FUNCTIONAL DISCOVERY IN THE OXIDATIVE D-GALACTURONATE ASSIMILATION PATHWAY AND DEVELOPMENT OF THE ENZYME SIMILARITY WEB TOOL

BY

JASON T. BOUVIER

DISSESSATION

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Biochemistry in the Graduate College of the University of Illinois at Urbana-Champaign, 2015

Urbana, Illinois

Doctoral Committee:

Professor John A. Gerlt, Chair
Professor John E. Cronan
Professor Satish K. Nair
Professor Peter Orlean
ABSTRACT

Sequencing technology has improved dramatically over the past few decades. Before the sequencing of complete genomes was possible, the sequencing of a gene was directly linked to the biochemical characterization of its product [1], however biochemical and genetic characterization has not benefited from being scaled up in the same way as has sequencing. Thus, the scientific community is confronted with exponentially growing sequence databases in which roughly half of the entries are either annotated incorrectly or not at all. Therefore, in order to realize the true potential of the data being generated by sequencing projects, something must be done about the way the functions of those sequences are being discovered and identified.

One approach to addressing the problem of the growing number of sequences without a known function is that set forth by the Enzyme Function Initiative (EFI). The goal of the EFI is to develop tools and strategies to characterize enzymes discovered in genome projects, and the EFI uses an interdisciplinary approach to address the problem. EFI labs include those with expertise in bioinformatics, computational biology, structural biology, enzymology, and biology, that work together to develop a systematic approach that starts with using bioinformatics to select enzyme candidates for structural elucidation, ligand docking to identify potential substrates, in vitro biochemistry to test those predictions, and microbiology to test for the physiological role of activities identified in vitro. The approach just described is the general approach taken, but other tools and approaches also have been tested and developed in each of the areas mentioned (e.g., bioinformatics, computational biology). Bioinformatics tools that...
have been further developed include sequence similarity networks (SSNs) and genomic context networks.

SSNs have a long history and are useful in visualizing trends across groups of related protein sequences, namely function. Before this work, access to SSNs by experimentalists with little bioinformatics training was limited. To provide the ability for experimentalist to generate an SSN for any protein family (~16,000 now in Pfam), we developed a web tool to generate SSNs quickly and easily. The networks can be viewed in Cytoscape and contain an aggregate of annotation data pulled from different sources (e.g., UniProt, GenomesOnline). The first part of this work (Chapter 2) describes the web tool and provides an example in which members of the enolase superfamily from Agrobacterium tumefaciens strain C58 are mined in a shotgun approach to discover novel enzymatic activities.

In the second part of this work, combined bioinformatics and experimental approaches are used to identify two novel enzymes in the oxidative pathway to degrade pectin, the abundant plant cell wall polysaccharide. In the first example (Chapter 3), genomic context and pathway reconstruction combined with in vitro biochemistry and gene expression analysis reveal a novel enzymatic activity of isomerizing the 6-member ring lactone of D-galacturonate (D-galA) to its 5-member ring lactone counterpart. An enzyme to catalyze this reaction had not been identified before this work. In the second example (Chapter 4), in a large scale screening of transporters we were lead to microbial gene neighborhoods containing many enzymes in the known D-galA oxidative pathway but noticed in a number of cases components of the known pathway were missing; in their place candidate enzymes were likely involved in an alternative pathway for metabolizing D-galA. This work lead us to the discovery of an enzyme that
hydrolyzed the 6-member ring lactone of D-galA to its acyclic diacid counterpart, meso-galactarate.

Reference

DEDICATION

Dedicated in memory of my dad, Ricky, and my maw maw, Jo Ann
ACKNOWLEDGEMENTS

First and foremost, I thank and acknowledge my family: Mom, Nanny, Danielle, Melissa, Michelle, the Gentrys et al., and the Bouviers et al. I also thank and acknowledge family members and friends not mentioned here. Without their continuous love, patience, support, and encouragement, none of this would be possible.

I acknowledge my advisor, Professor John Gerlt, for his hard work, dedication to problem solving, and for having sustained a long and successful career. I thank him for providing the space and resources in which to take risks, succeed, fail, and thrive yet still keep me on track. I also thank him for providing a rigorous learning environment and continually raising the bar. I found that he always took what I said seriously, and I continue to learn from this simple gesture.

I thank and acknowledge current and former labmates for guidance, technical advice, and camaraderie: Drs. Xinshuai Zhang, Dan Wichelecki, Katie Whalen, Ben Warlick, Ayano Sakai, Tiit Lukk, Heidi Imker, Hua Huang, Fiona Groninger-Poe, Salehe Ghasempur, Tobias Erb, and Bijoy Desai. I also thank and acknowledge my thesis committee: Drs. Peter Orlean, John Cronan, and Marty Burke. I thank and acknowledge members of Dr. Stephen Farrand’s (Microbiology Department, UIUC) lab for ongoing technical advice on A. tumefaciens, namely Maggie Wetzel, Yinping Qin, and Michael Barnhart, the staff at the School of Molecular and Cellular Biology, namely Shawna Smith, the Department of Biochemistry, namely Jeff Goldberg and Cara Day, the Carl R. Woese Institute for Genomic Biology (IGB), namely Lewis Sterling, Debbie Piper, and Dr. Xudong Guan, and Roger Adams Laboratory.
I thank and acknowledge EFI collaborators for their many contributions to this work: Drs. Brandan Hillerich, Nawar Al-Obaidi, and Matt Vetting from Steve Almo’s group (AECOM); Drs. Eyal Akiva, Shoshana Brown, Gemma Holliday, and Jeffrey Yunes from Patricia Babbitt’s group (UCSF); Drs. Michael Carter, Amy Jones, Ritesh Kumar, Brain San Francisco, Jose Solbiati, and McKay Wood from John Cronan’s group (UIUC); Drs. Chakrapani Kalyanaraman and Suwen Zhao from Matt Jacobson’s group (UCSF); and Dr. Dmitry Rodionov from Andrei Osterman’s group (SBMRI). I also thank and acknowledge the contributions from members of the EFI bridging projects.

There are too many former and current members of the IGB Mining Microbial Genomes theme to name individually, so I thank and acknowledge all of those from the Metcalf, Mitchell, van der Donk, and Zhao groups. The inter- and multi-disciplinary environment of the theme has been necessary and invaluable.
# TABLE OF CONTENTS

**LIST OF FIGURES**

- xi

**LIST OF TABLES**

- xiii

## CHAPTER 1: INTRODUCTION

1.1 The Genomics Era....................................................................................................1
1.2 Post-genomic Era Challenges................................................................................3
1.3 Strategies for Functional Assignment......................................................................6
1.4 EFI-Enzyme Similarity Tool (EFI-EST): A Web Tool for Making Protein SSNs..........................10
   1.4.1 EFI-EST Example Usage – Enolase Superfamily...........................................12
1.5 A. tumefaciens strain C58......................................................................................16
   1.5.1 D-GalA Assimilation in A. tumefaciens.........................................................18
1.6 SBP Initiated Regulon Reconstruction......................................................................20
1.7 Conclusion..............................................................................................................23
1.8 References............................................................................................................24

## CHAPTER 2: EFI – ENZYME SIMILARITY TOOL (EFI-EST): A WEB TOOL FOR MAKING PROTEIN SSNS

2.1 Introduction............................................................................................................29
2.2 Description and Implementation...........................................................................32
   2.2.1 Database Construction.....................................................................................32
   2.2.2 Stage 1: Dataset Generation...........................................................................34
   2.2.3 Stage 2: Analysis and Network Generation...................................................34
   2.2.4 Graph Interpretation.......................................................................................35
   2.2.5 Downloading and Visualizing Networks.......................................................37
2.3 EFI-EST Example Usage – Generation, Visualization, and Analysis of the Mandelate Racemase Subgroup of the Enolase Superfamily from A. tumefaciens ..................................37
   2.3.1 Option B: Identification of the Pfam Entry......................................................40
   2.3.2 EFI-EST: Start Page........................................................................................42
   2.3.3 Analyzing the BLAST Dataset: Specifying an Alignment Score Lower Limit for the SSNs.................................................................44
   2.3.4 Downloading the Full and Representative Node SSNs..................................49
   2.3.5 Visualizing and Analyzing the SSNs with Cytoscape....................................52
   2.3.6 Using Swiss-Prot Annotations........................................................................54
2.4 Experimental Testing Materials and Methods.....................................................57
   2.4.1 Target Cloning..................................................................................................57
   2.4.2 Target Expression and Purification...............................................................58
   2.4.3 High Throughput Screen for Enzyme Activity.............................................60
   2.4.4 Enzyme Activity Verification by NMR............................................................61
CHAPTER 3: GALACTARO-LACTONE ISOMERASE INVOLVED IN D-GALA ASSIMILATION BY A. TUMEFACIENS STRAIN C58

3.1 Introduction...........................................................................................................72
3.2 Materials and Methods........................................................................................76
  3.2.1 RNA-Seq of RNA Isolated from A. tumefaciens strain C58.........................76
  3.2.2 Cloning of Atu3138 from A. tumefaciens strain C58.................................76
  3.2.3 Expression and Purification of A9CEQ7......................................................77
  3.2.4 Construction, Expression, and Purification of the N240D Mutant..............78
  3.2.5 Preparation of D-Galactaro-1,5-lactone by Bromine Oxidation...............80
  3.2.6 Udh Assay....................................................................................................82
  3.2.7 Activity Assay by 1H-NMR Spectroscopy................................................83
  3.2.8 Metal Testing...............................................................................................83
  3.2.9 Kinetic Assay by Polarimetry.....................................................................84
  3.2.10 Crystallization and Structure Determination............................................84
3.3 Results and Conclusion.......................................................................................86
3.4 References...........................................................................................................93

CHAPTER 4: TRAP SBP INITIATED REGULON AND PATHWAY RECONSTRUCTION REVEALS NOVEL COMPONENTS OF D-GALA ASSIMILATION PATHWAYS IN SEVERAL CLASSES OF PROTEOBACTERIA

4.1 Introduction.........................................................................................................96
4.2 Materials and Methods.........................................................................................102
  4.2.1 Bioinformatics Methods for Metabolic Regulon and Pathway Reconstruction..102
  4.2.2 Bacterial Strains and Reagents..................................................................103
  4.2.3 Cloning, Expression, and Protein Purification of Lactone Hydrolase and TctC Solute Binding Protein.................................................................104
  4.2.4 Cloning, Expression, and Protein Purification of the Mutarotase..............104
  4.2.5 Cloning, Expression, and Protein Purification of the Regulators..............105
  4.2.6 Preparation of Lactones.............................................................................107
  4.2.7 DNA Binding Assays..................................................................................108
  4.2.8 Differential Scanning Fluorimetry.............................................................110
  4.2.9 Lactone Hydrolase Activity Screening by 1H-NMR Spectroscopy.............110
  4.2.10 Metal Testing............................................................................................111
  4.2.11 Lactone Hydrolase Kinetic Assay............................................................111
  4.2.12 Saturation Difference 1H-NMR Spectroscopy.........................................111
  4.2.13 Gene Expression Analysis.......................................................................112
4.3 Results and Conclusion.......................................................................................113
  4.3.1 Regulon and Metabolic Pathway Reconstruction....................................113
  4.3.2 Growth and Gene Expression Studies......................................................116
4.3.3 Lactone Hydrolase Activity by $^1$H-NMR Spectroscopy.................................118
4.3.4 Metal Testing........................................................................................................119
4.3.5 Lactone Hydrolase Kinetic Assay........................................................................120
4.3.6 Saturation Difference $^1$H-NMR Spectroscopy..............................................122
4.3.7 Differential Scanning Fluorimetry of TctC..........................................................123
4.4 References..................................................................................................................125

CHAPTER 5: CONCLUSION..........................................................................................130

5.1 Introduction...............................................................................................................130
5.2 Findings of This Work..............................................................................................130
5.3 Future Lines of Investigation....................................................................................134
5.4 References..................................................................................................................136
# LIST OF FIGURES

