542 and yigB orthologs). Taxonomic lineages were assigned to each species using the taxonomic lineage database described in Section 4.2.1.

For visualization purposes, the ortholog list was pared down by removing species with multiple strains; the strain with the highest hit sequence identity was chosen as the representative strain for that species while the others were removed from the visualization. These pared down results were used to generate another phylogenetic tree and MSA using COBALT and default parameters. FigTree was used to modify and annotate the MSA [http://tree.bio.ed.ac.uk/software/figtree/]. The final, annotated MSAs were visualized in a circular format using IToL, the Interactive Tree of Life (*19*).

### *4.2.3: Gene context acquisition*

In general, gene context was determined by investigating available NCBI gene records for query proteins and their putative orthologs. At the time this research was conducted (2012), RefSeq records with YP_ or NP_ accession numbers were typically associated with gene records; these gene records included neighboring genes and their annotations, if any. In such cases, the five sequences on either side of the query gene were recorded. However, non-RefSeq records and predicted proteins (e.g. those with ZP_ accession numbers) did not have gene records available; in these cases, the sequence was run through the STRING protein-protein interaction database (*20*) and inspected for recurrence of neighborhood proteins either globally or conserved within a taxonomic grouping. It should be noted that as of May 2015, most many of the aforementioned RefSeq accession prefixes have been folded into a new WP_ prefix. In the cases of paired *E. coli* and *B. thetaiotaomicron* sequences (yigB/BT2542 and yidA/BT3352), specific gene context clues were sought, as described below.

Because previous studies in this lab suggested that yidA's physiological substrate could be the 2-keto-3-deoxygluconate 6-phosphate intermediate in the gluconate degradation pathway, we tracked co-occurrence of a pathway member. We tracked 2-oxo-3-deoxygalactonate kinase (dgoK in *E. coli*, UniProt: P31459) by running a BLAST of dgoK in any species containing yidA or BT3352 orthologs. If a dgoK ortholog was found, we checked whether it was in the neighborhood (+/- 10 genes away) of yidA/BT3352 orthologs.

In the course of determining gene contexts for the FMN-active proteins, we discovered that BT2542 orthologs frequently were found adjacent to riboflavin kinase, ribF. Subsequently, we interrogated species containing BT2542 for orthologs to the *B. thetaiotaomicron* ribF (UniProt: Q8A4Q4) by running BLAST searches for ribF in BT2542/yigB ortholog-containing species and noting proximity to BT2542 orthologs.

### *4.2.4: Biological range of paired* E. coli *and* B. thetaiotaomicron *proteins*

We determined biological range by running BLAST searches for the query sequences against individual taxonomic groups as provided by NCBI (typically, species were grouped by domain, phylum, then order; however, class was sometimes also included); this method is now defunct. Hits with scores >50.0 and query coverage >80.0% were retained and tabulated for each species, resulting in a sequence identity threshold in mid to upper 20%. These hits were compared, by species, between BT3352 and yidA as well as between BT2542 and yigB to determine a) whether the species contained potential orthologs to both proteins or only one and b) in the event that potential orthologs were present for both proteins, whether the potential ortholog was the same for both proteins.

### *4.2.5: Degree of anomalous motif conservation in BT2542*

A multiple sequence alignment of all BT2542 orthologs was generated using the COBALT (*16*). Each alignment was visually inspected and the region matching the BT2542 D+GGVL motif was extracted. Departures from this motif were given a score based on how many positions they share with the motif; a score of 1 indicates complete conservation, a score of 0.83 indicates a change in one of the six positions, a score of 0.667 indicates a change in two positions, etc. Departures from the motif were sorted according to residue type and tabulated.

## 4.3: Results and discussion

### *4.3.1: Biological range and co-orthologs of yidA and BT3352*

The two proteins share a similar biological range in archaea and bacteria; however, orthologs are few in archaea and sequence identities never exceed 29% (Table 4.1). In bacteria, both typically return the same top ortholog. Indeed, there is only one exception to this ortholog sharing for orthologs with greater than 30% sequence identity: order Dehalococcoidetes of Chloroflexi. Ortholog sharing is seen for both medium and moderate sequence identity. Together with the well-conserved structure this is suggestive of divergent evolution and potential horizontal gene transfer between members of Firmicutes and Gammaproteobacteria.

The dgoK gene and other members in the pathway were only found adjacent to or within the neighborhood of the query in ~50% of orthologs belonging to Enterobacteriales. No other taxonomic groups contained nearby dgoK orthologs and, indeed, several taxonomic groups did not contain any dgoK orthologs. Thus, gene context cannot be reliably used to support annotation of yidA orthologs as belonging to the galactonate pathway.

| | | | **BT3352** | **BT3352** | **yidA** | **yidA** | **Ortholog** |
|---|---|---|---|---|---|---|---|
| **Phylum** | **Class** | **Order** | **avg** | **max** | **avg** | **max** | **sharing?** |
| Actinobacteria | Actinobacteria | Actinomycetales | 30.4 | 32.4 | 30.9 | 32.9 | y |
| Actinobacteria | Actinobacteria | Coriobacteriales | 31.3 | 34.2 | 31.4 | 33.6 | y |
| Actinobacteria | Actinobacteria | Bifidobacteriales | 30.7 | 31.0 | | | **N** |
| Actinobacteria | Actinobacteria | unclassified | | | 29.7 | 29.7 | **N** |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Actinobacteria | Actinobacteria | Rubrobacterales | 29.6 | 29.6 | | | **N** |
| Bacteroidetes | Bacteroidia | Bacteroidales | 55.5 | 100.0 | 31.3 | 34.0 | y |
| Bacteroidetes | Cytophagia | Cytophagales | 35.5 | 38.1 | 33.8 | 33.8 | y |
| Bacteroidetes | Flavobacteriia | Flavobacteriales | 41.2 | 61.8 | 31.9 | 32.7 | y |
| Chloroflexi | Chloroflexi | Chloroflexales | 31.7 | 32.1 | 33.6 | 33.8 | y |
| Chloroflexi | Chloroflexi | Herpetosiphonales | 31.6 | 31.6 | 30.4 | 30.4 | y |
| Chloroflexi | Dehalococcoidetes | unclassified | 30.1 | 30.1 | 30.0 | 30.3 | **N** |
| Chloroflexi | Ktedonobacteria | Ktedonobacterales | 31.3 | 31.6 | 32.7 | 33.9 | y |
| Chloroflexi | Thermomicrobia | Sphaerobacterales | 30.3 | 30.3 | | | **N** |
| Cyanobacteria | Gloeobacteria | Gloeobacterales | | | 32.6 | 32.6 | **N** |
| Cyanobacteria | Cyanobacteria | Nostocales | 30.5 | 30.5 | 32.6 | 34.7 | y |
| Cyanobacteria | Cyanobacteria | Chroococcales | | | 31.5 | 33.3 | **N** |
| Cyanobacteria | Cyanobacteria | Oscillatoriales | | | 32.2 | 33.4 | **N** |
| Deinococcus-Thermus | Deinococci | Deinococcales | 31.4 | 31.4 | 31.2 | 32.4 | y |
| Deinococcus-Thermus | Deinococci | Thermales | | | 29.7 | 29.7 | **N** |
| Dictyoglomi | Dictyoglomia | Dictyoglomales | 29.9 | 29.9 | | | **N** |
| Fibrobacteres | Fibrobacteria | Fibrobacterales | | | 29.8 | 29.8 | **N** |
| Firmicutes | Bacilli | Lactobacillales | 32.7 | 38.1 | 40.9 | 55.0 | y |
| Firmicutes | Bacilli | Bacillales | 34.7 | 47.6 | 35.3 | 53.0 | y |
| Firmicutes | Clostridia | Clostridiales | 35.0 | 51.0 | 34.0 | 51.0 | y |
| Firmicutes | Clostridia | Halanaerobiales | 31.4 | 31.9 | 33.9 | 34.8 | y |
| Firmicutes | Clostridia | Thermoanaerobacterales | 32.6 | 35.7 | 35.0 | 39.2 | y |
| Firmicutes | Clostridia | Natranaerobiales | 30.9 | 30.9 | 33.3 | 33.3 | y |
| Firmicutes | Erysipelotrichi | Erysipelotrichales | 34.5 | 47.6 | 32.1 | 36.0 | y |
| Firmicutes | Negativicutes | Selenomonadales | 32.6 | 35.2 | 32.7 | 40.7 | y |
| Fusobacteria | Fusobacteriia | Fusobacteriales | 32.4 | 42.0 | 31.9 | 39.3 | y |
| | Spirochaetales | | 32.9 | 39.0 | 31.7 | 34.0 | |
| Synergistetes | Synergistia | Synergistales | | | 32.2 | 34.4 | **N** |
| Tenericutes | Mollicutes | Acholeplasmatales | 27.7 | 28.9 | 28.1 | 28.1 | y |
| Tenericutes | Mollicutes | Entomoplasmatales | 28.8 | 31.1 | 28.0 | 28.5 | y |
| Tenericutes | Mollicutes | Mycoplasmatales | 26.2 | 30.7 | 26.4 | 29.4 | y |
| Thermotogae | Thermotogae | Thermotogales | 30.7 | 31.3 | 31.0 | 32.3 | y |
| unclassified | Unclassified | Haloplasmatales | 39.9 | 44.6 | 31.9 | 33.7 | y |

|  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|
| unclassified | unclassified | unclassified bacteria | 30.0 | 30.0 | 33.0 | 33.0 | y |
| Verrucomicrobia | Opitutae | Opitutales |  |  | 29.7 | 29.7 | **N** |
| Proteobacteria | Alpha | Others | 24.0 | 31.0 | 50.0 | 33.0 | y |
| Proteobacteria | Alpha | Rhizobiaceae | 27.9 | 32.2 | 28.3 | 32.1 | y |
| Proteobacteria | Beta | Burkholderiaceae | 24.9 | 27.0 | 29.0 | 30.4 | y |
| Proteobacteria | Beta | Neisseriaceae | 25.2 | 28.0 | 26.7 | 30.0 | y |
| Proteobacteria | Delta |  | 25.0 | 29.0 | 25.0 | 27.0 | y |
| Proteobacteria | Epsilon |  | 27.0 | 27.0 | 26.0 | 27.0 | y |
| Proteobacteria | Gamma | Enterobacteriales | 30.9 | 35.9 | 71.9 | 99.6 | y |
| Proteobacteria | Gamma | Others | 27.3 | 36.9 | 30.5 | 47.0 | y |
| Proteobacteria | Gamma | Pasteurellaceae | 27.8 | 32.6 | 27.7 | 32.1 | y |
| Proteobacteria | Gamma | Pseudomonadaceae | 27.7 | 30.0 | 28.6 | 32.0 | y |
| Proteobacteria | Gamma | Vibrionaceae | 35.5 | 37.6 | 39.7 | 47.0 | y |
| Proteobacteria | Gamma | Xanthomonadaceae | 30.5 | 31.0 | 53.8 | 56.0 | y |

**Table 4.1:** Average and maximum sequence identities for BLAST searches of BT3352 and yidA. Cells are colored according to sequence identity range: no orthologs (red), 20-30% SI (orange), 30-40% SI (yellow), 40-50% SI (darker green), 50-100% SI (lighter green). The Ortholog Sharing column indicates whether the majority of species containing orthologs to both queries were shared orthologs (yes or no).

### 4.3.2: Biological ranges of putative FMN hydrolases

Biological ranges were determined for the putative FMN hydrolases BT2542, yigB, ybjI, yigL, and Q8SV5. In comparing all five against one another, a sequence identity threshold of 40% was used. For each species, we tracked whether one query's orthologs also were orthologous to one of the other four queries. This sort of top-hit ortholog sharing provides additional evidence of divergent evolution. Table 4.2 summarizes our results, in which we account for maximum ortholog sequence identity, average ortholog sequence identity, and percent of ortholog hits for a given query within a taxonomic group.

|  |  |  |  | BT2542 |  | yigB |  | ybjI |  | yigL |  | Q8SV5 |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Phylum | Class | Order | Family | Avg | % | Avg | % | Avg | % | Avg | % | Avg | % |  |
| Actinobacteria | Coriobacteridae | Coriobacteriales | Coriobacterineae |  |  |  |  | 40 | 100 |  |  |  |  |  |
| Bacteroidetes | Bacteroidia | Bacteroidales | Bacteroidaceae | 67 | 100 |  |  |  |  |  |  |  |  |  |
| Bacteroidetes | Bacteroidia | Bacteroidales | Porphyromonadaceae | 48 | 100 |  |  |  |  |  |  |  |  |  |
| Bacteroidetes | Bacteroidia | Bacteroidales | Prevotellaceae | 42 | 100 |  |  |  |  |  |  |  |  |  |
| Bacteroidetes | Bacteroidia | Bacteroidales | Rikenellaceae | 42 | 100 |  |  |  |  |  |  |  |  |  |
| Bacteroidetes | Cytophagia | Cytophagales | Cytophagaceae | 42 | 100 |  |  |  |  |  |  |  |  |  |
| Bacteroidetes | Flavobacteriia | Flavobacteriales | Flavobacteriaceae | 41 | 50 |  |  |  |  | 41 | 50 | 40 | 50 | ~ |
| Firmicutes | Bacilli | Bacillales | Bacillaceae |  |  |  |  | 49 | 100 |  |  |  |  |  |
| Firmicutes | Bacilli | Bacillales | Listeriaceae |  |  |  |  | 41 | 100 |  |  |  |  |  |
| Firmicutes | Bacilli | Bacillales | Staphylococcaceae |  |  |  |  | 43 | 100 |  |  |  |  |  |
| Firmicutes | Bacilli | Lactobacillales | Enterococcaceae |  |  |  |  | 41 | 100 |  |  |  |  |  |
| Firmicutes | Bacilli | Lactobacillales | Lactobacillaceae |  |  |  |  | 41 | 100 |  |  |  |  |  |
| Firmicutes | Bacilli | Lactobacillales | Streptococcaceae |  |  |  |  | 41 | 100 |  |  |  |  |  |
| Firmicutes | Clostridia | Clostridiales | Clostridiaceae |  |  |  |  | 40 | 100 |  |  |  |  |  |
| Firmicutes | Clostridia | Clostridiales | unclassified |  |  |  |  | 41 | 100 |  |  |  |  |  |
| Firmicutes | Erysipelotrichi | Erysipelotrichales | Erysipelotrichaceae |  |  |  |  | 47 | 100 |  |  |  |  |  |
| Fusobacteria | Fusobacteria | Fusobacterales | Fusobacteriaceae |  |  |  |  |  |  | 40 | 100 |  |  |  |
| Fusobacteria | Fusobacteriia | Fusobacterales | Leptotrichiaceae |  |  |  |  | 48 | 100 |  |  |  |  |  |
| Proteobacteria | Betaproteobacteria | Burkholderiales | Burkholderiaceae |  |  |  |  |  |  | 40 | 100 | 40 | 67 |  |
| Proteobacteria | Betaproteobacteria | Neisseriales | Neisseriaceae |  |  |  |  |  |  | 44 | 100 | 43 | 100 |  |
| Proteobacteria | Gammaproteobacteria | Aeromonadales | Aeromonadaceae |  |  | 40 | 67 |  |  | 46 | 83 | 46 | 83 |  |
| Proteobacteria | Gammaproteobacteria | Alteromonadales | Moritellaceae |  |  |  |  |  |  | 45 | 100 | 44 | 100 |  |
| Proteobacteria | Gammaproteobacteria | Alteromonadales | Psychromonadaceae |  |  |  |  |  |  | 46 | 100 | 45 | 100 |  |
| Proteobacteria | Gammaproteobacteria | Alteromonadales | Shewanellaceae |  |  | 40 | 100 |  |  |  |  |  |  |  |
| Proteobacteria | Gammaproteobacteria | Chromatiales | Chromatiaceae |  |  |  |  |  |  | 40 | 50 | 41 | 100 |  |
| Proteobacteria | Gammaproteobacteria | Pasteurellales | Pasteurellaceae |  |  | 41 | 5 |  |  | 42 | 93 | 42 | 93 |  |
| Proteobacteria | Gammaproteobacteria | Pseudomonadales | Moraxellaceae |  |  |  |  | 42 | 100 |  |  |  |  |  |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Proteobacteria | Gammaproteobacteria | Vibrionales | Vibrionaceae | | | 43 | 49 | | | 50 | 96 | 49 | 96 |
| Class | Order | Family | Genus | Avg | % | Avg | % | Avg | % | Avg | % | Avg | % |
| Gammaproteobacteria | Enterobacteriales | Enterobacteriaceae | Brenneria | | | 79 | 100 | | | 79 | 100 | 78 | 100 |
| Gammaproteobacteria | Enterobacteriales | Enterobacteriaceae | Buchnera | | | | | | | 44 | 100 | 43 | 100 |
| Gammaproteobacteria | Enterobacteriales | Enterobacteriaceae | Citrobacter | | | 87 | 100 | 82* | 100 | 91 | 100 | 93 | 100 |
| Gammaproteobacteria | Enterobacteriales | Enterobacteriaceae | Cronobacter | | | 74 | 100 | 66* | 100 | 86 | 67 | 86 | 67 |
| Gammaproteobacteria | Enterobacteriales | Enterobacteriaceae | Dickeya | | | 67 | 50 | | | 74 | 100 | 72 | 100 |
| Gammaproteobacteria | Enterobacteriales | Enterobacteriaceae | Edwardsiella | | | 61 | 67 | | | 70 | 100 | 68 | 100 |
| Gammaproteobacteria | Enterobacteriales | Enterobacteriaceae | Enterobacter | | | 81 | 100 | 73* | 100 | 89 | 100 | 87 | 100 |
| Gammaproteobacteria | Enterobacteriales | Enterobacteriaceae | Erwinia | | | 62 | 80 | 54 | 40 | 72 | 20 | 69 | 80 |
| Gammaproteobacteria | Enterobacteriales | Enterobacteriaceae | Escherichia | | | 99 | 26 | 81* | 84 | 99 | 35 | 89 | 35 |
| Gammaproteobacteria | Enterobacteriales | Enterobacteriaceae | Klebsiella | | | 75 | 50 | 69* | 88 | 87 | 38 | 82 | 38 |
| Gammaproteobacteria | Enterobacteriales | Enterobacteriaceae | Pantoea | | | 62 | 100 | 51* | 100 | 71 | 100 | 69 | 100 |
| Gammaproteobacteria | Enterobacteriales | Enterobacteriaceae | Pectobacterium | | | 67 | 50 | | | 78 | 83 | 77 | 83 |
| Gammaproteobacteria | Enterobacteriales | Enterobacteriaceae | Photorhabdus | | | 58 | 67 | | | 67 | 67 | 66 | 67 |
| Gammaproteobacteria | Enterobacteriales | Enterobacteriaceae | Proteus | | | 51 | 50 | | | 64 | 100 | 63 | 100 |
| Gammaproteobacteria | Enterobacteriales | Enterobacteriaceae | Providencia | | | 59 | 100 | | | 60 | 100 | 60 | 100 |
| Gammaproteobacteria | Enterobacteriales | Enterobacteriaceae | Rahnella | | | 59 | 100 | 50 | 100 | 75 | 100 | 72 | 100 |
| Gammaproteobacteria | Enterobacteriales | Enterobacteriaceae | Salmonella | | | 87 | 26 | 77* | 89 | 90 | 26 | 99 | 26 |
| Gammaproteobacteria | Enterobacteriales | Enterobacteriaceae | Serratia | | | 65 | 100 | 99 | 17 | 76* | 83 | 75* | 83 |
| Gammaproteobacteria | Enterobacteriales | Enterobacteriaceae | Shigella | | | 100 | 36 | 88* | 93 | 99 | 57 | 89 | 57 |
| Gammaproteobacteria | Enterobacteriales | Enterobacteriaceae | Sodalis | | | 65 | 100 | | | 68 | 100 | 66 | 100 |
| Gammaproteobacteria | Enterobacteriales | Enterobacteriaceae | unclassified | | | 48 | 100 | | | 55 | 100 | 54 | 100 |
| Gammaproteobacteria | Enterobacteriales | Enterobacteriaceae | Xenorhabdus | | | 60 | 100 | | | 65 | 100 | 64 | 100 |
| Gammaproteobacteria | Enterobacteriales | Enterobacteriaceae | Yersinia | | | 64 | 20 | 54 | 47 | 74 | 93 | 74 | 93 |

**Table 4.2:** Biological ranges of putative FMN hydrolases arranged according to taxonomy. Most results are averaged for the family of interest but because four of the query sequences belonged to the same family (Enterobacteriaceae), orthologs for the overarching familywere averaged for the

genus of interest. Average sequence identity (Avg) for each group is reported; the cell is colored according to the maximum percent identity of all orthologs within the taxonomic group (green = 80-100%, yellow = 50-79%, red = 40-50%). Number of orthologs (%) for each query is reported as percent of all species within that taxonomic group containing an ortholog for any query. An asterisk indicates that the taxonomic group contained species with more than one ortholog for the query; these additional orthologs (all of which had sequence identities <50%) were excluded from the average sequence identity calculation. A tilde indicates that the reported query orthologs for the queries in question were not present in the same species.