**FIGURE 1.1** Genomics era timeline................................................................................................. 2
**FIGURE 1.2** Graph of sequence database growth.................................................................................. 4
**FIGURE 1.3** EFI pipeline..................................................................................................................... 9
**FIGURE 1.4** Example SSNs................................................................................................................ 11
**FIGURE 1.5** Conserved enolase structure............................................................................................ 13
**FIGURE 1.6** SSN of enolase superfamily acid sugar dehydratases.................................................. 17
**FIGURE 1.7** Oxidative D-galA pathway............................................................................................... 19
**FIGURE 2.1** Comparison between phylogenetic tree and SSN......................................................... 30
**FIGURE 2.2** Enolase superfamily reactions.......................................................................................... 39
**FIGURE 2.3** InterPro screenshot......................................................................................................... 41
**FIGURE 2.4** EFI-EST start page........................................................................................................... 43
**FIGURE 2.5** EFI-EST complete page.................................................................................................. 45
**FIGURE 2.6** Length histogram........................................................................................................... 46
**FIGURE 2.7** Number of edges histogram............................................................................................ 47
**FIGURE 2.8** Alignment length quartile plot.......................................................................................... 48
**FIGURE 2.9** Percent identity quartile plot........................................................................................... 49
**FIGURE 2.10** Download network files page......................................................................................... 50
**FIGURE 2.11** Network comparisons at different repnode percentages............................................. 52
**FIGURE 2.12** Enolase subgroup protein SSN....................................................................................... 53
**FIGURE 2.13** Enolase subgroup with activities labelled........................................................................ 56
**FIGURE 2.14** SDS page gel of purified enolase candidates................................................................. 60
**FIGURE 2.15** Acid sugars in screening library....................................................................................... 61
**FIGURE 2.16** Proton NMR spectra for L-fuconate dehydration.......................................................... 66
**FIGURE 3.1** Pectin............................................................................................................................... 73
**FIGURE 3.2** Oxidative pathway for D-galA degradation....................................................................... 73
**FIGURE 3.3** *A. tumefaciens* gene cluster............................................................................................ 74
**FIGURE 3.4** LgI substrate comparison................................................................................................ 75
**FIGURE 3.5** PAGE gel of A9CEQ7..................................................................................................... 78
**FIGURE 3.6** PAGE gel of A9CEQ7 mutant........................................................................................ 78
**FIGURE 3.7** NMR spectra for D-Galactaro-1,5-lactone preparation................................................ 82
**FIGURE 3.8** RNA-Seq results for gene cluster.................................................................................. 87
**FIGURE 3.9** $^1$H-NMR spectra of the isomerase substrate and product............................................ 88
**FIGURE 3.10** Reaction monitoring by polarimetry.............................................................................. 89
**FIGURE 3.11** A9CEQ7 structure......................................................................................................... 91
**FIGURE 3.12** Isomerization mechanism............................................................................................. 92
**FIGURE 4.1** Transporters, workflow, and TRAP network for SBP-initiated project.................. 97
FIGURE 4.2  Diagram of oxidative pathway.................................................................99
FIGURE 4.3  TRAP network and DSF results; regulon reconstruction......................101
FIGURE 4.4  Preparation of D-Glucaro-1,5-lactone...................................................108
FIGURE 4.5  Reconstructed pathways and regulons....................................................115
FIGURE 4.6  Gene neighborhood in R. pickettii 12J..................................................116
FIGURE 4.7  Growth on D-galA................................................................................117
FIGURE 4.8  Gene expression results........................................................................118
FIGURE 4.9  NMR Activity Screening......................................................................119
FIGURE 4.10  Metal Screening..................................................................................120
FIGURE 4.11  Polarimetry Results for Lactonase.......................................................121
FIGURE 4.12  SD NMR for Mutarotase Activity.......................................................123
FIGURE 4.13  TctC DSF results...............................................................................124
## LIST OF TABLES

| TABLE 1.1 | Enolase conserved residues | 14 |
| TABLE 2.1 | EFI-EST node attributes | 33 |
| TABLE 2.2 | Cloning primers | 58 |
| TABLE 2.3 | SSN publication list | 62 |
| TABLE 2.4 | Screening assay results | 64 |
| TABLE 2.5 | Atu2811 kinetic constants | 67 |
| TABLE 3.1 | Peak assignments for D-galactaro-1,5-lactone | 81 |
| TABLE 3.2 | Structure data collection | 85 |
| TABLE 4.1 | Cloning primers for regulators | 106 |
| TABLE 4.2 | FPA and EMSA primers | 109 |
| TABLE 4.3 | Kinetic constants for lactonase | 122 |
CHAPTER 1: INTRODUCTION

One of the most fundamental patterns of scientific discovery is the revolution in thought that accompanies a new body of data [1].

1.1 The Genomics Era

The first papers describing the structure of DNA were published in 1953 [2-4]. Since then, our ability to determine the nucleotide sequence of this important polymer has advanced dramatically [5]. In the past, determining a single gene sequence was worthy of publication on its own [6], but advances in sequencing technologies have made it almost routine for a researcher to determine the genome sequence of an entire organism (Figure 1.1) [7, 8]. The newest challenges being faced are in the area of sequencing the genomes of all organisms from an entire community (i.e., metagenome). The period of time encompassing these achievements is referred to as the genomics era.
Advances in sequencing technology are enabling researchers to generate a body of data of unprecedented size. (Those advances, such as reaction chemistries and bioinformatics pipelines, are not described in this work.) The Universal Protein Resource Knowledgebase (UniProtKB) is an example of one way researchers have attempted to organize and make this information accessible. After a genome sequence is determined, its open reading frames are translated into protein sequences, and those protein sequences are deposited into UniProtKB. UniProtKB is the preeminent source for publicly available protein sequences and their related information. As of this spring, UniProtKB contained more than 91 million sequences [9]. The
The database has grown nearly 10-fold since I began graduate school in 2010, and the trend continues.

The information in UniProtKB holds great promise as a source of basic discoveries with potential applications in medicine and industry. The importance of this information becomes even more apparent when it is considered that according to metagenomics studies, up to 90% of the microbes on earth remain uncultivated. Therefore, the only way to develop a more complete understanding of the critical roles microbes have and their impact on human health and the environment is by accessing their genome sequences. Research endeavors that seek a complete understanding of microbial physiology, the role of microbes in global nutrient cycles, and the role of the human microbiome in human health undoubtedly will depend on our ability to maximize the use of sequence data now and for many years to come [7].

1.2 Post-genomic Era Challenges

The rate at which genomes are being sequenced far outpaces the rate at which the proteins they encode can be characterized functionally (Figure 1.2). Thus, genomic era solutions have ushered in a host of post-genomic era problems. Researchers interested in protein function are faced with the challenge of determining what these newly deposited proteins do and interpreting the data sets such that they can be put to practical use. For experimentalists, the sequence datasets can be dizzying, and the question often becomes where to begin; something even as seminal as generating a hypothesis becomes a challenge [Dr. Abigail Salyers, personal communication, 2011]. Hence, fewer than 2.5% (2.3 million) of the sequences in UniProtKB are associated with direct experimental evidence, and the annotations for 74% (1.7
million) of those have yet to be reviewed by a curator. This leaves only 600,000 sequences (0.7%) with manually curated annotations, which are then to be used as the source for automatic annotation of the remaining 87.7 million sequences that lack experimental evidence [8]. Furthermore, conservative estimates suggest that approximately half of the 87.7 million sequences have no known functional homologue and thus are devoid of any associated functional information whatsoever [10].

Figure 1.2 Exponential growth of the sequence databases as a function of time. Unreviewed sequences depicted in red; reviewed sequences depicted in black. The red shaded area dwarfs the reviewed sequences [9].

The challenge of annotating uncharacterized protein sequences is further complicated by a number of factors. Sequence databases are extremely biased and do not represent true biological diversity. This bias is due to the fact that genomes are selected for sequencing based on their application potential and not their phylogenetic diversity [11]. Metagenomics studies have identified 60 major phyla; yet, 88% of all bacterial isolates come from just 4 of those
phyla: Proteobacteria, Firmicutes, Actinobacteria, and Bacteroidetes. These isolates represent the majority of sequenced genomes. Furthermore, 50% of the 60 identified phyla have no sequenced representative at all [12]. Such biases cause researchers to draw inaccurate or incomplete conclusions about particular taxonomic groups of organisms based on genomes that may not accurately represent the taxonomic group in question [7]. Therefore, to address this problem, researchers will need to expand the biological diversity represented in the sequence databases, and this expansion will cause dramatic growth of the databases.

Currently, the bacterial and archaeal type strain collections contain 11,000 different organisms. About 3,000 of those organisms are currently part of sequencing projects, but that 3,000 represents only about 4% of the known microbial diversity on the planet (73,000 species). Thus, we are only beginning to scratch the surface of understanding the diversity of organisms on the planet. An effort has been initiated to sequence the remaining 8,000 type strains, and that effort would help to approach roughly 15% of the known microbial diversity (at an average of 3,000 proteins encoded per genome, this project alone would add an additional 24 million sequences) [11].

Yet that would still leave 85% of known microbial diversity unexplored; it is estimated that exploration of that space would yield an additional 187 million sequences. Moreover, these estimates are based on known microbial diversity. 16s rRNA databases contain approximately 500,000 unique sequences, and some estimate that incorporation of unique species would push this number into the millions [11, 12]. To imagine a time when all genomes have been sequenced is difficult, but Galperin and Koonin note that a single genome from a particular taxon does not provide enough information to determine the biological diversity of
that taxon, and that even different strains of the same species can have very different genetic make ups resulting in unexpected discoveries [7]. Others are calling for strain sequencing to become a routine part of strain depositing. Currently, about 650 new species names are issued each year, and sequencing each new species would result in about 2 million additional protein sequences per year [11]. Considering 1) the backlog of unsequenced genomes, 2) the push for sequencing efforts to become routine, and 3) the fact that current estimates exclude the much larger genomes of eukaryotes, to imagine that protein sequence databases would contain hundreds of billions of sequences is not that far-fetched. Whether one measures the growth of these databases by the number of sequences or by biological diversity, the functional annotation community is at the very beginning stages of addressing this challenge.

There is no doubt that the information in the sequence databases holds great promise and that attempting to determine protein functions as sequence databases grow will continue to be a major challenge. However, so far, very little has been done to address these problems. It is clear that in order to realize the full potential of these datasets, advances in large-scale functional assignment are needed.

### 1.3 Strategies for Functional Assignment

Since experimentally characterizing each protein individually is time and cost prohibitive, current approaches rely heavily on computational strategies. In fact, the primary method used to assign functions to proteins on a large scale is automatic annotation based on sequence homology. This approach often fails, particularly for enzymes, to predict the correct substrates and subsequent molecular functions because homologous, or evolutionarily related,
enzymes do not necessarily catalyze the same reactions; sequence homology does not equal overall function homology, and in fact homologous families of proteins can perform/catalyze overall functions that vary considerably [10]. Annotation errors in public databases have been reported to be from 5-63% and can be up to 80% in some protein families, thus the limitation of using sequence homology to assign function to uncharacterized proteins is exacerbated by over-annotation, misannotation, and subsequent propagation [13].

Because the current approach relies heavily on computational strategies, the solution to the functional annotation problem is often framed as such. Large-scale efforts are underway to continue to develop and evaluate computational approaches through the Critical Assessment of Functional Annotation (CAFA) project [14, 15] (analysis of each of those computational approaches is outside the scope of this work). Computational approaches to solve the problem hold great promise for the future, but, in the interim, the work described in this text is focused on an approach that combines experiment with computation.

Protein function annotation that combines experiment with computation has a history dating back decades, but genome scale assignment of protein function is a more recent phenomenon. The buildup of sequenced genomes led to the call for community action in the mid-2000s. Richard Roberts, now of Computational Bridges to Experiments (COMBREX), was one of the first to call for community action in identifying function on a large scale to deal with the ballooning sequence data sets [16]. That same year, Peter Karp of the pathway database MetaCyc, focused on enzymes and called for an enzyme genomics initiative [17]. It was not long after, when in 2010, two major efforts were undertaken: COMBREX and the EFI. Both groups recognize that large-scale functional assignment is too big a problem for a single laboratory to
solve alone. Both COMBREX and the EFI have advocated for an interdisciplinary approach that brings together bioinformaticians, structural and computational biologists, experimental enzymologists, and biologists [10, 18].

“That the scientific community is inundated with DNA sequence data without comparable functional validation studies comes as no surprise . . .” according to Martin Steffen, principal developer of COMBREX [19]. COMBREX employs a unique, decentralized approach similar to “crowd sourcing” initiatives. In COMBREX, computational biologists deposit functional predictions into the COMBREX database. Experimentalists select predictions to test and then bid for small grants ($5,000-$10,000) to work on deciphering the protein’s function [18].

The EFI employs a multi-institution, interdisciplinary approach with the goal of developing a robust, integrated sequence/structure-based strategy to assign functions on a large scale to enzymes discovered in genome projects (Figure 1.3). The strategy currently being developed by the EFI relies on a number of approaches including the use of SSNs, high throughput screening, and in silico ligand docking to predict substrates for uncharacterized enzymes in functionally diverse superfamilies in which functions cannot be assigned based on sequence homology [10]; however, other approaches are also being tested. The goal of the EFI is to develop resources, tool, and technologies and to demonstrate the utility of those technologies.
Figure 1.3 A) Overall EFI pipeline. B) EFI Funnel depicting throughput and providing a description at each of the stages; also demonstrates the iterative nature of the pipeline [10].

The EFI, of which much of my dissertation research has been a part, has many strengths, including the fact that it draws on experts from different fields and makes use of a high throughput “pipeline” for cloning, protein purification, and structure determination, to name a few. The limitations of such a large scale approach include biases and limitations in physical libraries, challenges in discovering new metabolites, limited docking templates, and limited organisms with which to work. Collectively, the EFI libraries of physical substrates, genomic
DNA, and organisms are actually quite extensive, but these are dwarfed when compared to nature’s true biological diversity. Thus, reagent bias parallels the inherent bias of sequence databases.

One public criticism of the EFI is the amount of time and resources it takes to actually identify a new substrate for an unknown enzyme [20], although, in fact, that is exactly one of the problems EFI investigators are trying to address. One proposed solution has been to start with the backlog of identified activities for which no gene or protein sequence has been identified (i.e., orphan Enzyme Commission numbers); thus, the EFI has made orphan enzymes an active area of investigation. In the case of orphan enzymes, there is often a history of experimental work leading up to the identification of the activity, and this experimental work can be leveraged thereby reducing the overall time and cost of the functional investigation [20]. Another criticism is that the EFI initially focused on only five superfamilies, representing about 100 of the roughly 14,000 Pfam families [21]. Again, EFI investigators have worked to address this criticism; one example, described in more detail later, is the development of a protein SSN web tool to help researchers visualize relationships between sequence and function across protein superfamily or families. This tool is easy to use and free of charge to use by anyone in the community who would like to understand their favorite enzyme in the context of its protein family (family context).

1.4 EFI-Enzyme Similarity Tool (EFI-EST): A Web Tool for Making Protein SSNs

Protein SSNs are a powerful tool for visualizing relationships between sequence and function in protein superfamilies and families (Figure 1.4) [22] plus enable a researcher to
investigate the nuances of functional distribution at superfamily or family levels. To facilitate our ability to apply this tool to other superfamilies thereby addressing the criticism by Bastard et al., [21] and to enable researchers outside the EFI to generate networks for their favorite family, we developed a web tool for the non-specialist to generate SSNs easily and quickly. Automatic annotation at the family level is robust, whereas automatic annotation at the subfamily level is not. Thus automatic annotation could stop at the family level [7]. Thus, the EFI has provided a medium through which nearly any experimentalist can begin to navigate sequence datasets of his or her choosing across families and at subfamily levels. The functional annotation problem, at large, benefits from the expert knowledge of the researcher as he or she implements the use of these networks, and curation of precise functions at subfamily levels becomes largely a manual process done by the expert of the protein family in question.

Figure 1.4 A) Sample SSN of the mandelate racemase subgroup of the enolase superfamily at $10^{-40}$. B) Same network at a more stringent alignment score of $10^{-80}$. Note the clusters segregate into isofunctional clusters (each color represents a different function). All grey clusters lack functional information and are good starting points for experiments [10].
SSNs are useful for a number of reasons. SSNs aid in hypothesis generation, which is one of the chief problems identified by experimentalists with respect to a seemingly overwhelming amount of sequence data. SSNs also help one to visualize functional diversity of a protein family, because often what appear as distinct clusters in an SSN are typically found to be isofunctional. Those clusters of unknown function can be a starting point for experiments; thus, SSNs help to guide experiments. SSNs also enable a researcher to investigate a protein family at a resolution higher than that from the results of a conventional BLAST search and in a manner that is far more interactive (and less unwieldy when dealing with large superfamilies) than standard dendrograms [22].

1.4.1 EFI-EST Example Usage – Enolase Superfamily

As part of the EFI our lab started out investigating the enolase (EN) superfamily to help develop a systematic strategy to discover enzymes on a large scale. The superfamily is a well-characterized, mechanistically diverse set of evolutionarily related enzymes named after its most prominent member. The EN superfamily served as a good test system for the development of an integrated strategy for functional assignment because the function of many of its members are unknown [24, 25].

Work over the past 20 years has revealed that members of the EN superfamily are similar in a number of ways. Crystal structures reveal two domains: an N-terminal α+β capping domain containing the residues required for substrate specificity and a C-terminal modified TIM barrel [(β/α)_{7}β] domain containing the catalytic residues (Figure 1.5). Also, the reactions they
catalyze share a common first step: abstraction of a proton located adjacent (alpha) to a
 carboxylate group to generate a Mg$^{2+}$-stabilized enediolate intermediate. This step is facilitated
 by conserved residues in the active site, including residues used to bind the Mg$^{2+}$ that stabilizes
 the common intermediate. The active site is located at the interface between the two domains
 and is sequestered from solvent by loops that are ordered in the presence of substrate [26].

Figure 1.5 Conserved overall structure for all members of the enolase superfamily; alpha beta
capping domain (specificity residues) and barrel domain (catalytic residues) [26].

The members also differ in a number of ways, including the identities and positions of
the acid/base catalysts (Table 1.1). As a result, members catalyze over 20 unique overall
reactions including cycloisomerization, dehydration, deamination, epimerization, and
racemization. The superfamily can be divided into subgroups based on the differences in active
site residues. Currently, there are seven subgroups: enolase, muconate lactonizing enzyme
(MLE), mandelate racemase (MR), 3-methylaspartate ammonia lyase (MAL), D-mannionate
dehydratase (ManD), D-glucarate dehydratase (GlucD), and meso-galactarate dehydratase (GalrD) [27].

Table 1: Conserved Residues Across the Subgroups

<table>
<thead>
<tr>
<th>Subgroup</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
<th>5th</th>
<th>6th</th>
<th>7th</th>
<th>8th</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enolase</td>
<td>E</td>
<td>D</td>
<td>E</td>
<td>D</td>
<td>K</td>
<td>H</td>
<td>D</td>
</tr>
<tr>
<td>Methylaspartate ammonia lyase</td>
<td>H</td>
<td>D</td>
<td>E</td>
<td>D</td>
<td>K</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Muconate cycloisomerase</td>
<td>KxK</td>
<td>D</td>
<td>E</td>
<td>D</td>
<td>K</td>
<td>-</td>
<td>E/DxD/G</td>
</tr>
<tr>
<td>Mandelate racemase (Kx(K,R,H,D,Y,x))</td>
<td>D</td>
<td>E</td>
<td>E</td>
<td>D</td>
<td>H</td>
<td>E</td>
<td></td>
</tr>
<tr>
<td>Glucarate dehydratase</td>
<td>KxK</td>
<td>D</td>
<td>E</td>
<td>N</td>
<td>D</td>
<td>H</td>
<td>D</td>
</tr>
<tr>
<td>Mannonate dehydratase</td>
<td>RxN</td>
<td>D</td>
<td>E</td>
<td>E</td>
<td>R</td>
<td>H</td>
<td>E</td>
</tr>
<tr>
<td>Galactarate dehydratase</td>
<td>RxY</td>
<td>D</td>
<td>E</td>
<td>H</td>
<td>Y</td>
<td>H</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 1.1 Important residues at the ends of the 2nd through 8th beta strands in the barrel domain of enolase superfamily members. Note the conserved metal binding residues at the 3-5 beta strands, and variation at the other positions. The superfamily is partitioned on the basis of differences at these important positions.

The number of EN superfamily members identified in the nonredundant database now exceeds 50,000. The functions of ~50% of them can be assigned based on direct experimental evidence or inferred for sequences highly similar to those of experimentally verified members. Sequence and structural analyses suggest that uncharacterized members may catalyze reactions other than those currently known [28].

Our lab has studied the extent and limitations of divergent evolution in the enolase superfamily for over 20 years. The focus has been on the mechanistic course of the reactions that its members catalyze and the constellation and alterations in their active site residues that
confer the ability to catalyze those reactions. However, sequence data have opened up a new avenue for enzymologists to use their expertise in a more general fashion in order to contribute solutions to the problem of functional annotation. Thus, there has been a shift in our lab’s focus from determining the mechanistic course (mechanistic enzymology) of a single reaction to identifying the catalytic course of an entire metabolic pathway as an approach to discovering new functions. This latter research has been referred to as enzyme genomics or genomic enzymology [17, 25]. The research is two-pronged and addresses function discovery (tools and approaches to identify function on a large scale) and annotation propagation (high- (SSNs) and low- (alignments, dendograms) resolution defining of family boundaries). The gems of enzyme genomics include discovering novel reactions on novel substrates, old reactions on novel substrates, novel pathways for novel metabolites, and identifying novel activities for protein families with no known functions, to name a few. The allure of the unknown naturally has drawn an emphasis toward the most underexplored areas of biological sequence space in the protein universe (e.g., domains of unknown functions, unknown metabolites).

For historical and practical reasons (e.g., expertise, substrate libraries), this work begins with an investigation into the functional diversity in the enolase superfamily to contribute to the development of the systematic approach proposed by the EFI. Many members of the enolase superfamily accept sugar acids as their substrates: aldonic acids, aldaric acids, and/or uronic acids [27]. Thus, this work starts at the most intractable edges of the enolase superfamily and expands to other superfamilies related by the microbial pathways in which they operate – pathways for the assimilation of uronic and aldaric acids.
1.5 *A. tumefaciens* strain C58

One bacterium thought to be a source of EN superfamily members with novel functions is the soil dwelling plant pathogen and important member of the underexplored plant microbiome, *A. tumefaciens*. Its genome was sequenced in 2001 and encodes 12 EN superfamily members [29]. Because it lives in close proximity to plants, which are primary producers of carbohydrates, we reasoned that *A. tumefaciens* was likely to possess enolase superfamily members of novel function. Furthermore, *A. tumefaciens* is easy to grow in the lab and genetically tractable, so growth experiments, gene expression analyses, and gene knockouts can be used to verify assigned functions.

SSN analysis suggests that eight of the 12 superfamily members are part of the mandelate racemase subgroup of the superfamily, which the majority function to dehydrate sugar acids (Figure 1.6).
Figure 1.6 SSN of putative acid sugar dehydratase subgroups in the enolase superfamily at 10°−80°. Red numbered dots represent protein sequences from A. tumefaciens. Protein sequence labelled 7 is not a member of the mandelate racemase subgroup therefore is not described in this work. The function of 1RVK had already been identified; thus, the remaining seven “unknowns” were part of this investigation.

A small-scale research effort investigating a single member of the superfamily was initiated in 2005 by a former graduate student, but those screening efforts failed to identify any novel functions. As part of a large-scale protein structure initiative, a high resolution structure for another enolase superfamily member from A. tumefaciens was solved in 2003 (Protein Data Bank entry 1RVK), but its function remained unknown (SNF – structure no function). Subsequently, a different graduate student screened that protein against a library of sugar acids...
in order to assign a function to this SNF. She showed that 1RVK was able to dehydrate D-glucarate ($k_{\text{cat}} 0.26 \text{ s}^{-1}, K_M 0.012 \text{ mM}$, $k_{\text{cat}}/K_M 2.3 \times 10^4 \text{ M}^{-1}\text{s}^{-1}$) and L-idarate ($k_{\text{cat}} 0.36 \text{ s}^{-1}, K_M 0.11 \text{ mM}$, $k_{\text{cat}}/K_M 3.8 \times 10^3 \text{ M}^{-1}\text{s}^{-1}$). These represent new activities for members of the mandelate racemase subgroup, and thus confirm the presence of novel enolase superfamily members encoded by the A. tumefaciens genome.

The goal of this project at the time I became involved was to determine the functions of the remaining seven superfamily members hypothesized to dehydrate acid sugars and to incorporate lessons about the process of functional assignment into the EFI’s integrated strategy. The approaches combine the use of SSNs with high throughput screening and in vitro enzymology.

I cloned and purified the remaining seven, and I discovered a fuconate dehydratase (although this function was already known) and a galactarate dehydratase. Although the overall reaction of the galactarate dehydratase was not novel, the active site residues were different from those described for other galactarate dehydratases in the enolase superfamily. The galactarate dehydratase was biochemically characterized in a subsequent study by another graduate student [37], while my attention was drawn to how A. tumefaciens assimilates D-galA because of a discovery by another group on the function of one of the initial targets of this study.

1.5.1 D-GalA Assimilation in A. tumefaciens

D-GalA can be used as a carbon source by many bacteria. A few pathways are known for assimilating D-galA, including the isomerase pathway and the oxidative pathway [30]. A.
*tumefaciens* strain C58 can use D-galA as a sole source of carbon via the oxidative pathway (Figure 1.7).

![Diagram of oxidative assimilation pathway](image)

Figure 1.7 The D-galA oxidative assimilation pathway described in *A. tumefaciens*. The conversion of 1,5-lactone to 1,4-lactone was believed to occur spontaneously prior to the discovery of the enzyme catalyst for the reaction.