In every taxonomic group containing orthologs to both yigL and Q8SV5E, the majority of orthologs were shared by both queries. While both queries had orthologs in species lacking an ortholog to the other, every species containing both returned the same ortholog for both. Indeed, the shared ortholog also had very similar sequence identity for both queries with an average sequence identity difference of 2.5%. The greatest difference in shared ortholog sequence identity was 7-10% and occurred only in the opposite query genus. In other words, in *Salmonella* (the genus to which Q8SV5 belongs), a shared ortholog might have 90% sequence identity to yigL compared to 100% sequence identity to Q8SV5. Conversely, in *Escherichia* (the genus to which yigL belongs), a shared ortholog might have 90% sequence identity to Q8SV5 compared to 100% sequence identity to yigL.

### *4.3.2: Gene contexts of putative FMN hydrolases*

Gene contexts were determined for the five putative FMN hydrolases but only minimal shared context was discovered. Each individual query exhibited some degree of conserved gene context but none of them shared gene context with one another, except Q8SV5 and yigL. Indeed, given that Q8SV5 and yigL nearly always returned the same orthologous protein, they shared exactly the same gene context and are treated as one when discussing gene context below.

In BT2542, 79.5% of orthologs were found to have an adjacent protein annotated as ribF and/or having better than 40% SI to the BT2543 ribF. Given the chemical relationship between ribF and FMN hydrolase (ribF produces FMN while FMN hydrolase catalyzes the reverse reaction), the proximity of ribF supports the assignment of BT2542 as an FMN hydrolase.

Of the 68 yigB orthologs with interrogated gene contexts, all were found adjacent to diaminopimelate epimerase, 14.7% were found in the neighborhood of cya-Y frataxin-like proteins, 89.7% were found near xerC site specific tyrosine recombinases, and 70.5% were found in the neighborhood of DNA-dependent helicase II. The ybjI orthologs had significantly less conserved and minimally useful gene context: 75% of the orthologs with >80% sequence identities were found adjacent to another HADSF protein, particularly of the Cof-like hydrolase family (45.5% of all orthologs).  No definitive context could be determined for 34% of the ybjI orthologs and the rest exhibited no conservation of context.

Orthologs to yigL and Q8SV5 were shared between the two but not consistent across the range of available gene contexts.  The neighborhood contained: lysophospholipase (65%), ATP-dependent helicase (43%), homoserine lactone efflux protein (42%), threonine efflux system or pump (38%), 5-methyltetrahydropteroyltriglutamate/homocysteine S-methyltransferase (35%) and various regulators (10% for LysR, 14% for metE and metH regulators).

### *4.3.3: Divergence from anomalous DxG motif in BT2542*

In order to further probe the relationship between BT2542 and yigB, the biological range constraints were relaxed to 20% sequence identity in order to capture any possible links between the two queries.  Even given these much more lenient constraints, BT2542 and yigB only rarely appeared in the same order, and only then at very low sequence identities for both sequences (Table 4.3). They shared an ortholog in only some of these cases, with BT2542 having higher sequence identity to the shared ortholog and yigB having very low sequence identity. This further underscores that the two proteins share a superfamily and fold but clearly no recent evolutionary relationship.

| Phylum | Class | Order | yigB avg SI | yigB max SI | BT2542 avg SI | BT2542 max SI | Same ortholog | BT2542 divergence from D+GGVL motif |
|---|---|---|---|---|---|---|---|---|
| Acidobacteria | Acidobacteriia | Acidobacteriales | | | 26 | 26 | N | 1 |
| Acidobacteria | Solibacteres | Solibacterales | | | 29 | 29 | N | 0.85 |
| Actinobacteria | Actinobacteria | Actinomycetales | 27.3 | 29 | 24.9 | 28 | N | 0.9 |
| Bacteroidetes | Bacteroidia | Bacteroidales | | | 45.2 | 100 | N | 0.75 |
| Bacteroidetes | Cytophagia | Cytophagales | | | 35.2 | 42 | N | 0.55 |
| Bacteroidetes | Flavobacteriia | Flavobacteriales | 24.5 | 25 | 34.5 | 41 | **Y** | 0.4 |
| Bacteroidetes | Sphingobacteriia | Sphingobacteriales | | | 37.6 | 39 | N | 0.5 |
| Chlorobi | Chlorobia | Chlorobiales | | | 25.8 | 28 | N | 0.5 |
| Elusimicrobia | Elusimicrobia | Elusimicrobiales | | | 27 | 27 | N | 0.85 |
| Fibrobacteres | Fibrobacteria | Fibrobacterales | | | 34 | 34 | N | 1 |
| Firmicutes | Bacilli | Lactobacillales | 26.5 | 28 | 27.3 | 25 | N | 0.75 |
| Firmicutes | Clostridia | Clostridiales | 25 | 27 | 26 | 31 | **Y** | 0.75 |
| Firmicutes | Clostridia | Halanaerobiales | | | 31 | 31 | N | 0.8 |
| Fusobacteria | Fusobacteriia | Fusobacteriales | 20 | 20 | 29.1 | 32 | **Y** | 0.85 |
| Planctomycetes | Planctomycetia | Planctomycetales | | | 27.7 | 32 | N | 0.6 |
| Thermotogae | Thermotogae | Thermotogales | 23 | 23 | 27.2 | 30 | **Y** | 0.9 |
| Verrucomicrobia | Spartobacteria | | | | 30 | 30 | N | 0.85 |
| Verrucomicrobia | Verrucomicrobiae | Verrucomicrobiales | | | 28 | 28 | N | 0.85 |

**Table 4.3**: Biological range overlap, including average and maximum sequence of orthologs computed for each taxonomic group, for BT2542 and yigB. Ortholog sharing is also noted (yes or no) as is divergence of BT2542 orthologs from the anomalous active site motif. For divergence from the BT2542 motif, 1 indicates the motif is completely conserved and <1 indicates the percentage of the motif conserved, taken as an average for each taxonomic group.

Though previous studies demonstrated that the anomalous DxG motif is indeed functional (*9*) the question of how prevalent it is remains. The BT2542 motif can be represented as D+GGVL. Departures from this motif can be given a score based on how many positions they share with the motif (1 minus 1/6th for each departure from the motif). Scores were averaged for each taxonomic group and recorded according to biological range (Table 4.3) where we see that the new motif largely persists across the BT2542 orthologs. The common deviations, described in Figure 4.4, indicate that even though the canonical HADSF motif is Dx***D***, the D+**G**GVL motif is strictly conserved. Of ~450 orthologs with >20% sequence identity, only two contained a deviation from the aberrant Gly, switching it "back" to Asp; these two cases occurred in orthologs with 20% and 21% sequence identity. The first D is absolutely conserved, which is consistent with its important role in catalysis but the absolute conservation of the non-functional glycine is puzzling, given that it replaces the second Asp, a general acid/base.

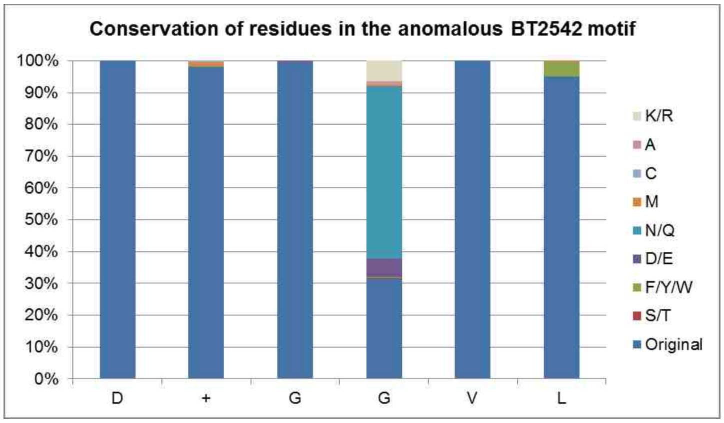


**Figure 4.4:** Conservation BT242 orthologs to the anomalous D+GGVL motif. Conservation is reported as percentage of the six residues maintained compared to the original motif. Departures from the motif are categorized according to amino acid type.

## 4.4: Conclusions

I have demonstrated that yigB and BT2542 share no significant biological range, even at very lenient cutoffs. It is clear from the persistence of the anomalous DxG catalytic motif, lack of query sequence identity, and very minimal ortholog sharing and shared ortholog sequence identity that these two proteins are an example of convergent evolution within the HADSF to acquire FMN hydrolase activity. This FMN hydrolase activity is further supported in BT2542 by the conserved adjacent riboflavin kinase protein.

YigL and Q83SV5 are clearly closely related orthologs separated almost entirely by speciation—they share high sequence identities (89.5%), similar identities to shared orthologs, and, due to the latter, shared conservation of gene contexts, as well. The relationship among yigB, yigL, and ybjI is less clear; by virtue of being from the same species, they share some degree of biological range. However, that yigB is a Cap 1 type HAD and yigL/ybjI are Cap 2 type HADs suggests different evolutionary history and convergent evolution at the cap divide level. The sole extension of ybjI into Firmicutes is curious and should be further explored; it may indicate gene transfer from Proteobacteria into Firmicutes.

The shared biological range suggests that BT3352 and yidA are related by divergent evolution. Both have orthologs with unexpectedly high sequence identities in the phylum Firmicutes compared to other non-native (e.g., Bacteroidetes/ Proteobacteria) taxonomic groups suggesting that there may have been gene transfer among the three phyla. Gene context is not sufficiently conserved to support or infer physiological activity.

### 4.5: Referencess

1. 1. Powers, H. J., Riboflavin (vitamin B-2) and health. *The American journal of clinical nutrition* **2003,** *77*, 1352-60.
2. Efimov, I.; Kuusk, V.; Zhang, X.; McIntire, W. S., Proposed steady-state kinetic mechanism for Corynebacterium ammoniagenes FAD synthetase produced by Escherichia coli. *Biochemistry* **1998,** *37*, 9716-23.
3. Mack, M.; van Loon, A. P.; Hohmann, H. P., Regulation of riboflavin biosynthesis in Bacillus subtilis is affected by the activity of the flavokinase/flavin adenine dinucleotide synthetase encoded by ribC. *J Bacteriol* **1998,** *180*, 950-5.
4. Manstein, D. J.; Pai, E. F., Purification and characterization of FAD synthetase from Brevibacterium ammoniagenes. *J Biol Chem* **1986,** *261*, 16169-73.
5. Barile, M.; Brizio, C.; De Virgilio, C.; Delfine, S.; Quagliariello, E.; Passarella, S., Flavin adenine dinucleotide and flavin mononucleotide metabolism in rat liver--the occurrence of FAD pyrophosphatase and FMN phosphohydrolase in isolated mitochondria. *European journal of biochemistry / FEBS* **1997,** *249*, 777-85.
6. Sandoval, F. J.; Roje, S., An FMN hydrolase is fused to a riboflavin kinase homolog in plants. *J Biol Chem* **2005,** *280*, 38337-45.

7. Haase, I.; Sarge, S.; Illarionov, B.; Laudert, D.; Hohmann, H. P.; Bacher, A.; Fischer, M., Enzymes from the haloacid dehalogenase (HAD) superfamily catalyse the elusive dephosphorylation step of riboflavin biosynthesis. *Chembiochem : a European journal of chemical biology* **2013,** *14*, 2272-5.

8. Huang, H.; Pandya, C.; Liu, C.; Al-Obaidi, N. F.; Wang, M.; Zheng, L.; Toews Keating, S.; Aono, M.; Love, J. D.; Evans, B.; Seidel, R. D.; Hillerich, B. S.; Garforth, S. J.; Almo, S. C.; Mariano, P. S.; Dunaway-Mariano, D.; Allen, K. N.; Farelli, J. D., Panoramic view of a superfamily of phosphatases through substrate profiling. *Proc Natl Acad Sci U S A* **2015**.

9. Zheng, L. Catalytic mechanism and function evolvement studies of phosphatases within Haloacid Dehalogenase Superfamily (HADSF). http://hdl.handle.net/1928/23164, University of New Mexico, http://hdl.handle.net/1928/23164, 2013.

10. Robert, X.; Gouet, P., Deciphering key features in protein structures with the new ENDscript server. *Nucleic acids research* **2014,** *42*, W320-4.

11. Wang, M. Evolution of structure and function among hotdog-fold thioesterases and HAD family phosphatases. http://hdl.handle.net/1928/17513, University of New Mexico, 2011.

12. Cooper, R. A., The utilisation of D-galactonate and D-2-oxo-3-deoxygalactonate by Escherichia coli K-12. Biochemical and genetical studies. *Archives of microbiology* **1978**, *118*, 199-206.

13. Peekhaus, N.; Conway, T., What's for dinner?: Entner-Doudoroff metabolism in Escherichia coli. *J Bacteriol* **1998**, *180*, 3495-502.

14. Qin, J.; Li, R.; Raes, J.; Arumugam, M.; Burgdorf, K. S.; Manichanh, C.; Nielsen, T.; Pons, N.; Levenez, F.; Yamada, T.; Mende, D. R.; Li, J.; Xu, J.; Li, S.; Li, D.; Cao, J.; Wang, B.; Liang, H.; Zheng, H.; Xie, Y.; Tap, J.; Lepage, P.; Bertalan, M.; Batto, J. M.; Hansen, T.; Le Paslier, D.; Linneberg, A.; Nielsen, H. B.; Pelletier, E.; Renault, P.; Sicheritz-Ponten, T.; Turner, K.; Zhu, H.; Yu, C.; Li, S.; Jian, M.; Zhou, Y.; Li, Y.; Zhang, X.; Li, S.; Qin, N.; Yang, H.; Wang, J.; Brunak, S.; Dore, J.; Guarner, F.; Kristiansen, K.; Pedersen, O.; Parkhill, J.; Weissenbach, J.; Meta, H. I. T. C.; Bork, P.; Ehrlich, S. D.; Wang, J., A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **2010,** *464*, 59-65.

15. Bik, E. M., Composition and function of the human-associated microbiota. *Nutrition reviews* **2009,** *67 Suppl 2*, S164-71.

16. Eckburg, P. B.; Bik, E. M.; Bernstein, C. N.; Purdom, E.; Dethlefsen, L.; Sargent, M.; Gill, S. R.; Nelson, K. E.; Relman, D. A., Diversity of the human intestinal microbial flora. *Science* **2005,** *308*, 1635-8.

17. UniProt, C., UniProt: a hub for protein information. *Nucleic acids research* **2015,** *43*, D204-12.

18. Papadopoulos, J. S.; Agarwala, R., COBALT: constraint-based alignment tool for multiple protein sequences. *Bioinformatics* **2007,** *23*, 1073-9.

19. Letunic, I.; Bork, P., Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic acids research* **2011,** *39*, W475-8.

20. Szklarczyk, D.; Franceschini, A.; Kuhn, M.; Simonovic, M.; Roth, A.; Minguez, P.; Doerks, T.; Stark, M.; Muller, J.; Bork, P.; Jensen, L. J.; von Mering, C., The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research* **2011,** *39*, D561-8.

# APPENDIX

## A.1: Supplementary data

### *A.1.1: Functional annotations of hotdog-family members based on literature*

Literature search was conducted by Jie (Jenny) Zhang, former member of the Dunaway-Mariano lab.