Our attention was drawn to this pathway in *A. tumefaciens* strain C58 because it includes a reaction catalyzed by a member of the enolase superfamily that was part of our initial investigation into enolase superfamily members from *A. tumefaciens* but overlooked because our library lacked the true substrate for this enzyme. This enzyme, designated D-galactarolactone cycloisomerase (Gci), catalyzes the ring opening of D-galactaro-1,4-lactone to yield 5-keto-4-deoxy-D-galactarate (KDG) via a β-elimination reaction that is initiated by abstraction of the proton adjacent to the carboxylate group [31].
In this pathway, an NAD$^+$-dependent uronate dehydrogenase oxidizes the β-pyranose form of D-galA to D-galactaro-1,5-lactone (δ-lactone). The assumption had been that the δ-lactone nonenzymatically isomerizes to D-galactaro-1,4-lactone (γ-lactone). Following this isomerization, Gci catalyzes the ring opening reaction of the γ-lactone to produce KDG that undergoes dehydration and decarboxylation to form α-ketoglutarate semialdehyde which is oxidized to α-ketoglutarate, an intermediate in the citric acid cycle [30, 31].

Atu3139 is located in a gene cluster that includes Atu3138 and Atu3140. Based on sequence similarity, the closest functionally characterized homologue of the protein encoded by Atu3140 is a KDG dehydratase/decarboxylase (KdgD). Because KDG is the product of the reaction catalyzed by Gci, the proximity of these genes is not surprising. However, the function of the protein encoded by Atu3138 was unknown.

In this work, SSNs, genomic context, and pathway reconstruction were used in combination with in vitro biochemistry to identify a novel enzyme, D-galactarolactone isomerase (GLI), that catalyzes the isomerization of D-galactaro-1,5-lactone to D-galactaro-1,4-lactone. D-Galactaro-1,4-lactone is then dehydrated by the enolase superfamily member in the gene neighborhood to produce 5-keto-4-deoxy-D-galactarate. GLI is a member of the functionally diverse amidohydrolase superfamily.

### 1.6 SBP Initiated Regulon Reconstruction

One strategy used by the EFI and others to infer an enzyme’s function is to analyze its genomic context because proteins that are co-localized on a chromosome or co-regulated by the same regulator (i.e., part of the same regulon) often function together in a pathway.
Information about what other proteins do in the pathway can be used to infer the function of the “unknown” enzyme.

During the course of trying to identify the function for an enzyme of unknown function, the question of what substrate(s) to test arises. When the EFI began, projects started with trying to answer this question for enzymes in the middle of a potential pathway. However, enzymes that are part of a metabolic pathway are often encoded adjacent to transport systems that transport the initial metabolite into the cell, and an early EFI project was successful in using the solute binding protein of transport systems to identify the initial substrate in a pathway [32]. Armed with the identity of an initial substrate and the general function of the families to which the enzymes in the given gene neighborhood belong, one can then use this information to reconstruct a possible pathway. In reconstructing the pathway one can then make predictions about the possible substrate and reaction catalyzed by the uncharacterized enzymes in the neighborhood.

However, functionally related proteins or groups of proteins may be located in different regions in a genome; how to identify these is a challenge. One solution is to identify a conserved transcription factor binding site (TFBS) shared by proteins or groups of proteins in different locations because identical TFBSs suggests that they are co-regulated. To do this requires a comparative genomic context analysis based on identifying conserved chromosomal clusters and shared TFBSs in the genomes of closely related species [33, 34].

In a large scale thermal shift screening of the solute binding proteins (sbp) of tripartite ATP independent transporters against a metabolite library, a number of targets hit on D-
glucuronate (D-glcA) and/or D-galA (D-galA) providing the opportunity to investigate the assimilation of these carbon sources in other Proteobacteria other than A. tumefaciens [35].

D-GalA and D-glcA are highly abundant sources of carbon. A number of labs are interested in engineering microbes to convert these compounds into energy and chemicals, but those efforts are limited by the lack of enzymes that are soluble in their engineered strains [36]. This work identified two new families of lactone hydrolases, a new family of mutarotases, novel regulators, and novel transporters in work that extends beyond the characterization of strict “enzymes”. By improving our knowledge of pathways for metabolism of D-galA/D-glcA, and identifying new organisms that harbor these pathways, we expand the repertoire of enzymes that could be useful industrially and/or commercially.

There are two known pathways by which microbes harvest the carbon and energy from these compounds: the isomerase pathway and the oxidative pathway [30]. In this work I identified a missing enzyme in the oxidative pathway and a variant of the oxidative pathway.

This work used information from the sbp screen in combination with regulon and pathway reconstruction to predict the function of two families of D-glucaro- and D-galactaro lactone hydrolases from the beta propeller clan, a very large group of evolutionarily related proteins. The clan contains 60 Pfam family members; two subfamilies from two of the 60 Pfam families (PF08450 SMP-30/Gluconolactonase/LRE-like region and PF10282 Lactonase) were predicted to contain these lactonases. Members of the clan all share the same overall structural similarity: twisted beta sheets face-to-face in a propeller fashion. Also a part of this regulon is a novel uronic acid mutarotase, a member of the aldose 1 epimerase Pfam family. In vitro biochemistry was used to test the predictions, and gene expression analysis was used to
determine whether the genes encoding these enzymes are up regulated when the bacterial host is incubated with sugars that bound the sbps: D-galA and D-glcA. Here we identified a novel enzyme, D-galactarolactone hydrolase (Glh), that catalyzes the hydrolysis of D-galactaro-1,5-lactone to meso-galactarate and D-glucaro-1,5-lactone to D-glucarate, as well as a novel mutarotase responsible for interconverting the alpha and beta anomers of D-galA.

1.7 Conclusion

Advances in sequencing technology have generated massive amounts of sequence data. These data undoubtedly hold the potential to change our understanding of biology forever, in a fashion far beyond that of any previous generation. Many major basic discoveries have come from the study of life’s “simplest” creatures such as viruses and bacteria. Therefore, not only will the sequences of these genomes profoundly change the way we understand the role of microbes in the biosphere and in human health, but they also have the potential to be a treasure trove for fundamental discoveries in the biological sciences. However, these data have brought along with it a host of problems yet to be solved, and the scientific community is still at the very early stages of addressing these problems.

Before this work, the ability to generate SSNs required bioinformatics skills beyond those of many experimentalists. Now, anyone can generate SSNs easily on the web. Other complementary strategies also have been tested in this work (e.g., genomic context, pathway reconstruction). In addition to the development of strategies, this research has also contributed a number of basic discoveries in the area of how microbes convert abundant stores of carbon and energy into central metabolic intermediates. This information is useful to those researchers
interested in converting biomass to fuel and value-added chemicals. The work has also contributed to the discovery of several novel enzymatic activities in a number of protein families, a major post genomic era challenge.

The functional annotation field is still young. Perhaps it will benefit from technology like DNA sequencing has (e.g., miniaturization, automation, imaging, improved reaction chemistries), and through an interdisciplinary approach that includes engineering and informatics the rate at which protein sequences are annotated will begin to approach the rate at which genomes are sequenced. The problem can be declared solved when those rates are the same; then the true potential of the information locked within the massive and ever-growing sequence data sets will be realized, thus freeing up the field to move on to the next big revolution.

1.8 References


CHAPTER 2: EFI – ENZYME SIMILARITY TOOL (EFI-EST): A WEB TOOL FOR MAKING PROTEIN SSNS

Reprinted (adapted) with permission from:


2.1 Introduction

The number of sequences in the UniProt Knowledgebase (UniProtKB) is predicted to grow dramatically into the foreseeable future. Over 99% of those sequences have not been manually curated, and over 50% of those have unknown, uncertain, or incorrect functional annotations [1]. To put the information in UniProtKB to practical use, tools to scale up accurate function annotation are needed [2-4]. One approach experimental enzymologists and others find useful in scaling up function annotation is the protein SSN (Figure 2.1B). An SSN provides a quick, easy way to visualize relationships within groups of sequences. In its simplest form, sequences are represented by shapes (nodes), and connectivity between sequences is
represented as lines (edges) connecting the shapes. Connectivity is based on sequence similarity, and often similarity is determined using BLAST [5].

Figure 2.1 A) Rooted phylogenetic tree created with ClustalW grouping sequences by similarity. B) SSN nodes cluster by sequence similarity. Edges generated from all by all comparison aid in clustering tree clades. Results are qualitatively similar [6].

Common tools for exploring relationships between protein sequences include sequence alignments to observe conserved residues, and dendograms to enable inferences about evolutionary relatedness. For those interested in functional discovery and annotation propagation on a large scale and less interested in investigating evolutionary relatedness, a more robust, albeit lower resolution, approach are SSNs. SSNs not only allow experimentalists to interrogate and traverse the large sequence databases, they also aggregate information about each of the sequences in one place and enable a researcher to interact with the dataset.
through locally installed software. Information is aggregated as node attributes and includes host organism, whether or not the protein has been functionally characterized, whether or not a structure is available, etc. Interaction options include the ability to link out to external databases, label nodes, and filter the network by node connections in order to view at more stringent alignment scores. This enables clustering by sequence similarity, thus allowing visualizing relationships between sequence similarity and function for a particular protein family. For large datasets, they are currently less computationally intensive to construct and make it easier to visualize relationships between sequence and function than dendograms; they are qualitatively comparable to dendograms (Figure 2.1) [5, 6].

SSNs can be used to identify relationships between sequence similarity and function in protein families. They allow users to segregate groups of proteins into putative isofunctional clusters, and to easily identify those clusters without known functions, both good starting points for functional exploration. In fact, recent publications used SSNs not only to define specificity boundaries and investigate diversity within protein families but also to assign functions to “unknown” proteins with structures and to provide clues about the functions of operon encoded proteins [7, 8]. In those cases, the authors leveraged a combination of expert knowledge and SSNs; thus, research programs with expertise in a particular enzyme (or family/families) will benefit fast by considering family context provided by SSNs. Likewise, the function annotation problem benefits from the scaled up application of expert knowledge.

The gold standard for highly curated protein SSNs for a limited number of enzyme superfamilies is the Structure Function Linkage Database (SFLD) [9]. The SFLD also provides networks for a number of functionally diverse superfamilies. SFLD programmers also developed
a program for generating networks called Pythoscape, but, although Pythoscape is a powerful, extensible framework for generating SSNs for large groups of sequences [10], it requires basic computing skills that many biologists do not have. So in order to make it easier for biologists to obtain SSNs for their protein family of interest, and to make networks available any protein family, a more user-friendly method was needed.

To this end, the EFI developed a public web tool, EFI-Enzyme Search Tool (EFI-EST) [6]. EFI-EST enables users to make networks quickly and easily. In the past seven months, 362 jobs finished with an average run time of 5.5 hours requiring only modest input from the user. This chapter provides a description of the tool, and a tutorial, and an example use.

2.2 Description and Implementation

2.2.1 Database Construction

UniProtKB is the primary source for protein sequences and sequence related information. When this was written, EFI-EST was using the UniProt database release from 2015_02, and that database is updated every 4 weeks [11]. SSNs include sequence information from UniProtKB as node attributes, detailed in Table 2.1. Information is obtained from three locations: UniProtKB (Swiss-Prot and TrEMBL flat files downloaded from URL and merged); Genomes OnLine Database (URL) [12]; and EFI-specific from an in-house list of organisms for which the EFI has genomic DNA and an in-house list of identifiers for EFI targets.
Table 2.1 Annotations included as node attributes in EFI-EST networks.