| **UniProtID** | **General function type** | **Organism** | **Protein name** | **Source** |
|---|---|---|---|---|
| **Q9NPJ3** | Medium to long chain acyl-CoA | *Homo sapiens* | ACOT13 (THEM2) | (*1*) |
| **P77781** | Other | *Escherichia coli* | YdiI | (*2*) |
| **P76084** | Other | *Escherichia coli* | PaaI | (*3*) |
| **P56653** | Other | *Pseudomonas sp.* | 4HBT | (*4*) |
| **O34835** | Other | *Bacillus subtilis* | FapR | (*5*) |
| **P58137** | Long chain acyl-CoA | *Mus musculus* | Acot 8 | (*6*) |
| **Q8WYK0** | Broad range branched acyl-coA | *Homo sapiens* | ACOT 8 | (*7*) |
| **Q9Y305** | Long chain acyl-CoA | *Homo sapiens* | ACOT 9 | (*8*) |
| **Q9CQJ0** | Long chain acyl-CoA | *Mus musculus* | Acot15 (Them5) | (*9*) |
| **P14604** | Broad range straight chain acyl-CoA | *Rattus norvegicus* | Echs1 | (*10*) |
| **Q9R0X4** | Broad Range Acyl-CoA | *Mus musculus* | Acot 9 | (*8*) |
| **Q9CQR4** | Medium to long chain acyl-CoA | *Mus musculus* | Acot 13 (Them2) | (*11*) |
| **O00154** | Medium to long chain acyl-CoA | *Homo sapiens* | ACOT 7 (BACH) | (*8*) |
| **Q8WYK0** | Short chain acyl-CoA | *Homo sapiens* | ACOT12 (CACH) | (*12*) |
| **Q9DBK0** | Short chain acyl-CoA | *Mus musculus* | Acot 12 | (*13*) |
| **Q99NB7** | Short chain acyl-CoA | *Rattus norvegicus* | | (*14*) |
| **Q91V12** | Medium to long chain acyl-CoA | *Mus musculus* | Acot 7 | (*15*) |
| **Q8WXI4** | Medium to long chain acyl-CoA | *Homo sapiens* | ACOT11 (BFIT, Them1) | (*8*) |
| **Q9KBC9** | Other | *Bacillus halodurans* | BH1999 | (*16*) |
| **A6L315** | DHNA CoA | *Bacteroides vulgatus* | DHN-CoA | (*17*) |
| **P0A8Z3** | Short chain acyl-CoA | *Escherichia coli* | YbgC | (*18*) |
| **T2BL43** | Short chain acyl-CoA | *Haemophilus influenzae* | YbgC | (*19*) |
| **P77455** | Other | *Escherichia coli* | Paaz (MaoC) | (*20*) |

|  |  |  |  |  |
|---|---|---|---|---|
| **P77712** | Other | *Escherichia coli* | FadM (tesC, ybaW) | (*21*) |
| **Q9HTY7** | Other | *Pseudomonas aeruginosa* | PA5202 | (*22*) |
| **A9CFF2** | Other | *Agrobacterium tumefaciens strain C58* | hbdA | (*23*) |
| **Q84HI6** | Other | *Azoarcus evansii* | benzoyl-CoA thioesterase | (*24*) |
| **P0A8Y8** | Other | *Escherichia coli* | EntH (YbdB) | (*18*) |
| **Q6LS54** | Other | *Photobacterium profundum* | eicosapentaenoic acid synthesis gene cluster | (*25*) |
| **P96807** | Other | *Mycobacterium tuberculosis* | enoly-CoA hydratase | (*26*) |
| **P0A6Q3** | Other | *Escherichia coli* | fabA | (*27*) |
| **P0A6Q6** | Other | *Escherichia coli* | fabZ | (*27*) |
| **Q04416** | Other | *Arthrobacter sp.* | 4HBT-II (fcbC) | (*28*) |
| **Q93CG9** | Other | *Photobacterium profundum* | Orf6 | (*25*) |
| **Q9I042** | Other | *Pseudomonas aeruginosa* | PA2801 | (*22*) |
| **Q0R4E3** | Other | *Campylobacter jejuni* | Virulence protein | (*29*) |
| **O25174** | Other | *Heliobacter pylori* | Regulatory protein | (*30*) |
| **P0AEK4** | Other | *Escherichia coli* | fabI | (*31*) |
| **P0ADP2** | Other | *Escherichia coli* | YigI | (*32*) |
| **P0ADQ2** | Other | *Escherichia coli* | YiiD | (*33*) |
| **Q42561** | Medium to long chain acyl-CoA | *Arabidopsis thaliana* | FATA | (*34*) |
| **P0AGG2** | Medium chain acyl-CoA | *Escherichia coli* | tesB (TEII) | (*35*) |
| **Q41635** | Long chain acyl-CoA | *Umberllularia californica* | FATBI | (*36*) |
| **Q9SQI3** | Long chain acyl-CoA | *Gossypium hirsutum* |  | (*37*) |
| **P64685** | Long chain acyl-CoA | *M. tuberculosis* | RV0098 | (*38*) |
| **P0A8Z0** | Long chain acyl-CoA | *Escherichia coli* | YciA | (*39*) |
| **J0S389** | Long chain acyl-CoA | *Helicobacter pylori* |  | (*40*) |
| **Q1EMV2** | FLK | *Streptomyces cattleya* | FLK | (*41*) |
| **Q55777** | DHNA CoA | *Synechocystis sp.* | Slr0204 | (*42*) |
| **Q89YN2** | DHNA CoA | *Bacteroides thetaiotaomicron* | BF1314 | (*17*) |
| **Q7MU91** | DHNA CoA | *Porphyromonas gingivalis* | PG1653 | (*17*) |
| **Q8D151** | Broad Range Acyl-CoA | *Yersinia pestis* | TesB | (*43*) |
| **Q0P9Y4** | Broad Range Acyl-CoA | *Campylobacter jejuni* | Cj0915 | (*44*) |

| | | | | |
|---|---|---|---|---|
| **Q5T1C6** | Broad Range Acyl-CoA | *Homo sapiens* | THM4 (CTMP) | (*45*) |

**Table A.1:** Known hotdog-family functions from a literature search conducted in February, 2014.

## References

1. Cao, J.; Xu, H.; Zhao, H.; Gong, W.; Dunaway-Mariano, D., The mechanisms of human hotdog-fold thioesterase 2 (hTHEM2) substrate recognition and catalysis illuminated by a structure and function based analysis. *Biochemistry* **2009,** *48*, 1293-304.
2. Latham, J. A.; Chen, D.; Allen, K. N.; Dunaway-Mariano, D., Divergence of substrate specificity and function in the Escherichia coli hotdog-fold thioesterase paralogs YdiI and YbdB. *Biochemistry* **2014,** *53*, 4775-87.
3. Song, F.; Zhuang, Z.; Finci, L.; Dunaway-Mariano, D.; Kniewel, R.; Buglino, J. A.; Solorzano, V.; Wu, J.; Lima, C. D., Structure, function, and mechanism of the phenylacetate pathway hot dog-fold thioesterase PaaI. *J Biol Chem* **2006,** *281*, 11028-38.
4. Song, F.; Zhuang, Z.; Dunaway-Mariano, D., Structure-activity analysis of base and enzyme-catalyzed 4-hydroxybenzoyl coenzyme A hydrolysis. *Bioorganic chemistry* **2007,** *35*, 1-10.
5. Schujman, G. E.; Paoletti, L.; Grossman, A. D.; de Mendoza, D., FapR, a bacterial transcription factor involved in global regulation of membrane lipid biosynthesis. *Developmental cell* **2003,** *4*, 663-72.

6. Hunt, M. C.; Solaas, K.; Kase, B. F.; Alexson, S. E., Characterization of an acyl-coA thioesterase that functions as a major regulator of peroxisomal lipid metabolism. *J Biol Chem* **2002,** *277*, 1128-38.

7. Hunt, M. C.; Alexson, S. E., Novel functions of acyl-CoA thioesterases and acyltransferases as auxiliary enzymes in peroxisomal lipid metabolism. *Progress in lipid research* **2008,** *47*, 405-21.

8. Kirkby, B.; Roman, N.; Kobe, B.; Kellie, S.; Forwood, J. K., Functional and structural properties of mammalian acyl-coenzyme A thioesterases. *Progress in lipid research* **2010,** *49*, 366-77.

9. Zhuravleva, E.; Gut, H.; Hynx, D.; Marcellin, D.; Bleck, C. K.; Genoud, C.; Cron, P.; Keusch, J. J.; Dummler, B.; Esposti, M. D.; Hemmings, B. A., Acyl coenzyme A thioesterase Them5/Acot15 is involved in cardiolipin remodeling and fatty liver development. *Molecular and cellular biology* **2012,** *32*, 2685-97.

10. Muller-Newen, G.; Janssen, U.; Stoffel, W., Enoyl-CoA hydratase and isomerase form a superfamily with a common active-site glutamate residue. *European journal of biochemistry / FEBS* **1995,** *228*, 68-73.

11. Kang, H. W.; Niepel, M. W.; Han, S.; Kawano, Y.; Cohen, D. E., Thioesterase superfamily member 2/acyl-CoA thioesterase 13 (Them2/Acot13) regulates hepatic lipid and glucose metabolism. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* **2012,** *26*, 2209-21.

12. Suematsu, N.; Isohashi, F., Molecular cloning and functional expression of human cytosolic acetyl-CoA hydrolase. *Acta biochimica Polonica* **2006,** *53*, 553-61.

13. Suematsu, N.; Okamoto, K.; Isohashi, F., Mouse cytosolic acetyl-CoA hydrolase, a novel candidate for a key enzyme involved in fat metabolism: cDNA cloning, sequencing and functional expression. *Acta biochimica Polonica* **2002,** *49*, 937-45.

14. Suematsu, N.; Okamoto, K.; Shibata, K.; Nakanishi, Y.; Isohashi, F., Molecular cloning and functional expression of rat liver cytosolic acetyl-CoA hydrolase. *European journal of biochemistry / FEBS* **2001,** *268*, 2700-9.

15. Cantu, D. C.; Chen, Y.; Reilly, P. J., Thioesterases: a new perspective based on their primary and tertiary structures. *Protein Sci* **2010,** *19*, 1281-95.

16. Zhuang, Z.; Song, F.; Takami, H.; Dunaway-Mariano, D., The BH1999 protein of Bacillus halodurans C-125 is gentisyl-coenzyme A thioesterase. *J Bacteriol* **2004,** *186*, 393-9.

17. Wang, M. Evolution of structure and function among hotdog-fold thioesterases and HAD family phosphatases. http://hdl.handle.net/1928/17513, University of New Mexico, 2011.

18. Zhuang, Z.; Song, F.; Zhao, H.; Li, L.; Cao, J.; Eisenstein, E.; Herzberg, O.; Dunaway-Mariano, D., Divergence of function in the hot dog fold

enzyme superfamily: the bacterial thioesterase YciA. *Biochemistry* **2008,** *47*, 2789-96.

19. Zhuang, Z.; Song, F.; Martin, B. M.; Dunaway-Mariano, D., The YbgC protein encoded by the ybgC gene of the tol-pal gene cluster of Haemophilus influenzae catalyzes acyl-coenzyme A thioester hydrolysis. *FEBS letters* **2002,** *516*, 161-3.

20. Teufel, R.; Mascaraque, V.; Ismail, W.; Voss, M.; Perera, J.; Eisenreich, W.; Haehnel, W.; Fuchs, G., Bacterial phenylalanine and phenylacetate catabolic pathway revealed. *Proc Natl Acad Sci U S A* **2010,** *107*, 14390-5.

21. Nie, L.; Ren, Y.; Schulz, H., Identification and characterization of Escherichia coli thioesterase III that functions in fatty acid beta-oxidation. *Biochemistry* **2008,** *47*, 7744-51.

22. Gonzalez, C. F.; Tchigvintsev, A.; Brown, G.; Flick, R.; Evdokimova, E.; Xu, X.; Osipiuk, J.; Cuff, M. E.; Lynch, S.; Joachimiak, A.; Savchenko, A.; Yakunin, A. F., Structure and activity of the Pseudomonas aeruginosa hotdog-fold thioesterases PA5202 and PA2801. *Biochem J* **2012,** *444*, 445-55.

23. Dillon, S. C.; Bateman, A., The Hotdog fold: wrapping up a superfamily of thioesterases and dehydratases. *BMC bioinformatics* **2004,** *5*, 109.

24. Ismail, W., Benzoyl-coenzyme A thioesterase of Azoarcus evansii: properties and function. *Archives of microbiology* **2008,** *190*, 451-60.

25. Rodriguez-Guilbe, M.; Oyola-Robles, D.; Schreiter, E. R.; Baerga-Ortiz, A., Structure, activity, and substrate selectivity of the Orf6 thioesterase from Photobacterium profundum. *J Biol Chem* **2013,** *288*, 10841-8.

26. Johansson, P.; Castell, A.; Jones, T. A.; Backbro, K., Structure and function of Rv0130, a conserved hypothetical protein from Mycobacterium tuberculosis. *Protein Sci* **2006,** *15*, 2300-9.

27. Heath, R. J.; Rock, C. O., Roles of the FabA and FabZ beta-hydroxyacyl-acyl carrier protein dehydratases in Escherichia coli fatty acid biosynthesis. *J Biol Chem* **1996,** *271*, 27795-801.

28. Zhuang, Z.; Gartemann, K. H.; Eichenlaub, R.; Dunaway-Mariano, D., Characterization of the 4-hydroxybenzoyl-coenzyme A thioesterase from Arthrobacter sp. strain SU. *Applied and environmental microbiology* **2003,** *69*, 2707-11.

29. Yokoyama, T.; Paek, S.; Ewing, C. P.; Guerry, P.; Yeo, H. J., Structure of a sigma28-regulated nonflagellar virulence protein from Campylobacter jejuni. *J Mol Biol* **2008,** *384*, 364-76.

30. Tomb, J. F.; White, O.; Kerlavage, A. R.; Clayton, R. A.; Sutton, G. G.; Fleischmann, R. D.; Ketchum, K. A.; Klenk, H. P.; Gill, S.; Dougherty, B. A.; Nelson, K.; Quackenbush, J.; Zhou, L.; Kirkness, E. F.; Peterson, S.; Loftus, B.; Richardson, D.; Dodson, R.; Khalak, H. G.; Glodek, A.; McKenney, K.; Fitzegerald, L. M.; Lee, N.; Adams, M. D.; Hickey, E. K.; Berg, D. E.; Gocayne, J. D.; Utterback, T. R.; Peterson, J. D.; Kelley, J. M.; Cotton, M. D.; Weidman, J. M.; Fujii, C.; Bowman, C.; Watthey, L.;

Wallin, E.; Hayes, W. S.; Borodovsky, M.; Karp, P. D.; Smith, H. O.; Fraser, C. M.; Venter, J. C., The complete genome sequence of the gastric pathogen Helicobacter pylori. *Nature* **1997,** *388*, 539-47.

31. Bergler, H.; Wallner, P.; Ebeling, A.; Leitinger, B.; Fuchsbichler, S.; Aschauer, H.; Kollenz, G.; Hogenauer, G.; Turnowsky, F., Protein EnvM is the NADH-dependent enoyl-ACP reductase (FabI) of Escherichia coli. *J Biol Chem* **1994,** *269*, 5493-6.

32. Wu, Y. DETERMINING THE FUNCTION OF HOTDOG-FOLD THIOESTERASES. Masters thesis, University of New Mexico, Department of Chemistry and Chemical Biology, 2013.

33. Hayashi, K.; Morooka, N.; Yamamoto, Y.; Fujita, K.; Isono, K.; Choi, S.; Ohtsubo, E.; Baba, T.; Wanner, B. L.; Mori, H.; Horiuchi, T., Highly accurate genome sequences of Escherichia coli K-12 strains MG1655 and W3110. *Mol Syst Biol* **2006,** *2*, 2006 0007.

34. Salas, J. J.; Ohlrogge, J. B., Characterization of substrate specificity of plant FatA and FatB acyl-ACP thioesterases. *Arch Biochem Biophys* **2002,** *403*, 25-34.

35. Nie, L.; Ren, Y.; Janakiraman, A.; Smith, S.; Schulz, H., A novel paradigm of fatty acid beta-oxidation exemplified by the thioesterase-dependent partial degradation of conjugated linoleic acid that fully supports growth of Escherichia coli. *Biochemistry* **2008,** *47*, 9618-26.

36. Voelker, T. A.; Worrell, A. C.; Anderson, L.; Bleibaum, J.; Fan, C.; Hawkins, D. J.; Radke, S. E.; Davies, H. M., Fatty acid biosynthesis

redirected to medium chains in transgenic oilseed plants. *Science* **1992,** *257*, 72-4.

37. Pirtle, R. M.; Yoder, D. W.; Huynh, T. T.; Nampaisansuk, M.; Pirtle, I. L.; Chapman, K. D., Characterization of a palmitoyl-acyl carrier protein thioesterase (FatB1) in cotton. *Plant & cell physiology* **1999,** *40*, 155-63.

38. Wang, F.; Langley, R.; Gulten, G.; Wang, L.; Sacchettini, J. C., Identification of a type III thioesterase reveals the function of an operon crucial for Mtb virulence. *Chem Biol* **2007,** *14*, 543-51.

39. Kuznetsova, E.; Proudfoot, M.; Sanders, S. A.; Reinking, J.; Savchenko, A.; Arrowsmith, C. H.; Edwards, A. M.; Yakunin, A. F., Enzyme genomics: Application of general enzymatic screens to discover new enzymes. *FEMS microbiology reviews* **2005,** *29*, 263-79.

40. Behrens, W.; Bonig, T.; Suerbaum, S.; Josenhans, C., Genome sequence of Helicobacter pylori hpEurope strain N6. *J Bacteriol* **2012,** *194*, 3725-6.

41. Huang, F.; Haydock, S. F.; Spiteller, D.; Mironenko, T.; Li, T. L.; O'Hagan, D.; Leadlay, P. F.; Spencer, J. B., The gene cluster for fluorometabolite biosynthesis in Streptomyces cattleya: a thioesterase confers resistance to fluoroacetyl-coenzyme A. *Chem Biol* **2006,** *13*, 475-84.

42. Widhalm, J. R.; van Oostende, C.; Furt, F.; Basset, G. J., A dedicated thioesterase of the Hotdog-fold family is required for the biosynthesis of the naphthoquinone ring of vitamin K1. *Proc Natl Acad Sci U S A* **2009,** *106*, 5599-603.

43. Swarbrick, C. M.; Patterson, E. I.; Forwood, J. K., Crystallization of the acyl-CoA thioesterase TesB from Yersinia pestis. *Acta crystallographica. Section F, Structural biology and crystallization communications* **2013,** *69*, 188-90.

44. Yokoyama, T.; Choi, K. J.; Bosch, A. M.; Yeo, H. J., Structure and function of a Campylobacter jejuni thioesterase Cj0915, a hexameric hot dog fold enzyme. *Biochim Biophys Acta* **2009,** *1794*, 1073-81.

45. Zhao, H.; Lim, K.; Choudry, A.; Latham, J. A.; Pathak, M. C.; Dominguez, D.; Luo, L.; Herzberg, O.; Dunaway-Mariano, D., Correlation of structure and function in the human hotdog-fold enzyme hTHEM4. *Biochemistry* **2012,** *51*, 6490-2.

### *A.1.2: Function, subfamily, and/or structure annotations to hotdog-fold family SSN clusters*

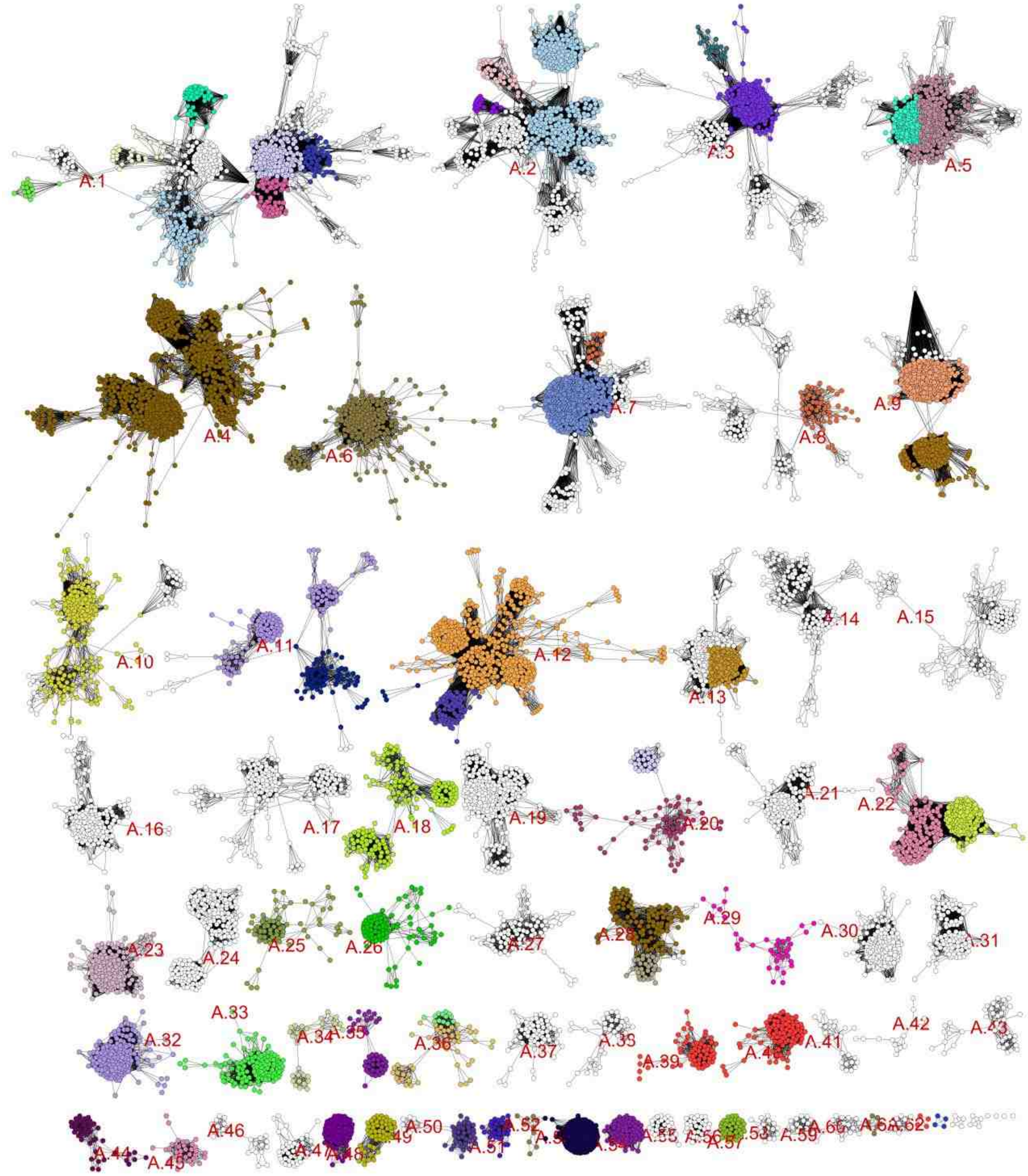


**Figure A.1:** Hotdog-fold family sequence similarity network, colored according to approximate regions that have been annotated in this study. Colors are meaningless except to denote approximate regions of annotation. White nodes are unannotated nodes.

| Major cluster(s) | Annotation types | Overall annotation results | Nodes (applied) | Sequences (applied) | Biological Assembly | HMM Cluster from Literature | Any additional clusters |
|---|---|---|---|---|---|---|---|
| **A.1** | L, S, P | Subfamily: MaoC-like hydratase. Contains subreagions with additional annotations | 765 | 10631 | Multiple | 3_MaoC-dehydratase-like, unknown (49, 82) | n/a |
| *bottom of right-most sub-cluster* | | *Acetyl/butyrl transfer? Co-occurs with PF01515* | *70* | *325* | | | |
| *Bottom of central sub-cluster* | | | *113* | *541* | *H2* | | |
| *Central/left of rightmost sub-cluster* | | | *171* | *1331* | *D* | | |
| *End of left-most branch* | | | *15* | *1747* | *TrdH* | *49_unknown* | |
| *left branch of central sub-cluster* | | *Mesaconyl-CoA hydratase* | *28* | *398* | | | |
| *Central sub-cluster, top* | | *Bifunctional PaaZ protein with aldehyde dehydrogenase region (PF00171)* | *46* | *2190* | | | |
| *Right-most sub-cluster, right* | | *Hormone biosynthesis? Co-occurs with short-chain dehydrogenase (PF00106)* | *59* | *227* | | | |
| **A.2 (all)** | S, L, C, T | Acyl-CoA cluster with additional subclusters | 663 | 16433 | Multiple | 1_Acyl-CoA thioesterases | n/a |
| *Upper eukaryote branch* | | *CACH/BFIT* | *44* | *203* | *TrdH* | | |
| *Lower Eukaryote branch* | | *BACH/ACOT7* | *20* | *121* | *H2* | | |
| *Uppermost bacterial sub-cluster* | | | *165* | *6312* | *H2* | | |
| *Central sub-cluster* | | | *212* | *6185* | *H2* | | |
| **A.3** | S, L, C, P | 4HBT-II cluster | 391 | 14929 | Subregion | 8_4HBT II | n/a |
| *Upper left branch* | | *Unknown function-- contains HAD domain, like B. thetaiotaomicron Q89YN2* | *21* | *118* | | | |
| *Central region and upper brach* | | *Cannot assign-- contains EntH and DHNA-CoA annotations* | *180* | *11973* | *TB* | | |
| **A.4** | L, S | YbgC-like cluster | 752 | 12488 | TA | 4_YbgC-like | n/a |
| **A.5** | S, L, P, D | FabZ cluster | 603 | 17572 | Subregion | 2_FabZ-like dehydratases | n/a |
| *Central cluster* | | *FabZ* | *599* | *16036* | *H1* | | |
| *Left region of central cluster* | | *FabZ with LpxX domains* | *69* | *918* | | | |
| **A.6** | S, L | FLK | 257 | 1055 | | n/a | n/a |
| **A.7** | S, L, P | TesB cluster | 651 | 11021 | Multiple | 7_tesB | n/a |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| *Center of cluster* | | *tesB with uncertain ACOT annotation (2, 8)* | *457* | *10548* | *DdhB* | | |
| *Right region above cluster* | | | *18* | *47* | *TB* | | |
| **A.8** | S | Not enough info to annotate everything | 174 | 627 | Subregion | 11_unknown | n/a |
| *Right sub-cluster* | | *ACOT13* | *59* | *189* | *TB* | | |
| **A.9** | S, D, L | Heavily split cluster among FabA/NodN | 500 | 5011 | Subregion | 5_FabA and 10_NodN-like | n/a |
| *Upper sub-cluster* | | *NodN* | *350* | *4724* | *D* | *10_NodN-like* | |
| *Lower sub-cluster* | | *Involved in polyunsaturated fatty acid biosynthesis (co-occurs withPF00109 and PF02801 BKAS domains, PF08659 KR domain, PF00550 PP domain, PF00698 Acyl_transf_1 domain)* | *150* | *287* | | *5_FabA* | |
| **A.10** | S, L | PaaI, primarily single domain protein., likely TB structure. | 265 (223) | 2264 (2112) | TB | 13_PaaI | n/a |
| **A.11** | L | Primarily YbgC | 213 | 2148 | Subregion | 4_YbgC-like, unknown (26) | n/a |
| *Left and upper sub-clusters* | | | *141* | *926* | *TA* | | |
| **A.12** | S, L, P, D | Fat subfamily with additional region of interest | 540 | 3404 | Subregion | 6_Fat subfamily (acyl-ACP thioesterases) | n/a |
| *Bottom-left outcropping* | | *Plant region with some additional PF12590 Acyl domains* | *70* | *421* | | | |
| *Everything else* | | *Fat subfamily* | *470* | *2983* | *DdhA* | | |
| **A.13** | S, L, P, D | Multiple Swiss-Prot annotations, primarily MaoC plus epimerases (see below) | 342 | 5201 | Subregion | 9_MaoC-like | n/a |
| *Right hemisphere* | | *Peroxisomal dehydratase/epimerases, likely assocaited with hormone biosynthesis* | *96* | *538* | *DdhA* | | |
| **A.14** | P | PF03061 (4HBT Thioesterase superfamily) | 144 | 1172 | | 21_unknown | n/a |
| **A.15** | P | PF01575 (MaoC-like domain) | 120 | 582 | | 33_unknown | n/a |
| **A.16** | P | PF03061 (4HBT Thioesterase superfamily) | 175 | 2285 | | 22_unknown | n/a |
| **A.17** | P | PF03061 (4HBT Thioesterase superfamily) | 182 | 1297 | | 19_unknown | n/a |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **A.18** | S, L, T | Mesenchymal stem cell protein/them6 in the eukaryote portion. Strongest for small, lower subcluster but applied across. | 170 | 512 | | 25_Mesenchymal stem cell protein | n/a |
| **A.19*** | L, P | MaoC-like; | 314 | 7170 | | 3_MaoC-dehydratase-like | A.19, B.68, B.184 |
| **A.20** | L, P | YiiD acetylransferase | 94 | 3948 | Subregion | 17_Acetyltransferase | n/a |
| *Upper cluster* | | | *29* | *3626* | *D* | | |
| **A.21** | P | general annotation of acetyltransferase 1 | 150 | 1859 | | 16_unknown | n/a |
| **A.22** | S, P, T | MaoC subfamily | 336 | 1540 | | 38_unknown | n/a |
| *Right region* | | *Mesaconyl-CoA hydratase* | *176* | *1171* | | | |
| *Left region* | | *Hydroxyacyl-thioester dehydratase type 2 mitochondrial 3-hydroxyacyl-ACP dehydratase* | *160* | *369* | | | |
| **A.23** | S, T | acyl coa thioesterase,with wide biological range. Swissprot = acyl coa 9 and 10 mitochondrial. Long and broad chain | 190 | 607 | | 1_Acyl-CoA thioesterases | n/a |
| **A.24*** | n/a | Uncharacterized | 722 | 2776 | | n/a | A.24, B.26, B.31, B.32, B.34, B.35, B.45, B.57, B.66, B.67, B.72, B.69, B.79, B.91, B.100, B.120, B.122, B.127, B.131, B.137, B.152, B.146, B.155, B.161, B.162, B.177, B.180, B.181, B.187, B.188, B.189, B.191, B.192, B.193, B.208, B.214, B.221, B.222, B.225, B.226 |
| **A.25, A.32*** | L | ybgc and ybgc/ybaw like | 248 | 3591 | TA | 4_YbgC-like | A.25, A.32 |
| **A.26** | L | PaaI | 90 | 283 | D | n/a | n/a |
| **A.27** | P | PF03061 (4HBT Thioesterase superfamily) | 83 | 296 | | 34_unknown | n/a |
| **A.28** | S, L, D | 3 hydroxyacyl coa dehydrogenase, with a small subregion | 148 | 5096 | D | 15_Hydroxyacyl-CoA dehydrogenaseassociated | n/a |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| *Bottom region* | | *L carnitine dehydrogenase, associates with 3HCDH domains* | *33* | *135* | *D* | | |
| *Top region* | | *3 hydroxyacyl coa dehydrogenase* | *115* | *4961* | *D* | | |
| **A.29** | S | 3-aminobutyryl-CoA ammonia lyase | 37 | 304 | | n/a | n/a |
| **A.30** | P | general maoc | 174 | 3258 | | 27_unknown | n/a |
| **A.31** | n/a | Unable to assign subfamily or function | 108 | 1299 | | 61_unknown | n/a |
| **A.33** | S, P | A-factor biosynthesis enzyme afsA | 120 | 209 | | n/a | n/a |
| **A.34** | L, S | 4hbt-I; TA structure is applied to the entire cluster due to internal consistency | 32 | 111 | TA | 42_4HBT-I | n/a |
| **A.35, A.54** | L, D | ybgC-like domains throughout, with additional ones | 115 | 1827 | | 4_YbgC-like | A.35, A.54 |
| **A.36** | L, D | ybgC-like domains throughout, with additional ones | 63 | 452 | | 4_YbgC-like | n/a |
| *Central sub-cluster, top* | | *Co-occurs with either acetyltransferase 1 or 7 domain* | *13* | *67* | | | |
| **A.37** | P | PF03061 (4HBT Thioesterase superfamily) | 68 | 297 | | 60_unknown | n/a |
| **A.38** | P | PF03061 (4HBT Thioesterase superfamily) | 38 | 326 | | 46_unknown | n/a |
| **A.39** | L | PaaI protein | 62 | 349 | D | 69_unknown | n/a |
| **A.40** | L | PaaI protein | 89 | 1807 | TB | 18_unknown | n/a |
| **A.41** | p | PF03061 (4HBT Thioesterase superfamily) | 26 | 292 | | 44_unknown | n/a |
| **A.42, B.9, B.10, B.11*** | P | PF03061 (4HBT Thioesterase superfamily) | 1024 | 3786 | | n/a | A.42, B.9, B.10, B.11, B.16, B.17, B.19, B.23, B.25, B.28, B.41, B.36, B.42, B.46, B.47, B.48, B.50, B.56, B.59, B.60, B.70, B.63, B.71, B.77, B.80, B.82, B.84, B.85, B.86, B.87, B.88, B.89, B.90, B.96, B.101, B.103, B.104, B.105, B.106, B.107, B.109, B.110, B.112, B.114, B.116, B.118, B.121, B.124, B.128, B.138, B.140, B.142, B.145, B.148, |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | | B.151, B.153, B.156, B.157, B.164, B.166, B.167, B.168, B.170, B.172, B.174, B.176, B.179, B.182, B.195, B.199, B.200, B.201, B.205, B.206, B.209, B.211, B.215, B.217, B.218, B.220, B.229, B.230, B.81, B.219 |
| **A.43, A.60, A.65** | P | PF03061 (4HBT Thioesterase superfamily) | | | | n/a | n/a |
| **A.44** | n/a | Cannot assign function: contains methylthioribose-1-phosphate isomerase Swiss-Prot annotation but paal assignment from literature | 39 | 510 | D | 31_unknown | n/a |
| **A.45** | S, T | Acyl-coA thioesterase, confirmed as Them4/Them5 by swissprot. | 34 | 111 | | 39_unknown | n/a |
| **A.46** | P | PF03061 (4HBT Thioesterase superfamily) | 23 | 113 | | unknown (50, 76) | n/a |
| **A.47** | P | PF03061 (4HBT Thioesterase superfamily) | 44 | 4308 | | 35_unknown | n/a |
| **A.48** | S, D, L | FapR, confirmed by SwissProt; combined with other domains including HTH_DEOR | 71 | 5089 | D | 2_FabZ-like dehydratases and 23_FapR | n/a |
| **A.49** | S | DHNA CoA | 54 | 129 | TA | 37_unknown | n/a |
| **A.50** | S, L | ybaw, specifically FadM; only applied to right region due to distance and lack of annotation in left | 44 (34) | 3460 (3253) | TA | 28_YbaW | n/a |
| **A.51** | S | DHNA CoA | 29 | 119 | | 8_4HBT II | n/a |
| **A.52** | L | FLK | 12 | 26 | | n/a | n/a |
| **A.53** | S, D, L | FabA, very messy but confirmed by swissprot (specfically fatty acid synthase subunit beta). Contains: acyl transferase 1, maoc, duf1729, nmo pf03060, pf00109 ketoacyl, pf02801 ketoacyl). T+ structure | 153 | 2478 | T | 5_FabA | n/a |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **A.55, B.24** | P,L | PF03061 (4HBT Thioesterase superfamily) | 67 | 577 | | 26_unknown | n/a |
| **A.56, B.30*** | L, P | 4HBT-II | 62 | 269 | | 8_4HBT II | A.56, B.30* |
| **A.57** | S, P | 3-hydroxydecanoyl-ACP dehydratase  confirmed by swissprot | 39 | 5116 | D | 5_FabA | n/a |
| **A.58** | p | PF03061 (4HBT Thioesterase superfamily) | 18 | 63 | | 53_unknown | n/a |
| **A.59** | p | PF03061 (4HBT Thioesterase superfamily) | 25 | 91 | | 57_unknown | n/a |
| **A.61** | L | FLK with D quaternary structure. | 7 | 25 | D | n/a | n/a |
| **A.62, B.154** | P,L | PF03061 (4HBT Thioesterase superfamily) | 16 | 681 | | 41_unknown | n/a |
| **A.63** | L | PaaI | 4 | 389 | TB | 56_unknown | n/a |
| **A.64** | S, D, C | fabz, swissprot labeled as coronafacic acid dehydratase | 4 | 24 | | 2_FabZ-like dehydratases | n/a |
| **B.1** | P | PF01575 (MaoC-like domain) | 261 | 8375 | | unknown (24, 33) | n/a |
| **B.2** | P, D | Fatty acid synthase (fatty acid biosynthesis) based on BKAS N nand C, and FabA domains | 221 | 2063 | | 30_unknown | n/a |
| **B.3** | P | 4hbt, but tentative (contains some multi domains but nothing consistent) | 177 | 430 | | n/a | n/a |
| **B.4** | P | 4hbt but tentative (contains a central area with additional domains) | 219 | 537 | | 47_unknown | n/a |
| **B.5** | P | PF03061 (4HBT Thioesterase superfamily) | 120 | 2245 | | unknown (45, 81) | n/a |
| **B.6** | too diverse | Contains a mix of 4hBT ad acyl PF01643 | 128 | 1391 | | 43_unknown | n/a |
| **B.7** | P, L | Associates with AMP-binding subfamily, though chaotically. | 124 | 1689 | | 29_AMP-binding subfamily | n/a |
| *Top region* | | 1 or 2 AMP-binding Pfam families (PF00501 AMP and/or PF13193 AMP), sometimes with FabA. | 102 | 787 | | 29_AMP-binding subfamily | n/a |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| *Bottom region* | | FabA only, according to Pfam but AMP-binding according to Dillon/Bateman | 22 | 902 | | 29_AMP-binding subfamily | n/a |
| **B.8** | n/a | Unable to assign subfamily or function | 159 | 417 | | 78_unknown | n/a |
| **B.12*** | p | PF07977 (FabA-like domain) | 110 | 622 | | n/a | B.12, B.44, B.62, B.126, B.134, B.144, B.186, B.190, B.198, B.212, B.227 |
| **B.13** | L | PaaI | 77 | 1943 | DdhB | 54_unknown | n/a |
| **B.14** | P | PF03061 (4HBT Thioesterase superfamily) | 58 | 315 | | 52_unknown | n/a |
| **B.15** | L | PaaI | 45 | 387 | TB | 22_unknown | n/a |
| **B.18** | too diverse | multiple stuff, including dufs | 78 | 3996 | | 58_unknown | n/a |
| **B.20** | P | PF03061 (4HBT Thioesterase superfamily) | 28 | 125 | | 74_unknown | n/a |
| **B.21*** | p | PF01575 (MaoC-like domain) | 135 | 566 | | n/a | B.21, B.58, B.76, B.102, B.111, B.117, B.123, B.135, B.136, B.147, B.163, B.160, B.173, B.185, B.196, B.203, B.204, B.165 |
| **B.22** | P | PF03061 (4HBT Thioesterase superfamily) | 21 | 171 | | 18_unknown | n/a |
| **B.27** | L | PaaI probably | 20 | 57 | | 13_PaaI | n/a |
| **B.29** | P | PF03756 (A-factor biosynthesis hotdog domain) | 32 | 61 | | n/a | B.29 |
| **B.33*** | P | PF01643+Acyl | 68 | 105 | | n/a | B.33, B.52, B.74, B.119, B.129, B.194, B.213, B.223, B.224 |
| **B.37, B.55** | P,L | PF03061 (4HBT Thioesterase superfamily) | 89 | 546 | | 20_unknown | n/a |
| **B.38** | n/a | Unable to assign subfamily or function | 8 | 256 | | 59_unknown | n/a |
| **B.39** | p | PF03061 (4HBT Thioesterase superfamily) | 28 | 87 | | 85_unknown | n/a |
| **B.40** | L | general acetyltransferase | 72 | 1224 | D | 65_unknown | n/a |
| **B.43** | p | PF03061 (4HBT Thioesterase superfamily) | 43 | 2053 | | 55_unknown | n/a |