<table>
<thead>
<tr>
<th>Node attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACC</td>
<td>UniProt accession(s)</td>
</tr>
<tr>
<td>Uniprot_ID</td>
<td>UniProt ID(s)</td>
</tr>
<tr>
<td>GN</td>
<td>Gene name(s)</td>
</tr>
<tr>
<td>GI</td>
<td>GI numbers</td>
</tr>
<tr>
<td>STATUS</td>
<td>Reviewed — manually annotated, in Swiss-Prot; unreviewed — automatically annotated, in TrEMBL</td>
</tr>
<tr>
<td>Description</td>
<td>Protein name(s)/annotation(s) in UniProtKB</td>
</tr>
<tr>
<td>SwissProt_Description</td>
<td>Protein name(s)/annotation(s) in UniProtKB for SwissProt reviewed entries</td>
</tr>
<tr>
<td>IPRO</td>
<td>InterPro family(ies)</td>
</tr>
<tr>
<td>PFAM</td>
<td>Pfam family(ies)</td>
</tr>
<tr>
<td>PDB</td>
<td>Protein Data Bank entry</td>
</tr>
<tr>
<td>CAZY</td>
<td>Carbohydrate-Active enZymes (CAZY) family name(s)</td>
</tr>
<tr>
<td>EC</td>
<td>EC number(s)</td>
</tr>
<tr>
<td>GO</td>
<td>Gene Ontology classification(s)</td>
</tr>
<tr>
<td>Sequence_Length</td>
<td>Number(s) of amino acid residues</td>
</tr>
<tr>
<td>Domain</td>
<td>Domain of life to which the organism(s) belong(s)</td>
</tr>
<tr>
<td>PHYLUM</td>
<td>Phylogenetic phylum/phyla of the organism(s)</td>
</tr>
<tr>
<td>CLASS</td>
<td>Phylogenetic class(es) of the organism(s)</td>
</tr>
<tr>
<td>ORDER</td>
<td>Phylogenetic order(s) of the organism(s)</td>
</tr>
<tr>
<td>FAMILY</td>
<td>Phylogenetic family(ies) of the organism(s)</td>
</tr>
<tr>
<td>GENUS</td>
<td>Phylogenetic genus/genera of the organism(s)</td>
</tr>
<tr>
<td>SPECIES</td>
<td>Phylogenetic species of the organism(s)</td>
</tr>
<tr>
<td>Organism</td>
<td>Organism genus/genera and species</td>
</tr>
<tr>
<td>Taxonomy_ID</td>
<td>NCBI taxonomy identifier(s)</td>
</tr>
<tr>
<td>HMP_Body_Site</td>
<td>Location(s) of organism(s) in/on the body, if human microbiome organism</td>
</tr>
<tr>
<td>HMP_Oxygen</td>
<td>Oxygen requirement(s), if human microbiome organism</td>
</tr>
<tr>
<td>EFI_ID</td>
<td>Enzyme Function Initiative database ID(s)</td>
</tr>
<tr>
<td>GDNA</td>
<td>Availability of gDNA(s) at EFI Protein Core</td>
</tr>
<tr>
<td>Shared name</td>
<td>Full network — UniProt accession; Rep Node network — UniProt accession for the longest sequence in the representative node</td>
</tr>
<tr>
<td>Name</td>
<td>UniProt accession for the longest sequence in the representative node</td>
</tr>
<tr>
<td>Cluster size$^1$</td>
<td>Number of proteins represented by the representative node</td>
</tr>
</tbody>
</table>

$^1$ Representative node SSNs only.

Pfam [13] and InterPro [14] family membership assignments for UniProt entries that belong to families come from InterPro’s match_complete.xml file available from InterPro’s website. When this was written, EFI-EST was using release 47.0. Currently there are over 14,300 Pfam families.
2.2.2 Stage 1: Dataset Generation

EFI-EST runs in two stages; the first stage has five steps. In step one, a user has two options for defining a set of sequences. A user may either BLAST [15] a protein sequence against a local version of the UniProt protein sequence database [11] or specify a Pfam and/or InterPro family or a combination thereof. The family to which a sequence belongs can be identified at the InterPro website. In the first option, 5000 most similar sequences are retrieved. In the second option, all sequences corresponding to the specified Pfam or InterPro family(ies) are retrieved from an in-house SQLite database. The program only works for protein sequences; sequence sets up to 100,000 are allowed.

In step two, the pairwise BLAST is performed. All BLAST steps use NCBI BLAST 2.2.26. BLAST results are returned for all pairwise comparisons with an E-value less than 1E-5 (i.e., pairwise comparisons with E-values greater than 0.00001 are not performed). E-values are used as edge weights and are included in the networks edge attributes. Percent identity and alignment length are also included as edge attributes.

In step three, the BLAST output is compiled and filtered to remove self (sequence A → sequence A) and reciprocal (sequence A → sequence B, sequence B → sequence A) comparisons. Although the values for reciprocal comparisons may not be equal, for our purposes we keep the occurrence with the smallest E-value. In step four, annotations are retrieved from the SQLite database. In step five, graphs (vide infra) are plotted.

2.2.3 Stage 2: Analysis and Network Generation
In stage two, the user inputs an E-value threshold to filter BLAST results; edges with E-values greater than the threshold will be excluded from the networks (how to determine starting alignment score is described below). Next, the program uses the filtered BLAST results and the annotation file to generate a full network. Representative node (repnode) networks are generated by clustering sequences by percent identity with the program CD-HIT [16]. A limitation to opening a network has to do with the number of edges contained in a full network and the memory of a user’s computer. Therefore representative nodes allow one to open larger networks by condensing sequences of specified percent identity into single nodes, thus reducing the total number of nodes and thus edges in a network. The repnode network file is then written from the filtered BLAST results, the annotation file, and the CD-HIT output. Repnode networks are described below. Finally, the number of edges and nodes as well as the file size for each network is also tallied. The format of all network files is eXtensible Graph Markup and Modeling Language (XGMML). This file type can be imported and displayed by Cytoscape [17].

2.2.4 Graph Interpretation

EFI-EST provides four graphs that guide the selection of an E-value cutoff to produce an informative SSN [5]. The “number_of_edges” histogram displays the relationship between the number of edges in a full network and the E-value. At high E-values (lower percent identity) the number of edges is high. As the E-values decrease (higher percent identity) the number of edges will decrease. For large datasets, this graph helps to determine an E-value cutoff that will
produce a full network with few enough edges to be rendered on most computers (e.g. ~500,000 edges for 4GB RAM).

The “length_histogram” displays the number of sequences for each sequence length in the dataset. For EFI-EST jobs initiated with a single protein sequence, this graph typically displays a tight distribution of sequence lengths around the length of the query sequence. For jobs initiated with a Pfam or InterPro family, the “length_histogram” is valuable for determining the sequence length distribution of that family. For instance, it will inform the user about the presence of fragments or fusions that could be excluded from analysis. Additionally, the graph will signal the presence of sequences containing domain repeats. Users should inspect this histogram to determine if upper and lower sequence length cutoffs are needed.

The “alignment_length” quartile plot displays the relationship between the alignment length and E-value. At larger E-values, partial alignments can occur over small portions of the sequences. As the E-value decreases, alignments occur over longer portions of the sequences. Thus, the “alignment_length” quartile plot will appear to begin at zero on the y-axis and increase asymptotically toward the longest sequence length in the dataset. We typically select a low enough E-value cutoff to exclude the presence of edges due to partial alignments.

The “percent_identity” quartile plot displays the relationship between percent identity and E-value. The median percent identity generally increases linearly with decreasing E-value. This plot provides the more familiar metric, percent identity, for designating an initial E-value cutoff. We find that an E-value cutoff equivalent to 30-40% sequence identity is a good starting point; it is low enough to prevent over-fractionating orthologous clusters yet high enough to remove uninformative edges. Networks can be filtered further in Cytoscape, but only to more
stringent (lower) E-values. To generate networks at higher E-values, repeat Stage 2. Depending on the size of the family, full networks may be too large to open. Representative node networks address this problem by grouping sequences into a single node on the basis of percent identity.

2.2.5 Downloading and Visualizing Networks

After an E-value cutoff and optional length restrictions are specified, 14 network files are generated that represent the full network as well as various repnode networks. In a full network, each sequence is present, and edges are drawn between sequences that share a BLAST E-value less than the user-specified cutoff. In representative node (repnode) networks sequences are grouped into a single node on the basis of percent identity. For repnode networks, the program CD-HIT is used to group sequences that share the designated percent identity (e.g. 40, 50, 60) so that in the network they appear as a single node. Consolidating sequences into representative nodes primarily affects the density of edges and nodes within clusters but does not affect the number and general shape of clusters. Repnode networks are useful for observing a protein family in cases where a full network at the same E-value cutoff would be too large to open or too slow to work with effectively. The repnode networks have significantly fewer edges yet yield equivalent information.

2.3 EFI-EST Example Usage – Generation, Visualization, and Analysis of the Mandelate Racemase Subgroup of the Enolase Superfamily from A. tumefaciens

The mandelate racemase subgroup (a subset of Pfam families PF01188 MR_MLE, PF02746 MR_MLE_N, PF13378 MR_MLE_C) of the enolase superfamily is used to illustrate the
use of both Options A and B. The mandelate racemase subgroup is functionally diverse with 13 characterized reactions (Figure 2.2). Despite different substrates/products, reactions, and mechanisms, the members of this superfamily have a conserved structure with two domains: an N-terminal α+β capping domain containing the residues required for substrate specificity and a C terminal modified TIM (β/α)$_7$β barrel domain containing the catalytic residues. The active site is located at the interface between the two domains. The mechanisms of the reactions catalyzed by members of the mandelate racemase subgroup have been investigated in our lab over the past 20 years [18].
Figure 2.2 Reactions found within the seven subgroups (blue boxes) of the enolase superfamily. This work is focused on the mandelate racemase subgroup, upper right-hand blue box. There are 13 reactions (families) in this subgroup.
The strategies used here can be applied universally to other sequence datasets (closest neighbors for Option A and complete families for Option B). Option B will be described to provide an overview of structure/function space in members of the enolase superfamily, with emphasis on the mandelate racemase subgroup. Option A can be used to enable more focused examinations of structure/function space in this subgroup.

2.3.1 Option B: Identification of the Pfam Entry

The structurally and mechanistically characterized mandelate racemase subgroup member from *A. tumefaciens* strain C58 (Atu3453; UniProt accession Q7CSI0; Figure 2.3 is used to demonstrate the identification of the Pfam/InterPro entries for generating the SSNs. Five InterPro entries are identified using InterProScan.
Figure 2.3 InterProScan5 output using UniProt accession Q7CSIO as the query protein sequence.

Two structure-based InterPro domains are identified, IPR029017 (Enolase N-terminal domain-like; 35,115 sequences) and IPR029065 (Enolase C-terminal domain-like; 37,876 sequences), that are defined by the CATH/Gene3D and SCOP/Superfamily databases.

One sequence-based InterPro domain, IPR001354 (Mandelate racemase/muconate lactonizing enzyme/methylaspartate ammonia-lyase; 22,317 sequences), is defined by entries
from Panther (PTHR13794), another, IPR013341 is defined by Pfam only, and IPR013342 is defined by Pfam (PF01188) and SMART (SM00922). Not all InterPro entries are defined by multiple databases; however, when they are, the number of sequences in the InterPro entry likely will be larger than the numbers identified by the individual database entries because different InterPro member databases use different bioinformatics approaches to classify sequences, thus leveraging the expertise of multiple groups. InterProScan also identifies a Pfam family (PF13378) and a Panther family (PTHR13794) that are not incorporated into an InterPro entry. The availability of multiple InterPro entries for a family allows the EFI-EST users to be inclusive in identifying sequences for generating SSNs.

2.3.2 EFI-EST: Start Page

To generate the SSN for the subgroup as defined by the three Pfam families (PF02746, PF01188, PF13378) identified using InterPro scan, first the user enters the Pfam entry identifiers in the Option B box on the EFI-EST Start Page (Figure 2.4) and an e-mail address, and next clicks on the GO button to collect the sequences, perform the all-by-all BLAST, and generate the graphs to inform selection of the alignment score lower limit (E-value upper limit/percent identity lower limit) for outputting the SSN files. Entries that occur in more than 1 Pfam family will only be included once.
Using the InterPro 51.0/UniProt 2015_02 releases, EFI-EST collects 56,199 sequences. However, the user could have entered IPR013341 and IPR013342, that are derived from PF02746 and PR01188. Note that depending on the definitions of a family by the databases, the number of sequences may be much larger for InterPro entries defined by multiple databases.
(an InterPro entry) than that defined by a single database (e.g., Pfam). The subgroup is described by the five InterPro entries noted above, so the user could have entered all three identifiers in the Option B box.

2.3.3 Analyzing the BLAST Dataset: Specifying an Alignment Score Lower Limit for the SSNs

The Data Set Completed page (Figure 2.5) provides links for displaying and downloading to the user’s desktop the four graphs used to select 1) the alignment score lower limit (E-value upper limit) for generating the SSN files (required), and 2) minimum and/or maximum length limits to exclude fragments and/or multidomain proteins (optional). The user should download and save all four of the graphs for future reference. Interpretation of the graphs for the three Pfam families is provided in this section.
Figure 2.5 EFI-EST Dataset Completed Page.
The Length Histogram is displayed in Figure 2.6. The majority of the sequences have lengths between 300-450 residues (the minimum length for a member of the enolase superfamily is ~300 residues); additional residues are N- and C-terminal extensions as well as internal loops within the capping domain and/or at the end of the beta strands beyond the 300 residue minimum. The length histogram reveals the presence of shorter sequences (≤ 300 residues; fragments) and very few longer sequences greater than 450. The presence of longer sequences indicates fusions or potential gene calling errors.

![Figure 2.6 Length Histogram.](image)

The Number of Edges Histogram is displayed in Figure 2.7. This graph reveals that the majority of the edges are associated with relatively small alignment scores, with few at very large alignment scores. This subgroup is populated by multiple divergent families.
The Alignment Length Quartile Plot is displayed in Figure 2.8. At the smallest alignment scores, the alignment length is a fraction of the minimum length for a protein in the enolase superfamily; the alignment length then increases to ~300 residues and remains consistent across all alignment scores. In the length histogram (vide supra), the alignment scores greater than 15 correspond to the sequences between roughly 300-400 residues in length. Over this range, the alignment score (E-value) is calculated over the full length of the protein, so these E-values are reliable measures of the pair-wise sequence similarity. The Alignment Length Quartile Plot allows the user to select an alignment score range that corresponds to alignment of full-length sequences thereby enabling proper interpretation of the subsequent Percent Identity Quartile Plot.
Figure 2.8 Alignment Length Quartile Plot

The Percent Identity Quartile Plot is displayed in Figure 2.9. On average the percent identity increases monotonically toward 100% as the alignment score increases. An alignment score corresponding to 35% sequence identity usually is a good choice for generating the initial SSN. Simultaneous visualization and analysis of the resulting network in the context of the node attributes should allow the user to increase, if necessary, the alignment score lower limit to achieve segregated isofunctional clusters.
From the Percent Identity Quartile Plot, 35% sequence identity corresponds to an alignment score of 60. This value is entered in the Choose E-value for output field on the Data Set Completed Page (Figure 2.5). For this example, the minimum and maximum length fields are left blank to include all sequences. In the Provide Network Name field, the user provides a name for the network when it is opened in Cytoscape, e.g., MR_subgroup_E-60. Finally, the user starts generation of the SSN XGMML files.

2.3.4 Downloading the Full and Representative Node SSNs

The Download Network Files page (Figure 2.10) displays the total number of sequences used in the analysis and provides links for downloading the SSNs as well as summary of the number of nodes and edges in each SSN file. For this example, a total of 56,199 sequences was identified in these Pfam families and used to calculate the edges with alignment scores > 5.
Figure 2.10 Download Network Files Page.
Note that the 100% repnode network has over 5 million edges and is 1.4 gigabytes in size (Figure 2.10). This network would be too large to open on many computers; thus, to view this network requires a lower percentage repnode network. A 60% repnode and an 80% repnode are shown in Figure 2.11. The 60% repnode contains 2,131 nodes and 25,100 edges while the 80% repnode contains 5,028 nodes and 241,800 edges. Both are accessible with a 4 GB computer.
Figure 2.11 A) 60% repnode. B) 80% repnode. Both are at an alignment score of 60. Note the 60% repnode has fewer nodes and thus fewer edges. This approach allows for opening larger networks and those that are suitable for a user’s computer based on its RAM.

2.3.5 Visualizing and Analyzing the SSNs with Cytoscape
This example starts with an alignment score of 60 (Figure 2.11B). Figure 2.12A is the same network as 2.11B but those nodes representing sequences from *A. tumefaciens* strain C58 are colored red. The starred cluster and the boxed cluster were isolated, and new network was made just from those clusters (Figure 2.12B).

Figure 2.12 A) Protein SSN of Pfam families PF02746, PF01188, and PF13378. *A. tumefaciens* proteins are colored red. The starred cluster was the original query sequence used to determine what Pfam families to use. The boxed cluster contains the seven mandelate racemase subgroup members from *A. tumefaciens*. The three other red nodes are members of other subgroups within the enolase superfamily and are not a part of this study. B) The child network was made using just the boxed and starred clusters.
The alignment score required to separate isofunctional clusters is 80, thus filtering of the alignment score was performed in Cytoscape. The conclusion that the clusters are isofunctional is based on the annotations in Swiss-Prot.

### 2.3.6 Using Swiss-Prot Annotations

Cytoscape can be used to select sequences in the SSNs that contain specific information in the node attributes. For the purpose of assessing when an SSN is segregated into isofunctional clusters, the pertinent node attributes are STATUS (Swiss-Prot reviewed or unreviewed) and Description (protein name/annotation) both of which are obtained from the UniProtKB.

In the 80% rep node network with a minimum alignment score of 80 (Figure 2.13):

1) One metanode has a Swiss-Prot reviewed status (STATUS node attribute) in UniProt and a Description node attribute that includes the phrase, D-arabinonate dehydratase; the cluster is labeled. This suggests that this cluster is populated by D-arabinonate dehydratases.

2) One metanode has a reviewed status in UniProt and a Description node attribute that includes the phrase, D-arabinonate dehydratase; the cluster is labeled. This suggests that this cluster is populated by D-arabinonate dehydratases.

3) Twenty-five metanodes share a reviewed status in UniProt and a description that includes the phrase, D-galactonate dehydratase; these are located in two clusters. These matching shared attributes suggest that these clusters are both populated by D-galactonate dehydratase family.
4) Seven metanodes share a reviewed status in UniProt and a description that includes the phrases, D-mannonate dehydratase. These matching shared attributes suggest that the cluster is populated by the mannonate dehydratases. Mannonate dehydratases are members of a different subgroup in the enolase superfamily but share a Pfam family with members of the mandelate racemase subgroup.

5) One metanode has a reviewed status in UniProt and a Description node attribute that includes the phrase, D-xylonate dehydratase; the cluster is labeled. This suggests that this cluster is populated by D-xylonate dehydratases.

6) One metanode has a reviewed status in UniProt and a Description node attribute that includes the phrase, L-fuconate dehydratase; the cluster is labeled. This suggests that this cluster is populated by L-fuconate dehydratases.

7) One metanode has a reviewed status in UniProt and a Description node attribute that includes the phrase, L-galactonate dehydratase; the cluster is labeled. This suggests that this cluster is populated by L-galactonate dehydratases.

8) One metanode has a reviewed status in UniProt and a Description node attribute that includes the phrase, L-talarate/galactarate dehydratase; the cluster is labeled. This suggests that this cluster is populated by L-talarate/galactarate dehydratases.

9) Five metanodes have a reviewed status in UniProt and a Description node attribute that includes the phrase, L-rhamnonate dehydratase; the cluster is labeled. This suggests that this cluster is populated by L-rhamnonate dehydratases.

Although there are 13 families in the mandelate racemase subgroup (Figure 2.2), only eight (mannonate dehydratases are a different subgroup) currently with a reviewed status are
depicted in this network. The unknowns from *A. tumefaciens* (Figure 2.13B) were selected for further investigation.

![Diagram](image)

**Figure 2.13** A) 80% repnode at an alignment score of 80. At this alignment score functions segregate into isofunctional clusters. Pink nodes represent Swiss-Prot (reviewed) status. Each cluster containing reviewed sequences is labeled with the substrate for those enzymes. B) The same network labeled with *A. tumefaciens* targets in red. The starred cluster contains the original query sequence. The arrow indicated a function for an *A. tumefaciens* protein that was identified by another research group during the course of this work. Note that five of the remaining unknowns (red nodes) are in clusters with no known function (no pink nodes).
2.4 Experimental Testing Materials and Methods

2.4.1 Target Cloning

DNA sequences were amplified via the polymerase chain reaction (PCR) from genomic DNA isolated from *A. tumefaciens* strain C58 using Platinum Pfx DNA Polymerase (Invitrogen). The PCR mixture contained 5 μL of 10X Pfx amplification buffer, 0.3 mM dNTPs, 1 mM MgSO₄, primers at 0.3 μM each (Table 2.2), 1 unit of Pfx DNA polymerase, and 50 ng of genomic DNA in a total volume of 50 μL. The amplification was performed using a PTC-200 gradient cycler (MJ Research) with the following cycling profile: 94 °C for 5 min; followed by 35 cycles of 94 °C for 15 s, 60 °C for 30 s, and 68 °C for 1 min and 30 s; followed by a final extension at 68 °C for 10 min. The amplified genes were inserted into pET-15b (N-terminal His-tag) and pET-17b (tagless) vectors using conventional cloning methods (PCR amplification followed by cloning into multiple cloning restriction sites in the vectors). Targets were heterologously expressed by growing 2-12 L of *Escherichia coli* strain BL21(DE3) cells in LB broth (supplemented with 100 μg/mL ampicillin) at 37 °C (30 °C for Atu3139) while the sample was shaken at 220 rpm for 24 h (48 h for Atu3139). The cultures were not induced with IPTG.
<table>
<thead>
<tr>
<th>UniProt AC</th>
<th>Locus Tag</th>
<th>Cloning Primer Sequences</th>
</tr>
</thead>
</table>
| Q7D1T6     | Atu0270   | Agrtu_20_for_NdelCTCTGGAGGGAGACCATATGAAAAATCACAAAAAATCG  
|            |           | Agrtu_20_rev_BamHI GCAACCTGGGATCCTCATGAGACGCACGCGCATCC  |
| A9CJ44     | Atu1406   | Agrtu_32_for_Ndel GCGGGAGGGAGCATATGAAAATTACGTCTG  
|            |           | Agrtu_32_rev_BamHI CTGATTCGCTCATGAGGATCCTCACAGCCTGTG  |
| Q7CW73     | Atu2811   | Agrtu_92_for_Ndel CGCGAGTAACCAGCATATGACAAAAATTACAGATC  
|            |           | Agrtu_92_rev_BamHI GCGCTCGATCGATCCTCCTCACTCTCGAAAGTGTAAG  |
| A9CEQ8     | Atu3139   | Atu3139 forward GAAAAACGGGAGGGACATATGAAAAATCACGGCGGTGC  
|            |           | Atu3139 reverse CTCAGTTTTTCCTCGAGTTCGCTCATGGGTCTGGCATC  |
| Q7CU39     | Atu4120   | Agrtu_55_for_Ndel CGAGGAATGACCCGATATGAAAAATAACTGCTGTGC  
|            |           | Agrtu_55_for_BamHI CGTCGCAAGCGGATTCATTACGCCCTTGCG  |
| A9CG74     | Atu4196   | Agrtu_86_for_Ndel CATGAGGAAGACTGACATATGAAAAATCGATCGCATGC  
|            |           | Agrtu_86_rev_XhoI CGATGAAGCTCGAGTCGAGGCAGCATAGAAACC  |
| A9CL63     | Atu5458   | Agrtu_for_Ndel_2NQL CTTGTTGCCGCCGGAGGCCCATATGA  
|            |           | Agrtu_rev_BamHI_2NQL GTTTCCCTTCCTCCAAAAAGGATTCAGATACCGTTCAATC  |

Table 2.2 Cloning primers.

2.4.2 Target Expression and Purification
Cells were harvested by centrifugation (5000 x g and 4 °C) and resuspended in 30–40 mL of low-salt buffer [20 mM Tris-HCl (pH 7.9) and 5 mM MgCl₂]. Cells were lysed by sonication and the lysate pelleted by centrifugation (31000 x g and 4 °C) to remove cell debris. The pH of all buffers were adjusted using either HCl or NaOH.

Targets were purified using standard methods either by Ni²⁺-affinity or ion exchange chromatography depending on the level of expression and degree of solubility. For tagless purification the supernatant was loaded onto a 125 mL DEAE column equilibrated with 1250 mL of low-salt buffer. The column was washed with 800 mL of low-salt buffer, and the protein was eluted with a linear 1800 mL gradient of 0 to 50% high-salt buffer [1 M NaCl, 20 mM Tris-HCl (pH 7.9), and 5 mM MgCl₂] followed by 300 mL of 100% high-salt buffer. The purity was confirmed by sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS–PAGE). Fractions containing the protein of the appropriate size were pooled and loaded onto a 20 mL Q-Sepharose column equilibrated with 200 mL of low-salt buffer. The column was washed with 100 mL of high-salt buffer, and the protein was eluted with a linear 700 mL gradient of 100 to 0% high-salt buffer. The purity was checked by SDS–PAGE.

For His-tagged purification, the clarified supernatant containing the His-tagged protein was then loaded onto a chelating Sepharose Fast Flow column charged with nickel and equilibrated with 10 column volumes of binding buffer. The column was washed with 5 column volumes of 15% elution buffer (0.5 M imidazole, 20 mM Tris-HCl pH 7.9, 0.2 M NaCl, and 5 mM MgCl₂) containing 85% binding buffer and eluted over a linear gradient of 15 to 100% elution buffer. Purity was checked by SDS-PAGE gel electrophoresis. Fractions containing protein were then pooled and dialyzed against a solution of 20 mM Tris-HCl pH 7.9 and 5 mM MgCl₂. Protein
was flash frozen drop wise into liquid nitrogen, and the beads were transferred to a cryogenic vial and stored at -80 °C. (Figure 2.14)

Figure 2.14 SDS PAGE gel of seven targets purified in this study. Protein in lanes 3 and 5 were not part of this study.