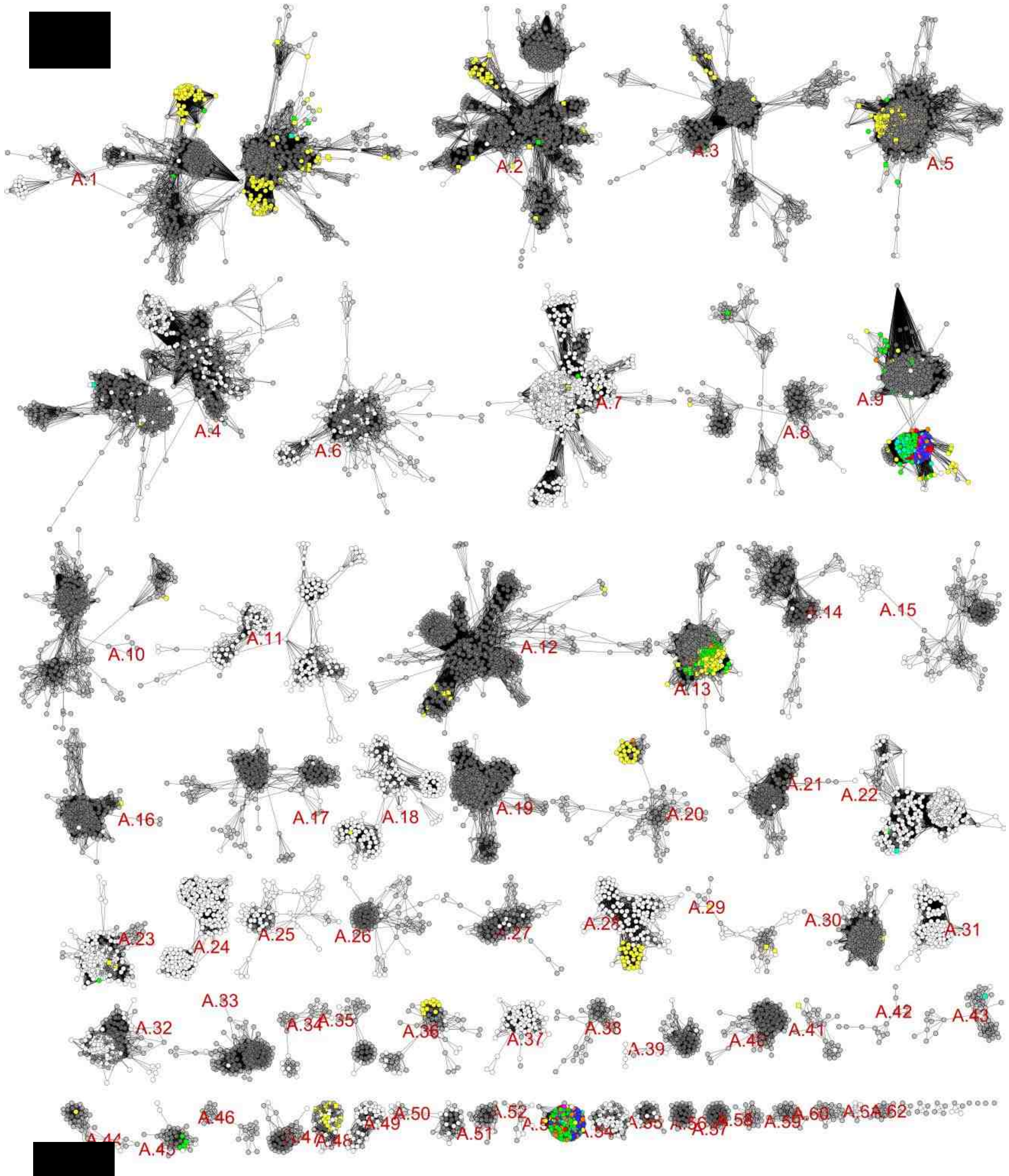
| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **B.49*** | P | PF14539 (Domain of Unknown Function 4442) | 83 | 460 | | n/a | B.49, B.75, B.94, B.98, B.115, B.202, B.228 |
| **B.51** | n/a | Unable to assign subfamily or function | 28 | 73 | | 11_unknown | n/a |
| **B.53, B.197** | P,L | PF03061 (4HBT Thioesterase superfamily) | 18 | 69 | D | 32_unknown | n/a |
| **B.54** | n/a | Unable to assign subfamily or function | 24 | 418 | | 68_unknown | n/a |
| **B.61** | p | PF01643 (Acyl-ACP thioesterase) | 34 | 3426 | | 40_unknown | n/a |
| **B.64** | L | Unable to assign subfamily or function | 24 | 1694 | H3 | n/a | |
| **B.65** | p | PF03061 (4HBT Thioesterase superfamily) | 14 | 38 | | 71_unknown | n/a |
| **B.73** | D, L | CBS-associated: contains CBS domains along with pf07085 drtgg domains. | 74 | 6456 | | 12_CBS-associated | n/a |
| **B.78** | P | some characterization: PF03328 HpcH_HpaI | 6 | 8 | | n/a | B.78 |
| **B.83** | p | PF14539 (Domain of Unknown Function 4442) | 38 | 984 | | 83_unknown | n/a |
| **B.92** | n/a | Unable to assign subfamily or function | 17 | 75 | | 14_unknown | n/a |
| **B.93*** | P,L | PF03061 (4HBT Thioesterase superfamily) | 16 | 4087 | | 48_unknown | B.93, B.178 |
| **B.95*** | L, P | FabZ subgroup of FabA-like domain | 62 | 207 | | 2_FabZ-like dehydratases | B.141, B.108, B.95, B.183 |
| **B.97** | L, P | PF14539 (Domain of Unknown Function 4442) | 18 | 252 | D | | |
| **B.99** | p | PF01575 (MaoC-like domain) | 7 | 59 | | 66_unknown | n/a |
| **B.113** | p | PF03061 (4HBT Thioesterase superfamily) | 8 | 39 | | 80_unknown | n/a |
| **B.125** | P | Unclear function-- MaoC domain with adh short chain dehydrogenase PF00106 | 7 | 25 | | n/a | B.125 |
| **B.130** | n/a | Cannot assign-- multiple domains (4hbt and acyl) on different sequences | 6 | 10 | | | B.130 |
| **B.139** | P | PF03061 (4HBT Thioesterase superfamily) | 12 | 23 | TA | n/a | n/a |
| **B.143** | L | ybaw with Ta structure | 11 | 22 | TA | n/a | n/a |
| **B.149** | p | PF03061 (4HBT Thioesterase superfamily) | 13 | 939 | | 36_unknown | n/a |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **B.150** | p | PF14539 (Domain of Unknown Function 4442) | 8 | 81 | | 77_unknown | n/a |
| **B.158** | L | Hydroxyacyl-CoA dehydrogenase-associated thioesterases; | 12 | 26 | | 15_Hydroxyacyl-CoA dehydrogenaseassociated | n/a |
| **B.159** | L | Acyl-coA thioesterase; | 10 | 30 | | 1_Acyl-CoA thioesterases | n/a |
| **B.169** | p | PF01575 (MaoC-like domain) | 3 | 62 | | 51_unknown | n/a |
| **B.171** | n/a | Unable to assign subfamily or function | 6 | 49 | | 62_unknown | n/a |
| **B.175*** | P | AMP-binding | 7 | 10 | | | B.175 |
| **B.207** | L | AMP-binding subfamily; | 5 | 28 | | 29_AMP-binding subfamily | n/a |
| **B.210** | p | PF03061 (4HBT Thioesterase superfamily) | 5 | 198 | | 64_unknown | n/a |
| **B.216** | L | YbgC-like; TA structure | 5 | 26 | TA | 75_unknown | n/a |

**Table A.2:** Assignment of subfamily, function, and/or structure to all clusters in the hotdog-fold family sequence similarity network. An asterisk in column one indicates that the annotation(s) are applied to more than one cluster, the identities of which are listed in the right-most column. Column two indicates the method by which annotation was assigned: literature such as Dillon and Bateman, Pidugu *et. al*, literature search, or in-house FLK assignment (L); Pfam subfamily annotation from the UniProtKB (P); function or subfamily annotation from the manually curated Swiss-Prot database (S); inference from taxonomic context within the network, taxonomy being acquired from the UniProtKB (T); domain co-occurrence from combined Pfam subfamily annotations and literature descriptions of domain co-occurrence (D).

*A.1.3: Numbers of Pfam domains in the hotdog-fold family SSN*

[redacted]



[redacted]

| | | | |
|---|---|---|---|
| ▩ | 1 domain | 🟪 | 6 domains |
| 🟨 | 2 domains | 🟥 | 7-9 domains |
| 🟩 | 3 domains | 🟧 | Sequences containing different # of domains |
| ▩ | 4 domains | | No domain data |
| 🟦 | 5 domains | | |

[redacted]

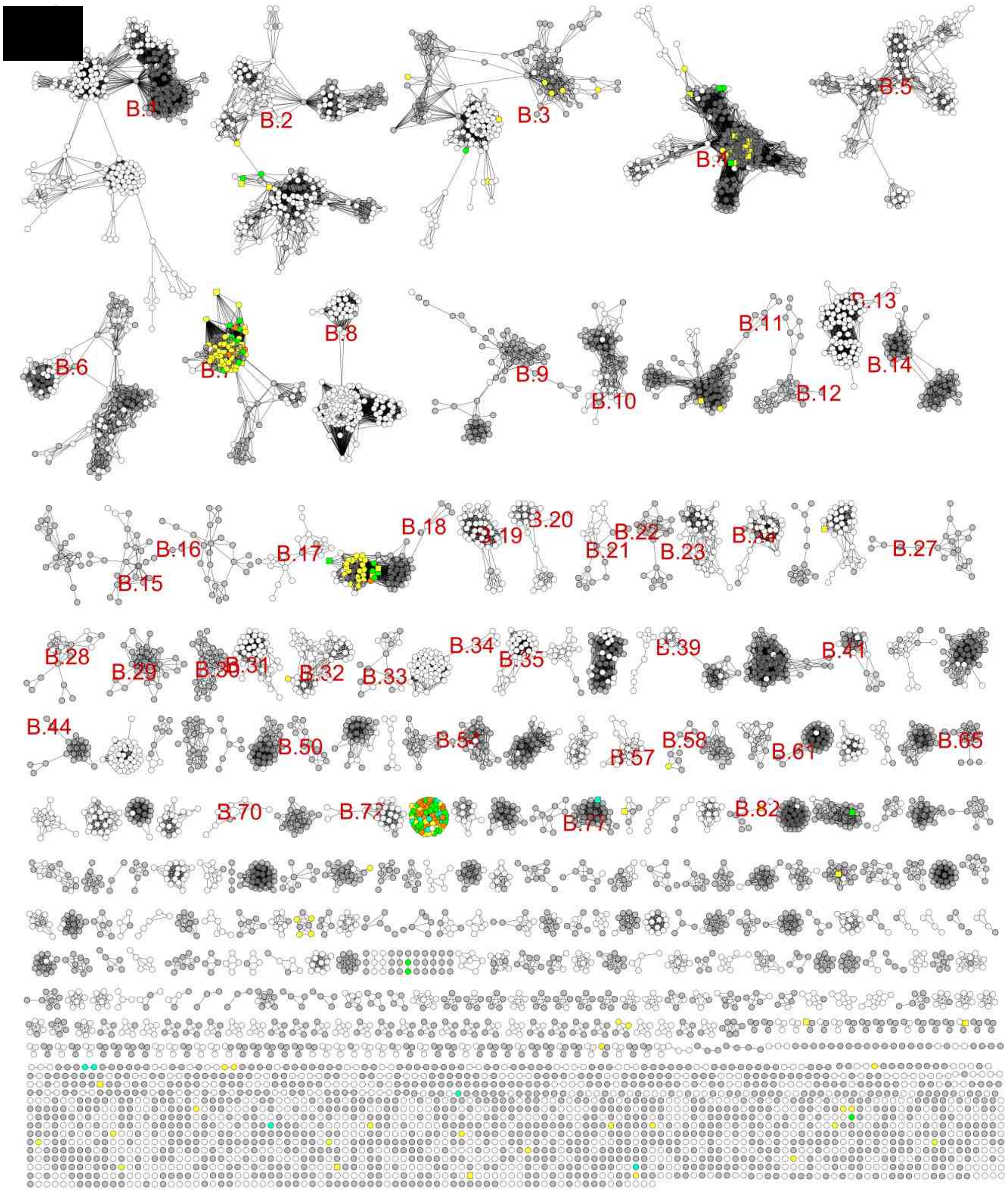
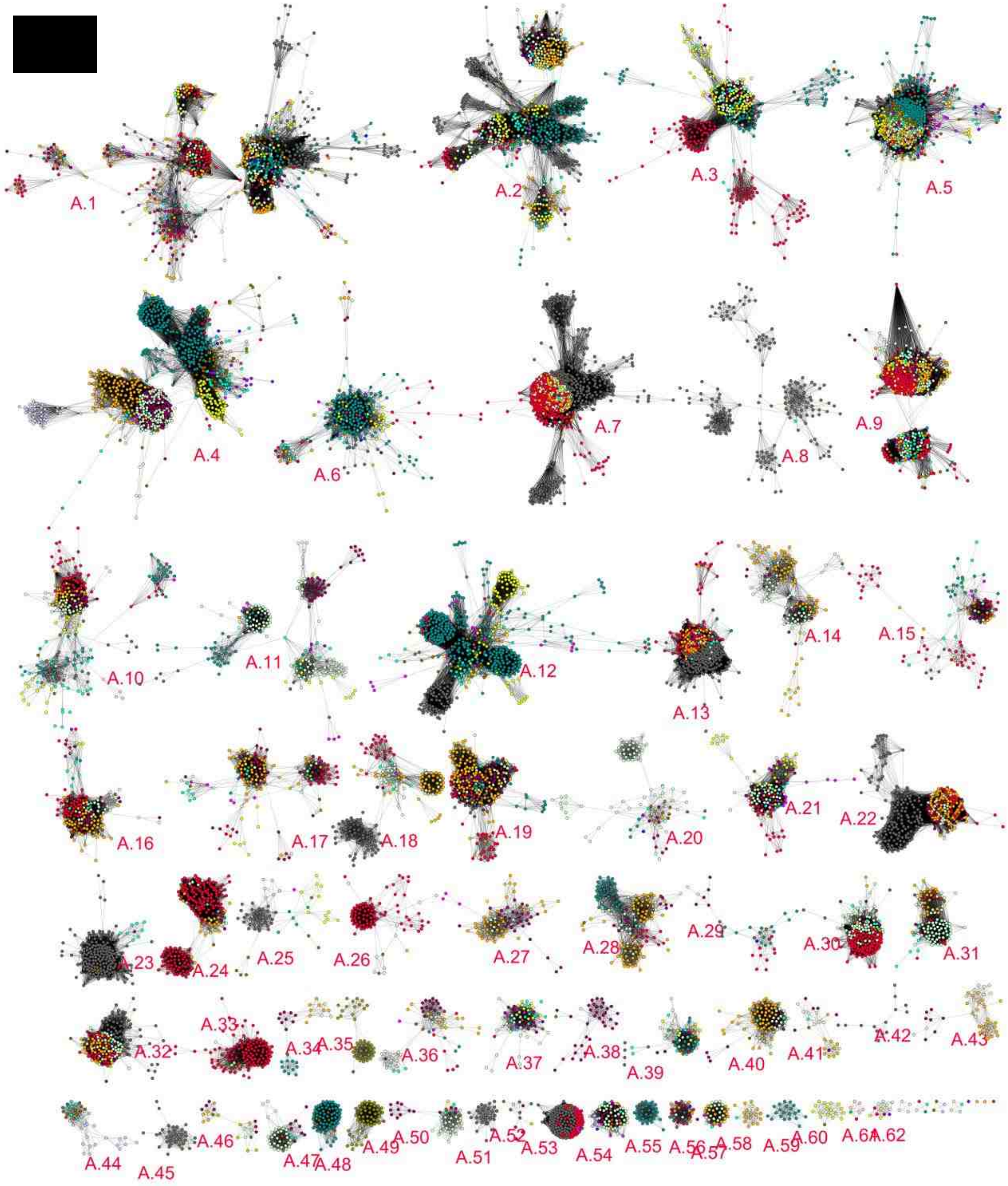


**Figure A.2:** The hotdog-fold SSN painted according to number of different Pfam domain in annotations acquired from the UniProtKB. Rectangular nodes indicate nodes with a combination of 1 domain and n domain nodes (e.g., 1 and 2, 1 and 3, but not 2 and 3); see key (B) for color assignments. Subnetworks A and B are represented in images (A) and (C), respectively.

### *A.1.4: Distribution of phyla in the hotdog-fold family SSN*

[redacted]



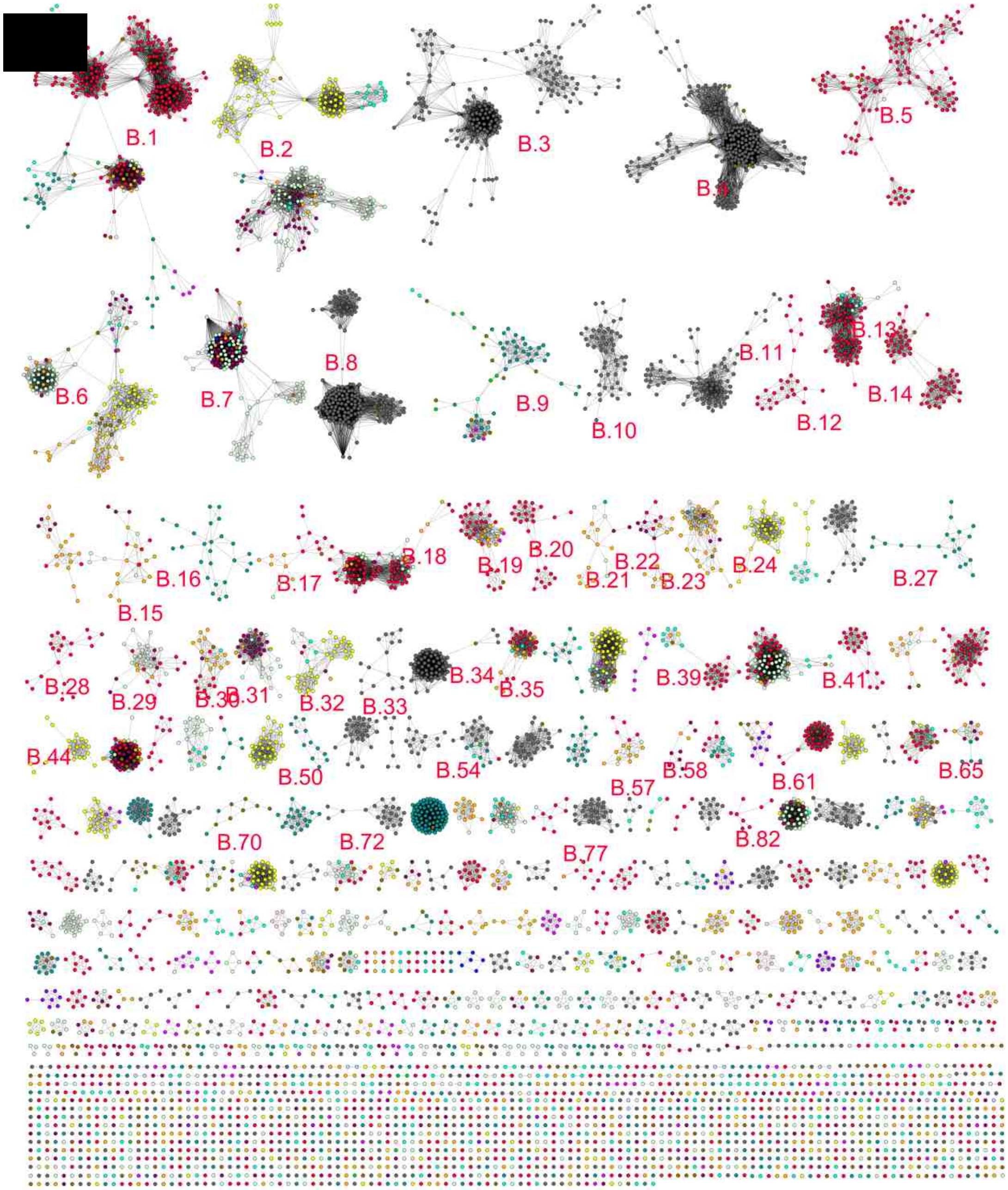
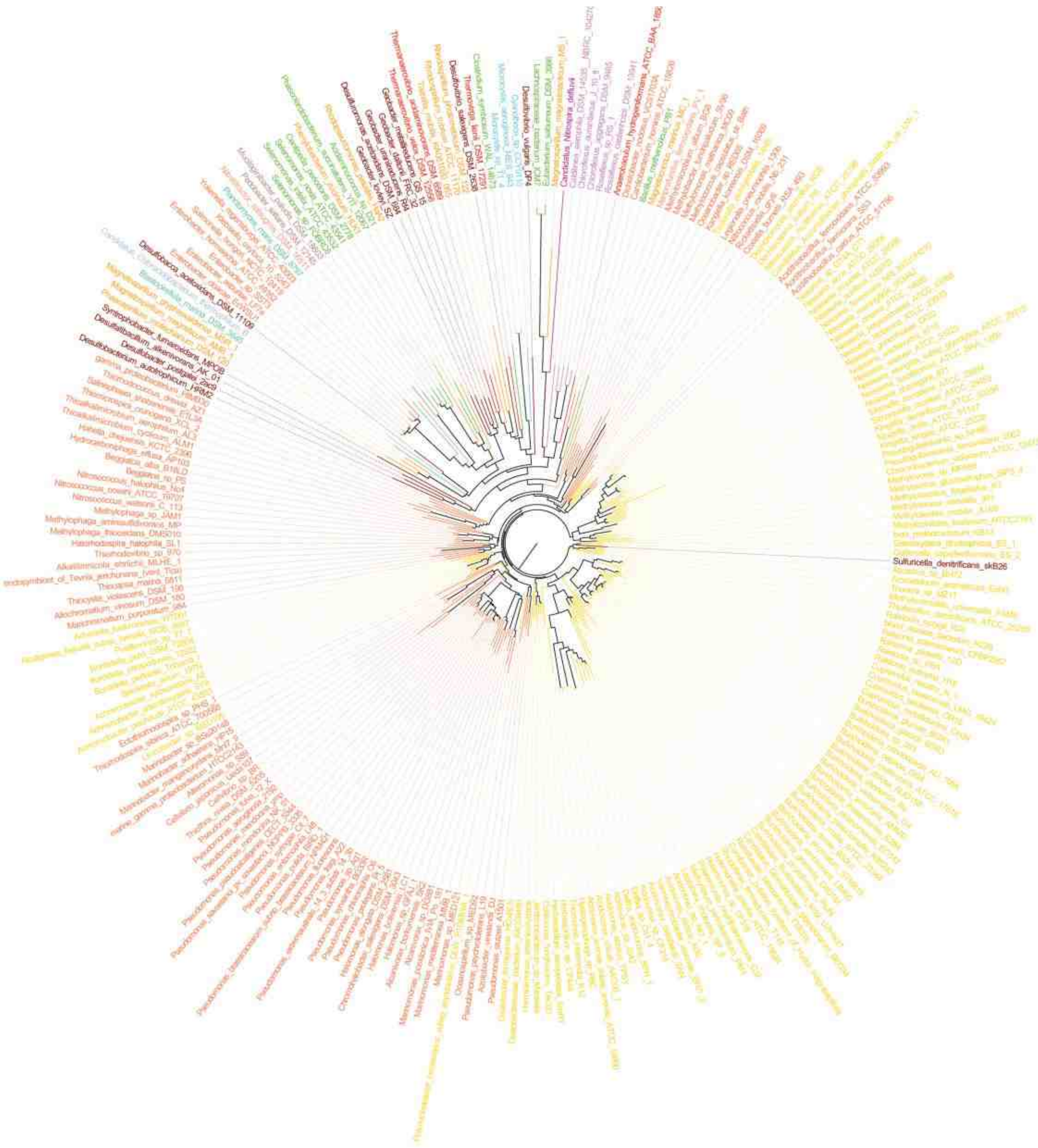