2.4.3 High Throughput Screen for Enzyme Activity

Proteins were screened against a library of 77 mono- and diacid sugars (Figure 2.15) as described previously in an absorbance assay using semicarbazide that reacts with the carbonyl carbon of the expected dehydration product to form a semicarbazone ($\varepsilon_{250} = 10200 \, \text{M}^{-1} \text{cm}^{-1}$) which is detectable at 250 nm [19]. Briefly, 50 μL reactions were performed in 96-plates that were UV transparent. The wells contained 20 mM Na-HEPES (pH 7.9), 5 mM MgCl₂, 1 mM substrate, and 1 μM enzyme. Reaction mixtures were incubated for 16 h at 30 C after which 250 μL of 1% sodium acetate/1% semicarbazide mixture (semicarbazide solution) was added.
Plates were incubated for 1 h and then the absorbance was read at 250 nM using a Tecan plate reader.

Figure 2.15 77 acid sugars in the screen.

2.4.4 Enzyme Activity Verification by NMR

Reactions showing complete turnover of substrate overnight were analyzed by 1D $^1$H-NMR for the presence of dehydration product. Briefly, a mixture of enzyme, substrate, MgCl$_2$, and buffer (pH 7.9) in H$_2$O at 30°C was incubated overnight. Next, the reaction mixture was lyophilized, the residue was dissolved in D$_2$O, and spectra of reaction mixtures with and without enzyme were recorded by NMR. I analyzed the spectra for loss of peaks associated with the proton adjacent to the carboxylate group of the substrate as well as for gain of peaks between 2.0 and 2.5 ppm, indicative of the presence of the dehydration product.
2.4.5 Determination of Kinetic Constants

Kinetic constants were determined using the end-point semicarbazide assay described above by varying the substrate concentration. Rates of product formation over time were derived from a linear fit of the quenched time points in Microsoft Excel, and kinetic constants were obtained by plotting rate as a function of substrate concentration using GraphPad Prism.

2.5 Results and Conclusion

SSNs have a proven track record of being used successfully for large scale function annotation (Table 2.3) [5, 6]. Experimental enzymologists and protein chemists are well poised to gain from their use. Large-scale function annotation also will benefit by harnessing expert knowledge in conjunction with the use of SSNs.

<table>
<thead>
<tr>
<th>Discovering novel chemistry within a superfamily</th>
<th>Defining substrate specificity</th>
<th>Coupling SSNs with structural insight</th>
<th>Coupling SSNs with genome context</th>
</tr>
</thead>
<tbody>
<tr>
<td>PMID 24684232[13]</td>
<td>PMID 23214453[33]</td>
<td>PMID 21222452[34]</td>
<td>PMID 25540822[16]</td>
</tr>
<tr>
<td>PMID 22069326[40]</td>
<td>PMID 24802635[41]</td>
<td>PMID 23493566[42]</td>
<td>PMID 25129028[14]</td>
</tr>
<tr>
<td>PMID 24401129[46]</td>
<td>PMID 25363770[15]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PMID 24947666[48]</td>
<td>PMID 24697646[51]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PMID 25151136[49]</td>
<td>PMID 24697329[52]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2.3 Publications using SSNs to explore sequence-function space.
SSNs enable the ability to transfer function annotations to closely related sequences reliably and at the family scale. To make sequence analysis tools for large datasets available to the scientific community, EFI-EST provides unrestricted network generation either via web tool or as a stand-alone program not outlined here. The stand-alone program offers additional features including the ability to generate large networks, regenerate networks, access the raw data files, etc. SSNs offer a unique opportunity comparable to going beyond the conventional BLASTing against a database and allow a researcher to visualize those results in a high throughput fashion.

An SSN of Pfam families, PF02746, PF01188, PF13378, was used to identify seven putative sugar acid dehydratases from A. tumefaciens. Initially six of them did not belong to a known family, however the function of 1 (Atu3139) was identified by another research group during the course of this work. Thus prior to the discovery of an activity for Atu3139, all seven putative acid sugar dehydratases were cloned, and the proteins were expressed and purified to homogeneity.

The results of the overnight screening assay are shown in Table 2.4. Two enzymes, Atu4196 and Atu2811 (1 μM) showed near complete turnover of 1 mM substrate overnight at 30°C. Atu3139, Atu0270, Atu4120, and Atu1406 showed some activity but none of them demonstrated complete conversion on any substrate overnight. Atu5458 was not active on any of the sugars in the screen.
<table>
<thead>
<tr>
<th>Locus Tag</th>
<th>Compound</th>
<th>Percent Turnover</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atu4196</td>
<td><em>meso</em>-Galactarate</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>D-Galacturonate</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>L-Rhamnonate</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>L-Galacturonate</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>L-Altrarate/L-talarate</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>L-Malate</td>
<td>5</td>
</tr>
<tr>
<td>Atu3139</td>
<td>D-Mannarate</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>D-Idarate</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>D-Altrarate/D-talarate</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>L-Malate</td>
<td>4</td>
</tr>
<tr>
<td>Atu0270</td>
<td>L-Idarate</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>D-Glucarate</td>
<td>21</td>
</tr>
<tr>
<td>Atu4120</td>
<td>D-Allonate</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>L-Fuconate</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>6-Deoxy-L-talonate</td>
<td>19</td>
</tr>
<tr>
<td>Atu5458</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>Atu1406</td>
<td>D-Gluconate</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>6-Deoxy-L-talonate</td>
<td>4</td>
</tr>
<tr>
<td>Atu2811</td>
<td>6-Deoxy-L-talonate</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td>L-Galactonate</td>
<td>89</td>
</tr>
<tr>
<td></td>
<td>L-Talonate</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>D-Ribonate</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>L-Fuconate</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td>D-Arabinonate</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>D-Arabinarate</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>D-Galacturonate</td>
<td>6</td>
</tr>
</tbody>
</table>

Table 2.4 Results of the screening assay. Percent turnover reflects the amount of substrate converted to product overnight at 30°C. The concentration of enzyme was 1 μM, and the concentration of substrate was 1 mM.
Atu4196 and Atu2811 were chosen for further study because of the high turnover percentages on several substrates. Those substrates that were turned over greater than or equal to 75% were analyzed by 1D \(^1\)H-NMR for the presence of dehydration product. The spectra were analyzed for loss of peaks associated with the proton adjacent to the carboxylate group of the substrate as well as for gain of peaks between 2.0 and 2.5 ppm, indicative of the presence of the dehydration product. Figure 2.16 shows sample data for confirmation of activity by NMR [20]. The conversion of 6-deoxy-L-talonate, L-galactonate, L-talonate, and D-ribonate were also confirmed by NMR to be dehydrated to the expected keto deoxy products (data not shown). These activities match activities for a previously discovered L-fuconate dehydratase [20]. This is the first fuconate dehydratase discovered in *A. tumefaciens*, although it is not a new function in the mandelate racemase subgroup. Atu4196’s conversion of meso-galactarate to a keto deoxy substrate also was confirmed by NMR.
Because Atu2811 was part of a known family of L-fuconate dehydratases we wanted to further confirm our findings by determining the kinetic constants for this enzyme with L-fuconate as the substrate. Atu2811 dehydrated L-fuconate with the following kinetic constants: $k_{cat}$ \(1.5 \text{ s}^{-1}\), $K_M$ \(0.27 \text{ mM}\), and $k_{cat}/K_M$ \(5.6 \times 10^3\). These are in the range of what one would expect for members of the mandelate racemase subgroup. The published data for L-fuconate dehydration are noted in Table 2.5. We cannot rule out that one of the other substrates, 6-deoxy-L-talonate, L-galactonate, L-talonate, or D-ribonate is the true substrate for the enzyme.
but since the function of the cluster to which Atu2811 belongs has already been assigned L-fuconate dehydration activity, we preferred to focus our attention on those clusters with no known function. Since this activity was already known, these results were sufficient to serve as a positive control for the process of using SSNs and high throughput screens to discover functions. This represents the first known L-fuconate dehydratase in *A. tumefaciens* and demonstrates the conservation of this activity across different species of bacteria. The fact that this enzyme is active on more than one substrate (promiscuous) provides a scaffold from which additional activities can evolve.

<table>
<thead>
<tr>
<th>Locus Tag</th>
<th>$k_{\text{cat}}$ (s$^{-1}$)</th>
<th>$K_M$ (mM)</th>
<th>$k_{\text{cat}}/K_M$ (M$^{-1}$ s$^{-1}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>XCC4069</td>
<td>15</td>
<td>0.33</td>
<td>4.5 x 10$^4$</td>
</tr>
<tr>
<td>Atu2811</td>
<td>1.5</td>
<td>0.27</td>
<td>5.6 x 10$^3$</td>
</tr>
</tbody>
</table>

Table 2.5 Kinetic constants of Atu2811 compared to the published results for XCC4069.

No further work was done with Atu4196 in this study, but another graduate student went on to characterize Atu4196. She showed that it dehydrates meso-galactarate to form 2-keto-3-deoxy-D-threo-hexarate ($k_{\text{cat}}$ 0.12 s$^{-1}$, $K_M$ 0.80 mM, and $k_{\text{cat}}/K_M$ 1.5 x 10$^3$). She also showed that this enzyme has different active site side chain catalysts than previously described meso-galactarate dehydratases in the enolase superfamily thus representing a new family within the superfamily [21].
While we were able to reproduce identification of an L-fuconate dehydratase while identifying a novel family of meso-galactarate dehydratases, and another group assigned the function of D-galactarolactone isomerase to Atu3139, this still left four mandelate racemase subgroup members from A. tumefaciens with no assigned functions. This highlights one of the few limitations to the approach taken in this work. First, putative orthologues were not investigated. It could be the case that simply choosing other proteins from the same cluster as an A. tumefaciens unknown would reveal the activity of that cluster (e.g., perhaps the protein targeted was improperly folded). This work also demonstrates a limitation in physical libraries: they simply may not contain the true substrate because it is currently not possible to obtain all known carbohydrates, and even doing that is limited by what is known. Thus, these remaining unknowns serve as good candidates for future work using other approaches being developed in the EFI such as large scale genomic context analysis, in silico ligand docking to pathways, the screening of knockout mutants for growth deficiencies against commercially available and custom phenotype microarrays that contain a more extensive metabolite library [22], the use of tagless proteins or cell lysates, and making use of the protein production pipeline to go after all possible targets in each of the unknown clusters. The challenging targets could help to drive further development of orthologous strategies being tested and developed in the EFI.

During the course of this work another research group assigned the function of D-galactarolactone isomerase to Atu3139 using traditional biochemical methods. The substrate for this enzyme was not in our screening assay thus we would not have been able to identify a function for it using a high through-put activity screen. Nevertheless we were drawn to the gene neighborhood in which Atu3139 resides, and that work is the topic of the next chapter.
2.6 References


CHAPTER 3: GALACTARO-LACTONE ISOMERASE INVOLVED IN D-GALA ASSIMILATION BY A. TUMEFACIENS STRAIN C58


3.1 Introduction

D-GalA is the primary constituent of pectin, an abundant polymer found in plant cell walls [1] (Figure 3.1). Following hydrolysis by pectinases, D-galA can be used as a sole carbon source by many soil bacteria, including *A. tumefaciens* strain C58. Several pathways are known for catabolism of D-galA, including the oxidative pathway shown in Figure 3.2 [2]. Our attention was drawn to this pathway in *A. tumefaciens* strain C58 because it includes a reaction catalyzed by a member of the functionally diverse enolase superfamily. This enzyme, designated D-galactarolactone cycloisomerase (Gci, Atu3139), catalyzes the ring opening of D-galactaro-1,4-lactone to yield 2-keto-3-deoxy-L-threo-hexarate (hereafter designated as 5-keto-4-deoxy-D-galactarate or 5-keto-4-deoxy-D-glucarate, KDG) via a β-elimination reaction that is initiated by abstraction of the proton adjacent to the carboxylate group [3].
Figure 3.1 D-Galactose (upper left) is oxidized at carbon 6 to give D-galA (upper right). Pectin (bottom center) is a polymer of D-galA.

Figure 3.2 Oxidative pathway for D-galA assimilation. PicA is a pectinase.

In the first step of this pathway, an NAD⁺-dependent uronate dehydrogenase (Udh, Atu3143) oxidizes the β-pyranose form of D-galA to D-galactaro-1,5-lactone (δ-lactone). The assumption has been that the δ-lactone non-enzymatically isomerizes to D-galactaro-1,4-lactone (γ-lactone). Following this isomerization, galactarolactone cycloisomerase (Gci) catalyzes the ring opening reaction of the γ-lactone to produce KDG that undergoes
dehydration and decarboxylation to form α-ketoglutarate semialdehyde, which is oxidized to α-ketoglutarate, an intermediate in the citric acid cycle [2-5].