**Figure A.3:** The hotdog-fold SSN (A and B) painted according to distribution of bacterial phyla. Nodes are: Actinobacteria (red), Alphaproteobacteria (orange), Bacteroidetes (yellow), Betaproteobacteria (maroon), Chloroflexi (dark green), Cyanobacteria (sand), Beinococcus-Thermus (pink), Beltaproteobacteria (cyan), Epsilonproteobacteria (lavender), Firmicutes (turquoise), Fusobacteria (blue), Gammaproteobacteria (mint), Planctomycetes (purple), Spriochaetes (magenta), other bacterial phyla (brown). Archaea and Eukaryota are greyed.

### *A.1.5: Biological ranges of EFI HTS proteins*



Desulfobacca_acetoxidans_DSM_11109
Syntrophobacter_fumaroxidans_MPOB
Desulfatibacillum_alkenivorans_AK_01
Desulfobacter_postgatei_2ac9
Desulfobacterium_autotrophicum_HRM2
Thiorhodococcus_drewsii_AZ1
Thiocapsa_marina_5811
Beggiatoa_alba_B18LD
Beggiatoa_sp_PS
Nitrosococcus_halophilus_Nc4
Nitrosococcus_oceani_ATCC_19707
Nitrosococcus_watsonii_C_113
Methylophaga_sp_JAM1
Methylophaga_thiooxydans_DMS010
Halorhodospira_halophila_SL1
Thiorhodovibrio_sp_970
Alkalilimnicola_ehrlichii_MLHE_1
endosymbiont_of_Tevnia_jerichonana_(vent_Tica)
Thiocystis_violascens_DSM_198
Allochromatium_vinosum_DSM_180
Marichromatium_purpuratum_984
Candidatus_Nitrospira_defluvii
Ammonifex_degensii_KC4
Desulfovibrio_vulgaris_DP4
Sulfuricella_denitrificans_skB26

**Figure A.4:** Phylogenetic representation of the biological range of putative orthologs of 501036, displayed at the phylum level. Biological range is confined to Bacteria in the following phyla. Acidobacteria (blue), Bacteroidetes (rosy brown), Chloroflexi (violet), Cyanobacteria (cyan), Firmicutes (green), Nitrospirae (dark violet), Planctomycetes (teal) and Synergistetes (red). Proteobacteria: Alpha (orange), Beta (gold), Gamma (coral), Delta (crimson) and Epsilon/Zeta (both rose).

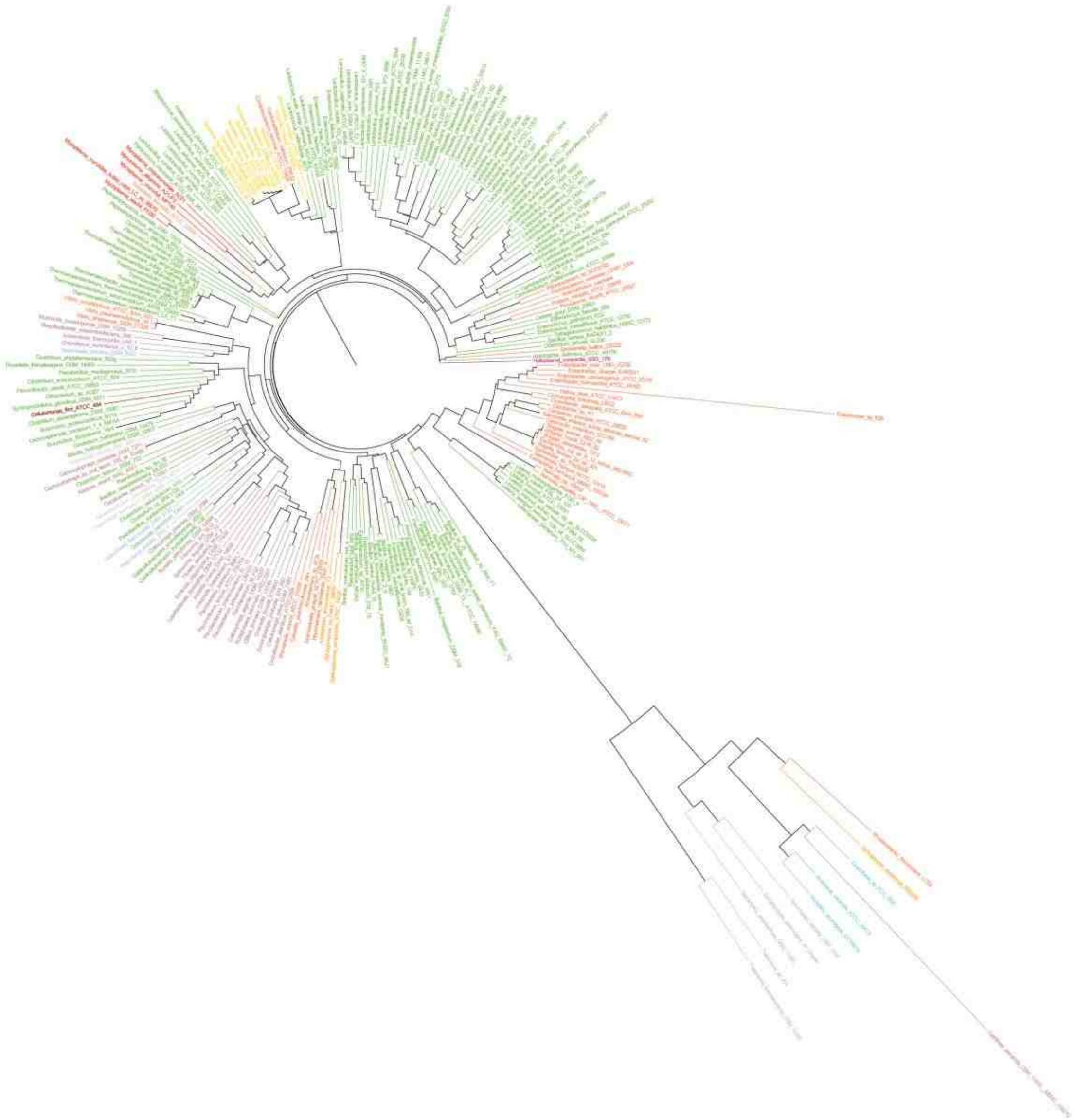


**Figure A.5:** Phylogenetic representation of the biological range of putative orthologs of 501172, displayed at the phylum level. Biological range is confined to Bacteria in the following phyla. Actinobacteria (crimson), Bacteroidetes (brown), Chloroflexi (violet), Cyanobacteria (cyan), Firmicutes (green), Fusobacteria (light brown), Spirochaetes (blue), Tenericutes (red), Thermotogae (grey), Verrucomicrobia (teal) and unclassified (dark violet). Proteobacteria: Alpha (orange), Beta (gold) and Gamma (coral).

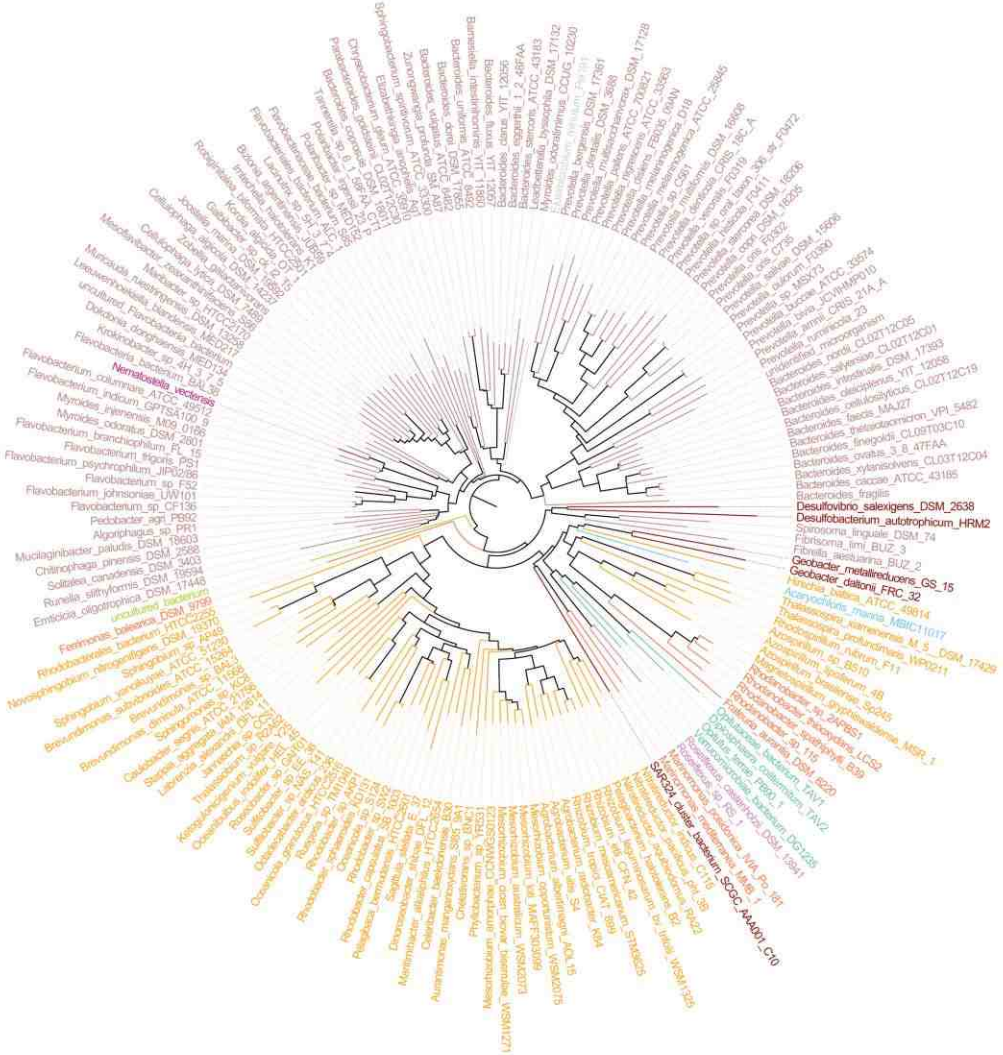

Prevotella_bivia_JCVIHMP010
Prevotella_amnii_CRIS_21A_A
Prevotella_ruminicola_23
unidentified_microorganism
Bacteroides_nordii_CL02T12C05
Bacteroides_salyersiae_CL02T12C01
Bacteroides_intestinalis_DSM_17393
Bacteroides_oleiciplenus_YIT_12058
Bacteroides_cellulosilyticus_CL02T12C19
Bacteroides_faecis_MAJ27
Bacteroides_thetaiotaomicron_VPI_5482
Bacteroides_finegoldii_CL09T03C10
Bacteroides_ovatus_3_8_47FAA
Bacteroides_xylanisolvens_CL03T12C04
Bacteroides_caccae_ATCC_43185
Bacteroides_fragilis
Desulfovibrio_salexigens_DSM_2638
Desulfobacterium_autotrophicum_HRM2
Spirosoma_linguale_DSM_74
Fibrisoma_limi_BUZ_3
Fibrella_aestuarina_BUZ_2
Geobacter_metallireducens_GS_15
Geobacter_daltonii_FRC_32
Hirschia_baltica_ATCC_49814
Acaryochloris_marina_MBIC11017
Thalassospira_xiamenensis_M_5__DSM_17429
Thalassospira_profundimaris_WP0211
Azospirillum_rubrum_F11
Azospirillum_sp_B510
SAR324_cluster_bacterium_SCGC_AAA001_C10
Nematostella_vectensis
Flavobacteria_bacterium_BAL38
Krokinobacter_sp_4H_3_7_5
Flavobacterium_columnare_ATCC_49512
Flavobacterium_indicum_GPTSA100_9
Myroides_injenensis_M09_0166
Myroides_odoratus_DSM_2801
Flavobacterium_branchiophilum_FL_15
Flavobacterium_frigoris_PS1
Flavobacterium_psychrophilum_JIP02/86
Flavobacterium_sp_F52
Flavobacterium_johnsoniae_UW101
Flavobacterium_sp_CF136
Pedobacter_agri_PB92
Algoriphagus_sp_PR1
Mucilaginibacter_paludis_DSM_18603
Chitinophaga_pinensis_DSM_2588
Solitalea_canadensis_DSM_3403
Runella_slithyformis_DSM_19594
Emticicia_oligotrophica_DSM_17448
uncultured_bacterium
Ferrimonas_balearica_DSM_9799
Rhodobacterales_bacterium_HTCC2255
Novosphingobium_nitrogenifigens_DSM_19370

**Figure A.6:** Phylogenetic representation of the biological range of putative orthologs of 501236, displayed at the phylum level. Biological range is confined to Bacteria in the following phyla. Bacteroidetes (rosy brown), Chloroflexi (violet), Cnidaria (dark violet), Cyanobacteria (cyan), Elusimicrobia (slate), Verrucomicrobia (teal) and unclassified Bacteria (green). Proteobacteria: Alpha (orange), Gamma (coral) and Delta (crimson).

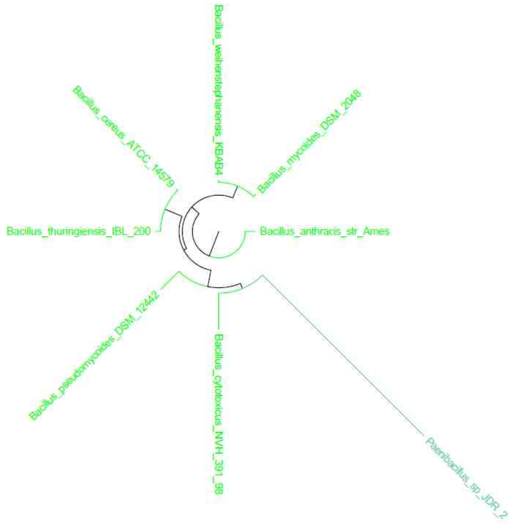

Bacillus_weihenstephanensis_KBAB4
Bacillus_cereus_ATCC_14579
Bacillus_mycoides_DSM_2048
Bacillus_thuringiensis_IBL_200
Bacillus_anthracis_str_Ames
Bacillus_pseudomycoides_DSM_12442
Bacillus_cytotoxicus_NVH_391_98
Paenibacillus_sp_JDR_2

**Figure A.7:** Phylogenetic representation of the biological range of putative orthologs of 501272, displayed at the genus level. Biological range is confined to Bacillales in Bacillaceae (green) and Paenibacillaceae (teal).

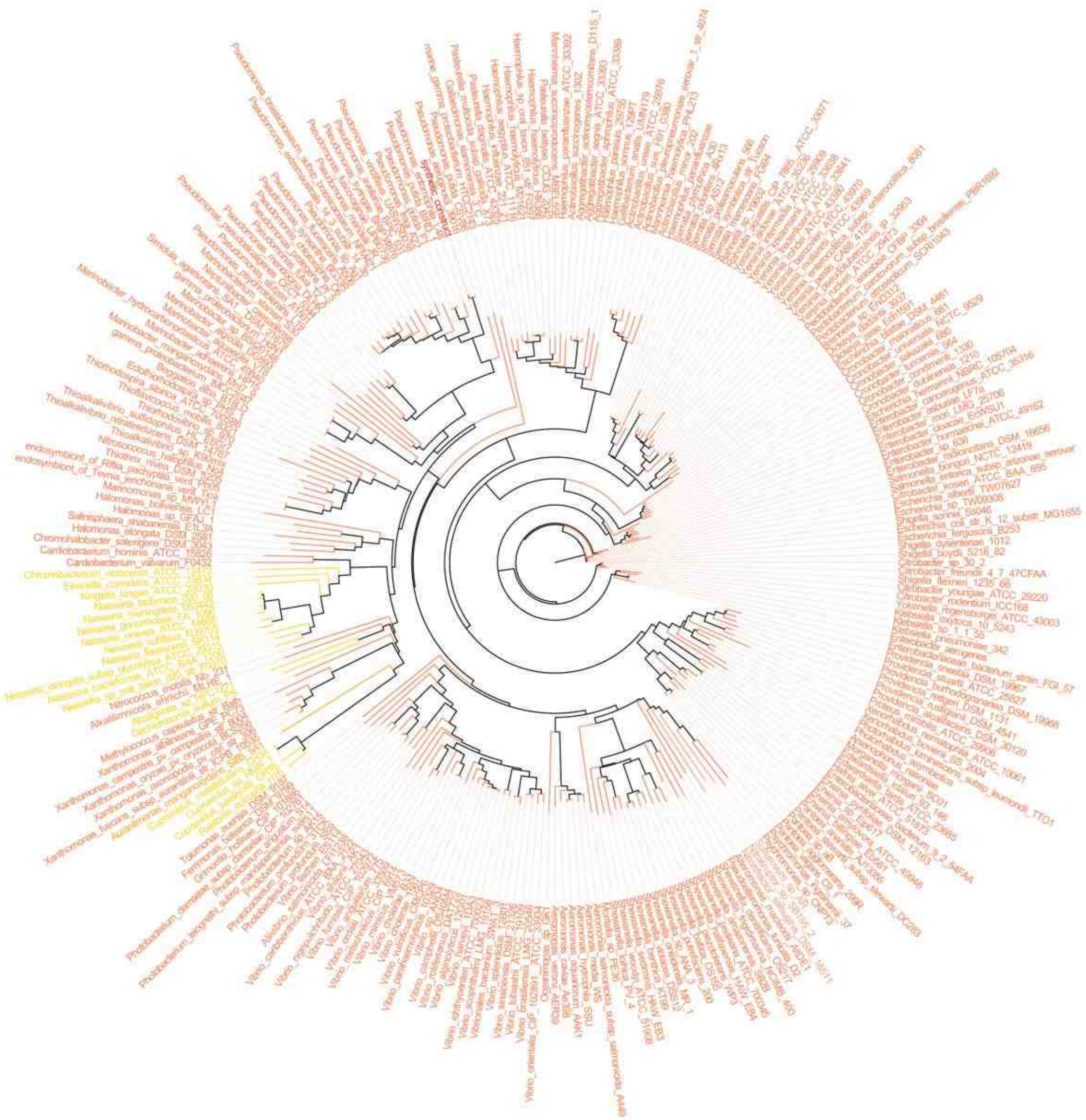


**Figure A.8:** Phylogenetic representation of the biological range of putative orthologs of 501279, displayed at the class level. Biological range is confined to the following Proteobacteria: Alpha (orange), Beta (gold), Gamma (coral), Epsilon (pink) and synthetic construct (red).

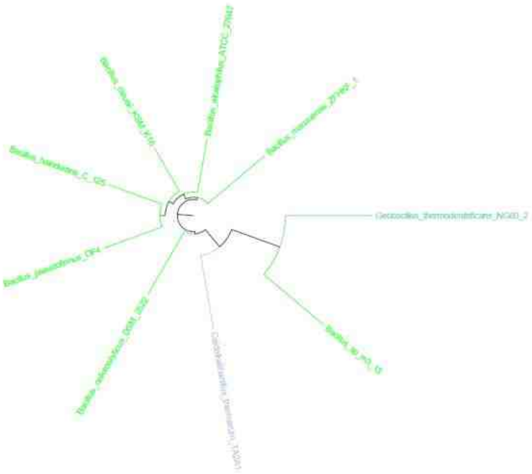

Bacillus_clausii_KSM_K16
Bacillus_alcalophilus_ATCC_27647
Bacillus_halodurans_C_125
Geobacillus_thermodenitrificans_NG80_2
Bacillus_pseudofirmus_OF4
Bacillus_cellulosilyticus_DSM_2522
Caldalkalibacillus_thermarum_TA2A1
Bacillus_sp_m3_13

**Figure A.9:** Phylogenetic representation of the biological range of putative orthologs of 501365, displayed at the genus level. Biological range is confined to Bacillaceae in Bacillus (green), Caldalkalibacillus (slate) and Geobacillus (teal).

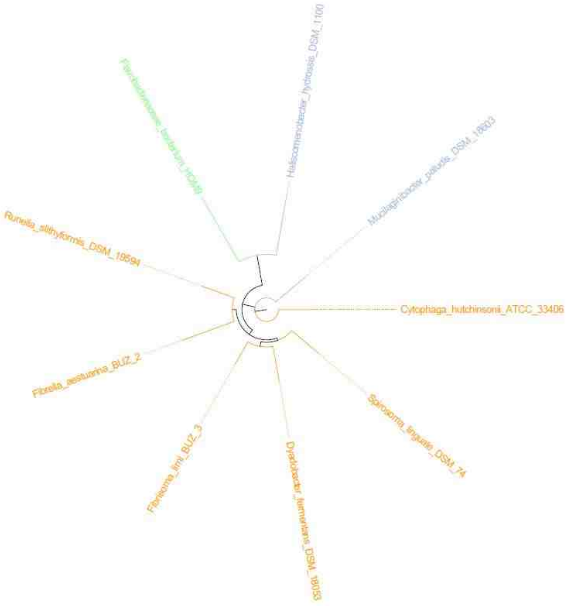

Haliscomenobacter_hydrossis_DSM_1100
Mucilaginibacter_paludis_DSM_18603
Runella_slithyformis_DSM_19594
Cytophaga_hutchinsonii_ATCC_33406
Fibrella_aestuarina_BUZ_2
Spirosoma_linguale_DSM_74
Fibrisoma_limi_BUZ_3
Dyadobacter_fermentans_DSM_18053

**Figure A.10**: Phylogenetic representation of the biological range of putative orthologs of 502337, displayed at the class level. Biological range is confined to phylum Bacteroidetes: Cytophagales (orange), Flavobacteriales (teal) and Sphingobacteriales (slate).

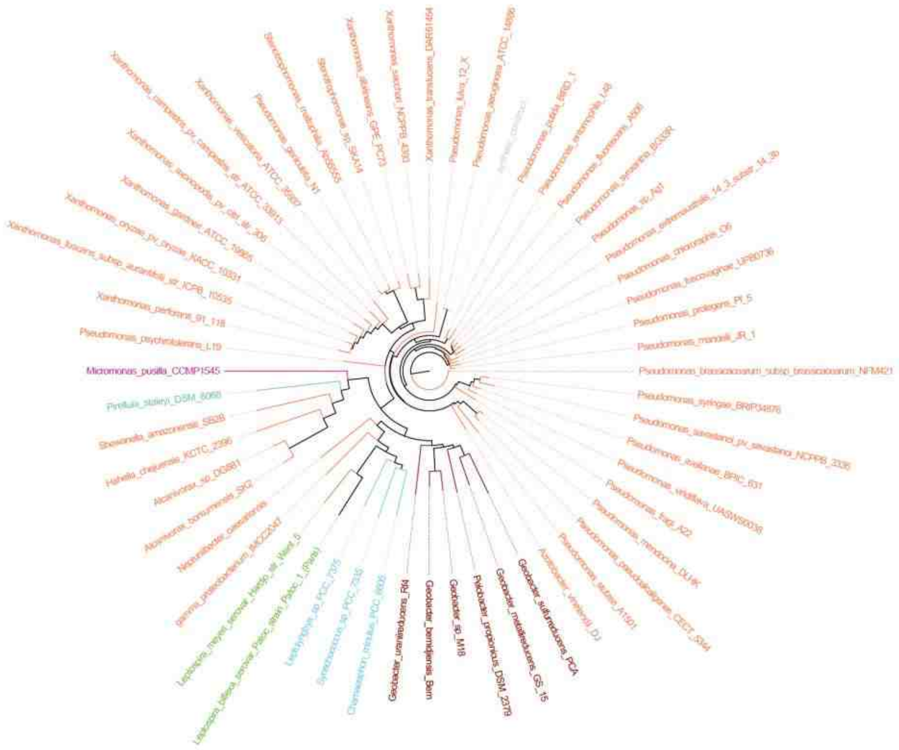
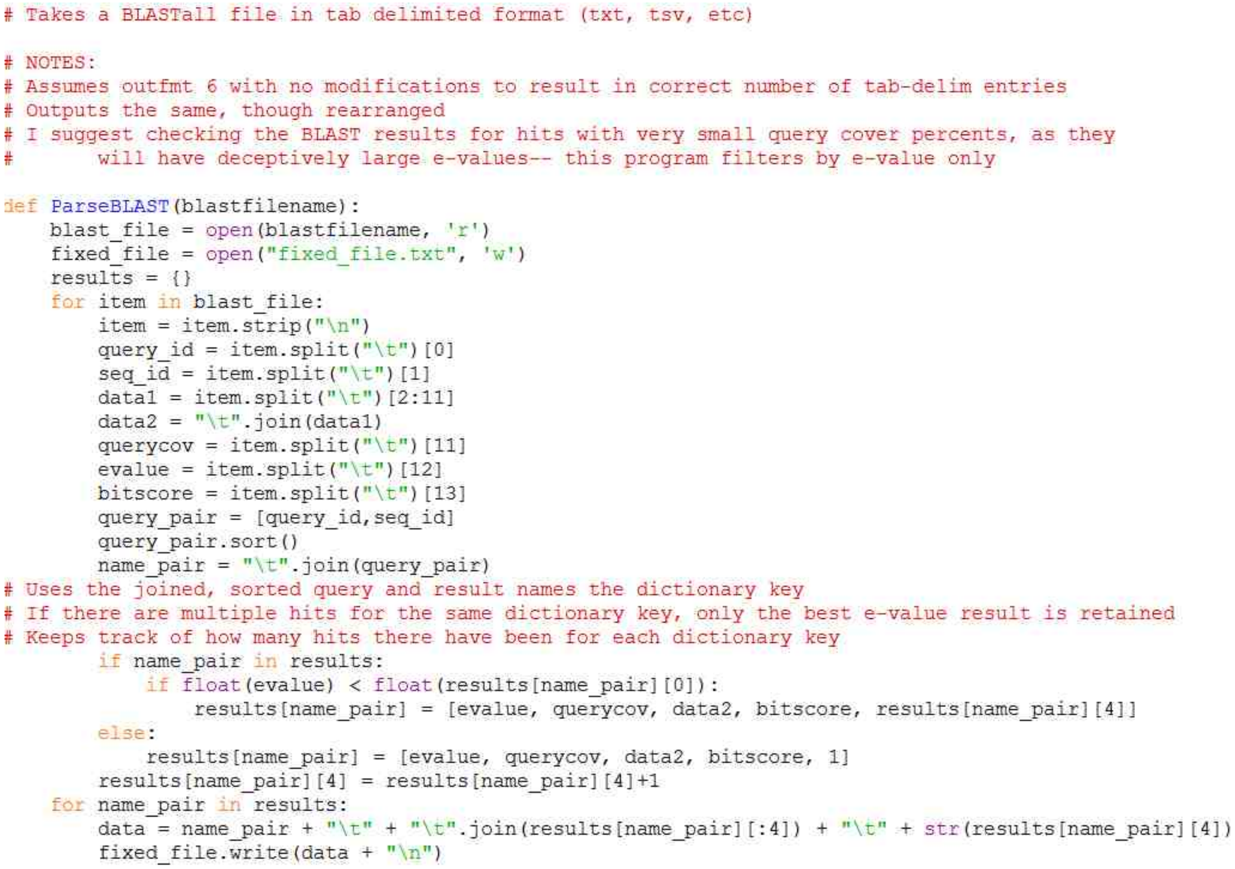

Xanthomonas_fuscans_subsp_aurantifolii_str_ICPB_10535
Xanthomonas_oryzae_pv_oryzae_KACC_10331
Xanthomonas_gardneri_ATCC_19865
Xanthomonas_axonopodis_pv_citri_str_306
Xanthomonas_campestris_pv_campestris_str_ATCC_33913
Xanthomonas_vesicatoria_ATCC_35937
Pseudomonas_geniculata_N1
Stenotrophomonas_maltophilia_R551_3
Stenotrophomonas_sp_SKA14
Xanthomonas_albilineans_GPE_PC73
Xanthomonas_sacchari_NCPPB_4393
Xanthomonas_translucens_DAR61454
Pseudomonas_fulva_12_X
Pseudomonas_aeruginosa_ATCC_14886
synthetic_construct
Pseudomonas_putida_BIRD_1
Pseudomonas_entomophila_L48
Pseudomonas_fluorescens_A506
Pseudomonas_synxantha_BG33R
Pseudomonas_sp_Ag1
Pseudomonas_extremaustralis_14_3_substr_14_3b
Pseudomonas_chlororaphis_O6
Pseudomonas_fuscovaginae_UPB0736
Pseudomonas_protegens_Pf_5
Pseudomonas_mandelii_JR_1
Pseudomonas_brassicacearum_subsp_brassicacearum_NFM421
Pseudomonas_syringae_BRIP34876
Pseudomonas_savastanoi_pv_savastanoi_NCPPB_3335
Pseudomonas_avellanae_BPIC_631
Pseudomonas_viridiflava_UASWS0038
Pseudomonas_fragi_A22
Pseudomonas_mendocina_DLHK
Pseudomonas_pseudoalcaligenes_CECT_5344
Pseudomonas_stutzeri_A1501
Azotobacter_vinelandii_DJ
Geobacter_sulfurreducens_PCA
Geobacter_metallireducens_GS_15
Pelobacter_propionicus_DSM_2379
Geobacter_sp_M18
Geobacter_bemidjiensis_Bem
Geobacter_uraniireducens_Rf4
Chamaesiphon_minutus_PCC_6605
Synechococcus_sp_PCC_7335
Leptolyngbya_sp_PCC_7375
Leptospira_biflexa_serovar_Patoc_strain_Patoc_1_(Paris)
Leptospira_meyeri_serovar_Hardjo_str_Went_5
gamma_proteobacterium_IMCC2047
Neptuniibacter_caesariensis
Alcanivorax_borkumensis_SK2
Alcanivorax_sp_DG881
Hahella_chejuensis_KCTC_2396
Shewanella_amazonensis_SB2B
Pirellula_staleyi_DSM_6068
Micromonas_pusilla_CCMP1545
Pseudomonas_psychrotolerans_L19
Xanthomonas_perforans_91_118

**Figure A.11:** Phylogenetic representation of the biological range of putative orthologs of 900338, displayed at the phylum level. In Bacteria: Cyanobacteria (cyan), Planctomycetes (teal), Spirochaetes (green), Delta Proteobacteria (crimson) and Gamma Proteobacteria (orange). In Eukaryota: Chlorophyta (dark violet). Also, one synthetic construct (grey).

## A.2: Python programs

Python programs were formatted for this manuscript using an online tool accessible at http://www.planetb.ca/syntax-highlight-word

### *A.2.1: ParseBLAST: a Python program to parse blastall results*



```
# Takes a BLASTall file in tab delimited format (txt, tsv, etc)

# NOTES:
# Assumes outfmt 6 with no modifications to result in correct number of tab-delim entries
# Outputs the same, though rearranged
# I suggest checking the BLAST results for hits with very small query cover percents, as they
#       will have deceptively large e-values-- this program filters by e-value only

def ParseBLAST(blastfilename):
    blast_file = open(blastfilename, 'r')
    fixed_file = open("fixed_file.txt", 'w')
    results = {}
    for item in blast_file:
        item = item.strip("\n")
        query_id = item.split("\t")[0]
        seq_id = item.split("\t")[1]
        data1 = item.split("\t")[2:11]
        data2 = "\t".join(data1)
        querycov = item.split("\t")[11]
        evalue = item.split("\t")[12]
        bitscore = item.split("\t")[13]
        query_pair = [query_id,seq_id]
        query_pair.sort()
        name_pair = "\t".join(query_pair)
# Uses the joined, sorted query and result names the dictionary key
# If there are multiple hits for the same dictionary key, only the best e-value result is retained
# Keeps track of how many hits there have been for each dictionary key
        if name_pair in results:
            if float(evalue) < float(results[name_pair][0]):
                results[name_pair] = [evalue, querycov, data2, bitscore, results[name_pair][4]]
        else:
            results[name_pair] = [evalue, querycov, data2, bitscore, 1]
        results[name_pair][4] = results[name_pair][4]+1
    for name_pair in results:
        data = name_pair + "\t" + "\t".join(results[name_pair][:4]) + "\t" + str(results[name_pair][4])
        fixed_file.write(data + "\n")
```

## *A.2.2: gi2taxid2lineage: a Python program to create taxonomic lineages from gi numbers or taxids*

```
1.  # Code to look up a batch of input gi numbers and output their taxIDs in an Excel-friendly format.  Takes
2.  # taxIDs and maps taxonomic lineages
3.  # PROGRAM REQUIREMENTS
4.  #       gi_taxid_prot.dmp (gi --> taxid dmp file from NCBI database)
5.  #       nodes.dmp and names.dmp from taxdump.zip (ncbi taxonomy ftp)
6.  #       an input file called "gi_list.txt", tab delimited txt) gi# queries, one gi per row, one column
7.  #       output files will be taxid_matches.txt and gi2tax2lineage.txt.  Format will be gi# /t taxid
8.  # taxid2lineage-specific notes
9.  # Code to look up a batch of input taxids and output their taxonomic lineages in an Excel-friendly format.
10. #   The bulk of the run time is spent converting the data from names.dmp into a lookup table with lineage information.
11. #   Once the code has been run, the function MakeLineages('inputname', 'outputname') can be called from the command line.
12. #   MakeLineages takes an input file consisting only of one taxid per line and converts it into a tab-delimited table of
13. #       those taxids and their taxonomic groups (when available; leaves an empty space otherwise).
14. #   The taxorder list below defines which taxonomic groups are tracked, and can be changed with no other alterations to the code.
15.
16. # PROGRAM REQUIREMENTS
17. #       nodes.dmp and names.dmp files downloaded from the NCBI taxonomy database.
18. #       an input file (tab delimited txt) containing the taxIDs to be assigned lineages.  One column, one taxID per row, no header.
19. #       name for an output file to be created (also tab delim).  Format will be taxID, kingdom --> species
20.
21.
22. # USER ACTIONS/FUNCTIONS
23. #       MakeLineages, currently disabled in line 153
24. #       TestLineages
25.
26. # BEGIN GI2TAXID
27. # imports necessary gzip file.
28. import gzip
29.
30. # opens gi number file and saves 'locally'
31. gi_set = set(line.strip() for line in open('gi_list.txt', 'r'))
32.
33. # sets namesfile to open and read ('r') from taxid.gz file
34. with gzip.open('gi_taxid_prot.dmp.gz', 'r') as taxidfile:
35.     # sets up an empty list called taxID_matches
36.     taxid_matches = []
37.     # sets up counting lines processed
38.     n = 0
39.     # Reads string and splits into items (lineparts) in a list based on delimitor '\t'
40.     for line in taxidfile:
41.         n = n + 1
42.         # Breaks each line at delimitors, and strips the '\t|\n' off the end of the line
```

```
        linepart = line.strip('\t|\n').split('\t')
        if linepart[0] in gi_set:
            taxid_matches.append(line)
        # prints a line every 1 million lines.
        if n%1000000 == 0:
            print str(n/1000000) + ' million lines'
    write_file = open('taxid_matches.txt','w')
    for line in taxid_matches:
        write_file.write(line)
    write_file.close()
    taxidfile.close()

# BEGIN TAXID2LINEAGE
# sets namesfile to open and read ('r') from names.dmp
namesfile = open('names.dmp', 'r')

# sets up an empty dictionary called namesdict
namesdict = {}

# A small list, converted into a dict, that gives the ranking of taxonomic groups
taxorder = ['kingdom', 'phylum', 'class', 'order', 'family', 'genus', 'species']
rank = {}
for n in range(len(taxorder)):
    rank[taxorder[n]] = n

# Given an argument, returns a list of n copies of that argument, where n is the number of taxonomic groups under consideration
def emptylist():
    return [None for n in range(len(rank))]
listoflists = [[] for n in range(len(rank))]

# Reads string and splits into items (lineparts) in a list based on delimitor '\t|\t'

for line in namesfile:
    # Breaks each line at delimitors, and strips the '\t|\n' off the end of the line
    lineparts = line.strip('\t|\n').split('\t|\t')
    # For each scientific name, makes that name the value associated with the taxid key (as an integer value) in namesdict
    # Also puts in an empty list of size 7 to hold lineage information later.
    if lineparts[3] == 'scientific name':
        namesdict[int(lineparts[0])] = [lineparts[1], emptylist(), None]
namesfile.close()


#sets nodesfile to open and read ('r') from nodes.dmp
nodesfile = open('nodes.dmp', 'r')
# A list of seven (currently empty) sublists into which we will sort everything by taxonomic group
taxon = listoflists
# Reads and splits each string into items, just like before
for line in nodesfile:
        lineparts = line.strip('\t|\n').split('\t|\t')
        # Anything in a major taxonomic group gets dropped into the appropriate sublist of groups, along with parent taxid
        # Otherwise the parent taxid goes straight into the original namesdict, replacing the lineage information
        if lineparts[2] in rank:
            taxon[rank[lineparts[2]]].append((int(lineparts[0]), int(lineparts[1])))
            namesdict[int(lineparts[0])][2] = int(lineparts[1])
        else:
            namesdict[int(lineparts[0])][1] = int(lineparts[1])
            namesdict[int(lineparts[0])][2] = lineparts[2]

    # Modifies the 'root' taxon to act like a kingdom, so the program will know to stop if it follows a lineage to the root
    namesdict[1][1] = emptylist()
    nodesfile.close()


    # Build taxonomic lineage
    # Handle the kingdoms (taxon 0) first, because they are the only ones with no parents as far as we're concerned
    for item in taxon[0]:
        # For all child,parent pairs in first (0th) item/taxonomic level in taxon, pick out only the first (child) taxID (0th).
        # In namesdict, return the emptylist (2nd piece) belonging to that child taxID; for kingdom, return the first (0th) entry.
        # Define this first (0th) entry as the name of the taxID (first/0th piece returned by searching namesdict for child's taxID)
        # KEY NOTE-- what we are changing is on the LEFT.  We are changing it TO what is on the RIGHT.
        namesdict[item[0]][1][0] = namesdict[item[0]][0]


    # For taxa beyond kingdom, build upon previous foundation-- change values for position in taxon list and position in namesdict empty list
    # Create a 'while loop' to continue.
    n = 1
    while len(rank) > n:
        for item in taxon[n]:
            parentID = item[1]
            # If the parent's lineage is just another taxid, that means it's in a category we're not looking at, so we follow it back.
            while type(namesdict[parentID][1]) == int:
                parentID = namesdict[parentID][1]
            # Parent first, then child
            # Parent has already been defined-- can just specificy empty list (second '1st' part of namesdict values)
            #   The [:] is telling PYTHON to make a new copy of the list that can be subsequently modified without changing the original.
            namesdict[item[0]][1] = namesdict[parentID][1][:]
            # Look at nth position in empty list from namesdisct, redefine said position with name of 0th (child) taxID in the nth taxon group
            namesdict[item[0]][1][n] = namesdict[item[0]][0]
        n = n+1

    # Allows testing of individual taxIDs for debugging and ctyoscape trouble-shooting purposes
    def TestLineages(taxIDinput):
        taxid = int(taxIDinput)
        # Looks up the taxid in the namesdict, then repeatedly looks up parents until it finds a proper lineage
        taxinfo = namesdict[taxid]
        while type(taxinfo[1]) == int:
            taxinfo = namesdict[taxinfo[1]]
        lineage = taxinfo[1][:]
        # For printing purposes, replaces all the 'None' entries with single spaces
        for n in range(len(lineage)):
            if lineage[n] == None:
                lineage[n] = ' '
        outputline = str(taxid)+'\t'+'\t'.join(lineage)+'\n'
        outputline_header = 'TaxID\t'+'\t'.join(taxorder)+'\n'
        print(outputline_header)
        print(outputline)


    # The function MakeLineages is called from the terminal after running the program.
    #   It takes two strings as arguments: the name of the input file (consisting of one taxid
    #   per row and nothing else) and the name of the output file.
    #def MakeLineages(inputfilename, outputfilename):
    inputfile = open('taxid_matches.txt', 'r')
    outputfile = open('gi2tax2lineage.txt', 'w')
    # Writes a header line consisting of the label 'TaxID' followed by taxonomic level names
    outputline = 'TaxID\t'+'\t'.join(taxorder)+'\n'
    outputfile.write(outputline)
    for taxid in inputfile:
        # Extracts the taxid as an integer from each line, or throws an error if that's not possible
        try:
            taxid = int(taxid.strip('\n').split('\t')[1])
            gi = taxid.strip('\n').split('\t')[0]
        except:
            print 'The taxid "'+str(taxid)+'" is not in the correct format.'
            quit()
        # Looks up the taxid in the namesdict, then repeatedly looks up parents until it finds a proper lineage
        taxinfo = namesdict[taxid]
        while type(taxinfo[1]) == int:
            taxinfo = namesdict[taxinfo[1]]
        lineage = taxinfo[1][:]
        # For printing purposes, replaces all the 'None' entries with single spaces
        for n in range(len(lineage)):
            if lineage[n] == None:
                lineage[n] = ' '
        # Turns the lineage list into a tab-delimited string and writes it to the output file
        outputline = gi + '\t' + str(taxid)+'\t'+'\t'.join(lineage)+'\n'
        outputfile.write(outputline)
    print 'done'
    inputfile.close()
    outputfile.close()
```

### *A.2.3: ContextBLAST: a Python program to run multiple BLAST searches for gene context generation*

```python
# Code to take a query protein or list of query proteins (format: gi or WP numbers) and search the immediate gene context by species

# PROGRAM REQUIREMENTS
#   a protein ID or list of protein IDs
#   Biopython
#   Internet connection, local BLAST database, or pre-rerun BLAST results

# NOTES
#   The bulk of the runtime here comes from running the BLASTs, especially if they contain many redundant species (eg, E. coli strains)

# USER ACTIONS/FUNCTIONS
#   Check or change the directory under "import os" (first lines of actual code)
#   Run and Run_multiple take the same arguments:(query_accession_number OR query_file, num_neighbors, BLAST_num, per_id, query_cov)
#       query_accession_number OR query_file: either a single input sequence (Run) or a .txt list of input sequences (Run_multiple)
#           with form WP___ or NP___ WITHOUT decimals
#       num_neighbors: the number of neighbors on BOTH sides of query to be used. Eg, for 15 neighbors on each side (30 total), use 15
#       num_BLAST: the max number of BLAST hits, 0 or a number.  10 000 is more rigorous but takes forever; I typically use 5000.  If BLASTs
#           have already been run and you are simply using different parameters, use 0
#       per_id and query_cover: percent ID and % query coverage.  Higher %ID = more stringent.  Use decimals, here, eg 0.30 and 0.70.

# Things to clean up and fix
    # remember to modify number of results for BLAST, or find a way to filter out undesired results
    # make RunBlast standalone-able

#set working directory
from Bio.Blast import NCBIWWW
import os
    #small HP at home
#os.chdir("C:\Users\BToews\Dropbox\Lab stuff\Hot Dog\data from shasha\operon searching\BLAST parser")
    #HP at work
os.chdir("C:\Users\BTdv7\Dropbox\Lab stuff\Hot Dog\data from shasha\operon searching\BLAST parser")
    # big computer athome
#os.chdir("E:\Dropbox\Lab stuff\Hot Dog\FLK (Luke)\BLAST parser\\flA")

# generates list of accession numbers from a query accession number.
def ImportProtein(query_accession_number, num_neighbors):
    # takes the query AC number, breaks into lead and #, populates a list
    # list contains range of ACnumber - 10 to ACnumber + 10
    accession_list = []
    # :3 and 3: are used because WP_#### and ACH###### both have three leading non-numerical values
    ac_number = int(query_accession_number[3:])
    ac_lead = query_accession_number[:3]
    zeroes = len(query_accession_number[3:]) - len(str(ac_number))
    # range 3 gives the number of neighbors to be generated.  Use 21 for 10 on each side.
    range_value = (num_neighbors*2)+1
    for n in range(range_value):
        # for n-1, choose n-(# of items on either side to be blasted)
        # This should already be accounted for in the "num_neighbors" definition
        next_name = ac_lead + "0"*zeroes + str(ac_number + n-num_neighbors)
        accession_list.append(next_name)
    accession_list_noquery = list(accession_list)
    accession_list_noquery.remove(ac_lead + "0"*zeroes + str(ac_number))
    return (accession_list, accession_list_noquery)
    #print accession_list
    #print accession_list_noquery

#Modifying to remove refseq_num and place that in ParseBlast_dict.  This is in order to deal with PDB code issues
def ParseTitle(line, alignment):
    primary_entry = alignment.title.split(" >")[0].strip("]")
    #split_entry = primary_entry.split("|")
    split_entry = primary_entry.split("|",4)
    refseq = split_entry[3]
    try:
        refseq_num = float(refseq[3:])
    #refseq_num = float(refseq[3:])
        description_info = split_entry[4].split(" [")
        #print str(refseq) + " and then " + str(description_info)
        description = description_info[0].strip(" ")
        species = description_info[1].strip("]")
        return (refseq, refseq_num, description, species)
    except:
        return None

def ParseHSPs(line, hsp, blast_record):
    percent_IDs = float(hsp.identities)/len(hsp.sbjct)
    query_cover = len(hsp.sbjct)/float(blast_record.query_letters)
    score = hsp.score
    e_value = hsp.expect
    return (score, query_cover, e_value, percent_IDs)

def ParseBlast_dict(AC, perc_id, quer_cov):
    result_handle = open(str(AC + "_BLAST.xml"))
    #result_handle = open(accession_number)
    from Bio.Blast import NCBIXML
    blast_record = NCBIXML.read(result_handle)
    # empty dictionary created
    dict_name = {}
    for alignment in blast_record.alignments:
        if len(alignment.accession) > 7:
            for hsp in alignment.hsps:
                Title_Parsed = ParseTitle(alignment.title, alignment)
                if Title_Parsed != None:
                    (refseq, refseq_num, description, species) = Title_Parsed
                    (score, query_cover, e_value, percent_IDs) = ParseHSPs(alignment.hsps, hsp, blast_record)
                    # ignores things with % identity less than 30%
                    if percent_IDs > perc_id:
                        if query_cover > quer_cov:
                            if species not in dict_name:
                                dict_name[species] = {}
                                blast_content = [refseq, description, str(score), str(query_cover), str(e_value), str(percent_IDs)]
                                dict_name[species][refseq_num] = blast_content
        return dict_name

    def ContextSearch(query_accession_number, accession_list_noquery, uberdict, num_neighbors):
        write_file = open(str(query_accession_number) + "_results.txt", "w")
        headings = MakeHeadings(num_neighbors)
        write_file.write("\t".join(headings)+"\n")
        query_dict = uberdict[query_accession_number + "_dict"]
        n = 0
        for species in query_dict:
            for qaccession in query_dict[species]:
                qaccession_info = [species] + [str(qaccession)] + query_dict[species][qaccession]
                BLAST_results = qaccession_info
                # looks at each different blast result
                for neighbor_accession in accession_list_noquery:
                    neighbor_dict = uberdict[neighbor_accession + "_dict"]
                    if species in neighbor_dict:
                        output_list = None
                        duplicates = False
                        for naccession in neighbor_dict[species]:
                            naccession_info = [str(naccession)] + neighbor_dict[species][naccession]
                            if naccession - 20 < qaccession and qaccession < naccession + 20:
                                if duplicates == False:
                                    distance = naccession - qaccession
                                    output_list = naccession_info + [str(distance)]
                                    duplicates = True
                                else:
                                    n_duplicate = ["multiple"]*8
                                    output_list = n_duplicate
                                    break
                            elif output_list == None:
                                output_list = naccession_info + ["distant"]
                    else:
                        n_info = ["n/a"]*8
                        output_list = n_info
                    BLAST_results = BLAST_results + output_list
                n = n + 1
                # print "done with" + str(n) + "accessions"
                write_file.write("\t".join(BLAST_results)+"\n")
        write_file.close()

    #pulled from prevoius operon search code
    def MakeHeadings(num_neighbors):
        neighbor_headings = ["Accession#","Accession_full","Description","Score","Query cover","E value","Ident","Distance"]
        query_headings = ["Species","Accession#","Accession_full","Description","Score","Query cover","E value","Ident"]
        all_headings = []
        range_values = (num_neighbors*2)+1
        for heading in query_headings:
            all_headings.append("query " + heading)
        for n in range(range_values):
            if n-num_neighbors!= 0:
                neighbor_number = "Neighbor " + str(n-num_neighbors) + " "
                for heading in neighbor_headings:
                    all_headings.append(neighbor_number + heading)
        return all_headings

    # takes a pre-existing file and counts instances of text under a given column
    def CountOperons(file_name, num_neighbors):
        from csv import DictReader
        range_value = (num_neighbors*2)+1
        write_file_name = file_name.split(".")[0] + "_summary.txt"
        write_file = open(write_file_name, "w")
        column_list = []
        results_list = []
        for n in range(range_value):
            if n-num_neighbors!= 0:
                column_name = "Neighbor " + str(n-num_neighbors) + " Distance"
                column_list.append(column_name)
        num_distant = 0
        num_20 = 0
        num_nomatch = 0
        num_multiple = 0
        header = ["column ID", "%distant", "% +/-20", "%nomatch", "%multiple", "total"]
        write_file.write("\t".join(header)+"\n")
        for column in column_list:
            read_file = open(file_name, "r")
            file_reader = DictReader(read_file, delimiter='\t')
            for line in file_reader:
                neighbor_info = line[column]
                if neighbor_info == "distant":
                    num_distant += 1
                elif neighbor_info == "n/a":
                    num_nomatch +=1
                elif neighbor_info == "multiple":
                    num_multiple +=1
                elif float(neighbor_info) > -20 and 20 > float(neighbor_info):
                    num_20 +=1
                total = float(num_distant + num_20 + num_nomatch + num_multiple)
                total_100 = float(total)/100
            #print "num_distant " + column + " " + str(num_distant/total)
            #print "num_20 " + column + " " + str(num_20/total)
            #print "num_nomatch " + column + " " + str(num_nomatch/total)
            #print "num_multiple " + column + " " + str(num_multiple/total)
            results = [column, str(num_distant/total_100), str(num_20/total_100), str(num_nomatch/total_100), str(num_multiple/total_100), str(total)]
            write_file.write("\t".join(results)+"\n")
            num_distant = 0
            num_20 = 0
            num_nomatch = 0
            num_multiple = 0
            num_str = 0
            total = 0

    def RunBlast(AC, num_BLAST):
        #testing expect stuff
        #result_handle = NCBIWWW.qblast("blastp", "nr", AC, hitlist_size = num_BLAST, expect = 1e-10)
        result_handle = NCBIWWW.qblast("blastp", "nr", AC, hitlist_size = num_BLAST)
        save_file = open(str(AC + "_BLAST.xml"),"w")
        save_file.write(result_handle.read())
        save_file.close()
        result_handle.close()

    def RunBlast_multiple(queries_file, num_BLAST):
        read_file = open(queries_file, "r")
        for AC in read_file:
            #testing expect stuff
            #result_handle = NCBIWWW.qblast("blastp", "nr", AC, hitlist_size = num_BLAST, expect = 1e-10)
            result_handle = NCBIWWW.qblast("blastp", "nr", AC, hitlist_size = num_BLAST)
            save_file = open(str(AC + "_BLAST.xml"),"w")
            save_file.write(result_handle.read())
            save_file.close()
            result_handle.close()

    def Run(query_accession_number, num_neighbors, num_BLAST, per_id, quer_cov):
        (accession_list, accession_list_noquery) = ImportProtein(query_accession_number, num_neighbors)
        if num_BLAST > 0:
            for AC in accession_list:
                RunBlast(AC, num_BLAST)
                print "BLAST complete for " + AC
        uberdict = {}
        for AC in accession_list:
            dict_name = AC + "_dict"
            uberdict[dict_name] = ParseBlast_dict(AC, per_id, quer_cov)
            print "Parsing done for " + str(AC)
        ContextSearch(query_accession_number, accession_list_noquery, uberdict, num_neighbors)
        results_file = query_accession_number + "_results.txt"
        CountOperons(results_file, num_neighbors)


    def Run_multiple(queries_file, num_neighbors, num_BLAST, per_id, quer_cov):
        read_file = open(queries_file, "r")
        for item in read_file:
            query = item.strip("\n")
            Run(query, num_neighbors, num_BLAST, per_id, quer_cov)
            print "done with " + query

    #code for testing
    #result_handle = open("WP_011573347_BLAST.xml")
    #from Bio.Blast import NCBIXML
    #blast_record = NCBIXML.read(result_handle)
    # for alignment in blast_record.alignments:
    #   for hsp in alignment.hsps:
    #       print('sequence:', alignment.title)
    #               (refseq, refseq_num, description, species) = ParseTitle(alignment.title)
```

### *A.2.4: AssignAttributes: a Python program to assign user-defined attributes to sequence similarity network nodes*

```python
# PROGRAM REQUIREMENTS
#   UNMODIFIED clusters file (eg, .NA attributes file) for a sequence similarity network
#   query file

# NOTES

# USER ACTIONS/FUNCTIONS
#   ParseProtein(clusters_file_name, query_file_name, outfile_name)
#       clusters_file_name: the pure, unmodified ACC list from a sequence similarity network (not sure-- I think it can still be a NA file).
#           Eliminates line splitting due to Excel overload
#       query_file_name: .csv query file with lines of ID (the query accession number or uniprot number) and OTHER (if a label is to be applied)
#       outfile_name: name of the output file, as .csv


from csv import DictReader

print 'Requires two inputs, a clusters file and a query file, the query in csv and the clusters in .NA with title "ACC (class"etc'
print 'Clusters file should have two columns with headers KeyID and ACC.  Acc should have form uniprot = uniprot::uniprot::uniprot'
print 'Query file should have two columns with headers ID and OTHER; ID is the query'

#clusters_file_name = 'ProteinIDs hotdog.csv'
#query_file_name    = 'To Find hotdog.csv'

def ParseProtein(clusters_file_name,query_file_name,outfile_name):
    # Parse the "To Find.csv" file.
    query_file    = open(query_file_name, 'r')
    query_reader = DictReader(query_file)
    query         = {line['ID']: line['OTHER'] for line in query_reader}
    query_ids     = set(query.keys())

    # Parse the "ProteinIDs.csv" file.
    clusters_file   = open(clusters_file_name, 'r')
    clusters_reader = DictReader(clusters_file)
    #clusters        = {line['ACC'].split(' = ')[0]: set(line['ACC'].split(' = ')[1].split('::')) for line in clusters_reader}
    clusters        = {line['ACC (class=java.lang.String)'].split(' = (')[0]: set(line['ACC (class=java.lang.String)'].split(' = (')[1].strip(')').split('::')) for line in clusters_reader}

    # Process the two sets of data.
    write_file = open(outfile_name,"w")
    for key_id in clusters:
      cluster_ids = set(clusters[key_id])
      match_ids   = query_ids & cluster_ids
      is_match    = len(match_ids) != 0
      match_datas = [query[mid] for mid in match_ids]
      #print ",".join([key_id, str(is_match), "::".join(match_ids), "::".join(match_datas)])
      write_file.write(",".join([key_id, str(is_match), "::".join(match_ids), "::".join(match_datas)])+"\n")

#ParseProtein('nodes_hotdog_e27.csv','hotdog_query.csv')
```