Atu3139 is located in a gene cluster that includes Atu3138 and Atu3140 (Figure 3.3). (This gene cluster also encodes orthologues of Kdul and KduD which are involved in an alternate D-galA utilization pathway that involves isomerization and reduction to produce 2-keto-3-deoxy-D-gluconate [6].) Based on sequence similarity, the closest functionally characterized homologue of the protein encoded by Atu3140 is a KDG dehydratase/decarboxylase (KdgD). Because KDG is the product of the reaction catalyzed by Gci, the proximity of these genes is not surprising. However, the function of the protein encoded by Atu3138 was unknown.

Figure 3.3 Gene cluster in A. tumefaciens.

The protein encoded by Atu3138 (UniProt accession number A9CEQ7) is a member of the amidohydrolase superfamily (AHS), a very large functionally diverse enzyme superfamily [7]. A9CEQ7 is a member of Pfam PF04909 (6287 sequences in release 27) and InterPro IPR006992 (11876 sequences in release 45.0). 2-Pyrone-4,6-dicarboxylate lactonase (PDC lactonase or LigI)
catalyzes the hydrolysis of 2-pyrone-4,6-dicarboxylate (PDC) to 4-oxalomesaconate and 4-carboxy-2-hydroxymuconate in lignin degradation, and is a structurally and mechanistically characterized member of these families [8].

PDC, the substrate for LigI, is structurally similar to D-galactaro-1,5-lactone: both are carboxy-substituted δ-lactones (Figure 3.4). In the pathway shown in Figure 3.2, D-galactaro-1,5-lactone has been proposed to isomerize to D-galactaro-1,4-lactone non-enzymatically. However, based on the proximity of the genes encoding A9CEQ7 and Gci, along with the structural similarity between PDC and D-galactaro-1,5-lactone, we hypothesized that A9CEQ7 catalyzes the 1,5- to 1,4-lactone isomerization and therefore is a “missing” enzyme in this pathway. At neutral pH, the non-enzymatic isomerization is fast, so previous investigators may have assumed there would be no need for a catalyst. Also, because Gci catalyzes the ring opening reaction of D-galactaro-1,5-lactone to KDG, no need for a lactonase is apparent, i.e., a reaction similar to that catalyzed by LigI. Here we report that A9CEQ7 is a galactaro-1,5-lactone isomerase (GLI), a novel activity for an AHS member.

![2-Pyrone-4,6-dicarboxylate (PDC) and D-Galactaro-1,5-lactone](image)

**Figure 3.4** LigI substrate on left; D-Galactaro-1,5-lactone on right.
3.2 Materials and Methods

3.2.1 RNA-Seq of RNA Isolated from *A. tumefaciens* strain C58 (Dr. Ritesh Kumar)

A single colony of *A. tumefaciens* strain C58 from a plated culture was used to inoculate 5 mL of 0.4% D-glucose-supplemented *A. tumefaciens* minimal medium (ABM) [3 g/L K₂HPO₄, 1 g/L NaH₂PO₄, 1 g/L NH₄Cl, 0.3 g/L MgSO₄·7H₂O, 0.15 g/L KCl, 0.01 g/L CaCl₂·2H₂O, and 0.0025 g/L FeSO₄·7H₂O (pH 7.0)] and grown to an OD₆₀₀ of 0.5. Cells were washed twice to remove the carbon source and then inoculated into 0.4% D-galA or 0.4% D-glucose minimal medium and grown for 2 h. Total RNA was isolated from the cells using an RNeasy kit (Qiagen). Samples were submitted to the Roy J. Carver Biotechnology Center for library preparation, data collection, and analysis.

3.2.2 Cloning of Atu3138 from *A. tumefaciens* strain C58 (Dr. Fiona Groninger-Poe, UIUC)

Atu3138 was PCR amplified from genomic DNA using Phusion DNA polymerase (New England Biolabs). The PCR reaction (50 µL) contained 1 ng of genomic DNA, 10 µL of 5X GC buffer, 1.5 µL of DMSO, 0.4 mM dNTPs, 20 pmol of each primer (forward primer 5’-CTGAATTCAGGATGCCAGACATATGAGCGAACTCGTCAGAAAAC -3’, reverse primer 5’-GGGGCCTCCCATTCCAACCATCTCGAGTAAAACTAGGTCGCC -3’), and 1 unit of Phusion DNA polymerase. The amplification used a PTC-200 gradient cycler (MJ Research) with the following parameters: 98 °C for 4 minutes followed by 35 total cycles of 98 °C for 20 seconds, 55 °C for 20 seconds, 72 °C for 30 seconds, and ending with a final extension at 72 °C for 7 minutes. The amplified gene was cloned into pET-15b vector.
3.2.3 Expression and Purification of A9CEQ7

Protein was expressed in an 8 L culture of *E. coli* strain BL21(DE3) cells. Expression consisted of growth in LB broth (supplemented with 100 µg/mL ampicillin) at 37 °C and 220 RPM for 18 hours. IPTG was not needed to induce expression. Cells were harvested by centrifugation (4650 x g, 4 °C) and resuspended in 30-40 mL of binding buffer (20 mM Tris-HCl pH 7.9, 0.2 M NaCl, and 5 mM MgCl₂). All cells were lysed by sonication, and the resulting lysate was pelleted to remove cell debris by centrifugation (31,000 x g, 4 °C). The clarified supernatant containing the His-tagged protein was then loaded onto a 5 mL chelating Sepharose Fast Flow column charged with nickel and equilibrated with 10 column volumes of binding buffer. The column was washed with 100 mL of 15% elution buffer (0.5 M imidazole, 20 mM Tris-HCl pH 7.9, 0.2 M NaCl, and 5 mM MgCl₂) containing 85% binding buffer and eluted over a 150 mL linear gradient of 15 to 100% elution buffer with an additional 100 mL at 100% elution buffer. Purity was checked by SDS-PAGE gel electrophoresis. Fractions containing protein were then pooled and dialyzed against a solution of 20 mM Na-HEPES pH 7.9, 5 mM MgCl₂, 0.1 M NaCl, and 10% glycerol. The mass of the polypeptide was confirmed by mass spectrometry (Figure 3.5). Protein was flash frozen drop wise into liquid nitrogen, and the beads were transferred to a cryogenic vial and stored at -80 °C.
Figure 3.5 Grayscale photo of an SDS-PAGE gel and mass spectrometry chromatogram for purified A9CEQ7. The expected mass of the his-tagged protein is 34016 daltons. Protein (arrow) in lane 1 and protein ladder in lane 2.

3.2.4 Construction, Expression, and Purification of the N240D Mutant

Asn 240 was mutated to aspartate using the following 50 μL reaction: 5 μL of 10X Pfx Amplification Buffer, 0.3 mM dNTP mixture, 1 mM MgSO₄, 0.25 μM of each primer (forward primer 5’- CATCGTCTGGGGCACCGACTGGCCGCATAATTC -3’, reverse primer 5’- GAATTATCGGCCAGTCGGTGCAGGATG -3’), 1.25 units of Platinum Pfx Platinum DNA Polymerase (Invitrogen), with varying amounts (6.25, 12.5, 25, and 50 ng) of template pET-15b with the gene encoding A9CEQ7 in H₂O. The amplification was performed using a PTC-200 gradient cycler (MJ Research) with the following cycling profile: 94 °C for 5 minutes; followed by 12 cycles of 94 °C for 30 seconds, 55 °C for 1 minute, and 68 °C for 7 minutes and 30 seconds. When complete, the reactions were pooled, 40 units of DpnI were added, and the mixture was incubated at 37 °C for one hour. Next the amplified product was gel-extracted and
electroporated into *E. coli* XL1-Blue cells. Single colonies were used to inoculate an overnight culture from which plasmid was isolated, and the correct sequence confirmed by Sanger sequencing (ACGT, INC.). Protein was expressed in a 6 L culture of *E. coli* BL21(DE3) cells. Expression consisted of growth in LB broth (supplemented with 100 µg/mL ampicillin) at 16 °C and 220 RPM for 48 hours. IPTG was not needed to induce expression. Cells were harvested by centrifugation (5000 x g, 4 °C) and resuspended in 75 mL of binding buffer (20 mM Tris-HCl pH 7.9, 0.5 M NaCl, and 5 mM MgCl₂). All cells were lysed by sonication, and the resulting lysate was pelleted to remove cell debris by centrifugation (30,000 x g, 4 °C). The clarified supernatant containing the His-tagged protein was then loaded onto a 50 mL chelating Sepharose Fast Flow column charged with nickel and equilibrated with 10 column volumes of binding buffer. The protein was eluted over a 1200 mL linear gradient of 0 to 25% elution buffer (1 M imidazole, 20 mM Tris-HCl pH 7.9, 0.5 M NaCl, and 5 mM MgCl₂). Purity was checked by SDS-PAGE gel electrophoresis. Fractions containing protein were then pooled and dialyzed against a solution of 20 mM Tris pH 7.9, 5 mM MgCl₂, and 0.15 M NaCl. The mass of the polypeptide was confirmed by mass spectrometry (Figure 3.6). Protein was flash frozen drop wise into liquid nitrogen, and the beads were transferred to a cryogenic vial and stored at -80 °C.
3.2.5 Preparation of D-Galactaro-1,5-lactone by Bromine Oxidation

D-Galactaro-1,5-lactone (δ-lactone) was prepared from the sodium salt of D-galacturonic acid using the procedure described by Isbell and Frush [9] with minor modifications: at ½ the scale in D$_2$O, 2X bromine, a reaction time of 22.5 minutes, and separation in a separatory funnel at 4 °C. The final product was stored at -80 °C (Table 3.1 and Figure 3.7).
Table 3.1 $^1$H-NMR chemical shifts for D-galactaro-$\delta$-lactone and D-galactaro-$\gamma$-lactone at pD 4.0.

<table>
<thead>
<tr>
<th>Compound</th>
<th>$\delta$ (PPM)</th>
<th>$J$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) D-Galactaro-$\delta$-Lactone</td>
<td>4.32 d</td>
<td>10.62</td>
</tr>
<tr>
<td></td>
<td>4.00 dd 2.76, 10.56</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4.40 t</td>
<td>2.31</td>
</tr>
<tr>
<td></td>
<td>4.71 d</td>
<td>1.98</td>
</tr>
<tr>
<td>2) D-Galactaro-$\gamma$-Lactone</td>
<td>4.11 d</td>
<td>1.74</td>
</tr>
<tr>
<td></td>
<td>4.47 dd 1.74, 8.52</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4.28 t</td>
<td>8.82</td>
</tr>
<tr>
<td></td>
<td>4.52 d</td>
<td>9.24</td>
</tr>
</tbody>
</table>
Figure 3.7 $^1$H-NMR spectra for (A) D-galactaro-$\delta$-lactone, (B) Udh reaction, and (C) D-galactaro-$\gamma$-lactone at pD 4.0. (A) Blue bars denote D-galA. Numbers correspond to the hydrogen attached to the carbon as numbered in structure number one of Table 3.1. (B) Arrows point to the $\delta$-lactone after acid quenching and removal of NAD$^+$/NADH from the oxidation of D-galA by the Udh. (C) Numbers correspond to the hydrogen attached to the carbon as numbered in structure number two of Table 3.1.

3.2.6 Udh Assay

To confirm that the $^1$H-NMR spectrum of the product from the bromine oxidation was indistinguishable from the $^1$H-NMR spectrum of the product from the Udh reaction, the spectra were compared (Figures 3.7A and 3.7B). To obtain a spectrum of the product of the Udh
reaction, 5 μM Udh was added to a reaction mixture containing 50 mM sodium phosphate buffer, 2 mM MgCl₂, 5 mM D-galA, and 6 mM NAD⁺ at pD 6.4 in a final volume of 500 μL. After a one minute incubation at room temperature the reaction was stopped by decreasing the pD to 4 with the addition of DCl [10]. The reaction mixture then was mixed with 2% (w/v) activated charcoal (100 mesh) for one minute to remove the NAD⁺/NADH and then centrifuged for 30 seconds in a microcentrifuge. The supernatant was filtered through a syringe filter to remove any additional charcoal; 300 μL was added to a 5 mm Shigemi NMR tube matched to D₂O. All NMR spectra were collected on an Agilent 600MHz spectrometer.

3.2.7 Activity Assay by ¹H-NMR Spectroscopy

To determine if D-galactaro-1,5-lactone is a substrate for A9CEQ7, the conversion of D-galactaro-1,5-lactone to D-galactaro-1,4-lactone (γ-lactone) was monitored by ¹H-NMR spectroscopy. A 650 μL reaction contained 6 mM D-galactaro-1,5-lactone, 10 nM enzyme, 50 mM sodium phosphate buffer, and 2 mM MgCl₂, at pD 6.4 in D₂O. The spectrum of D-galactaro-1,5-lactone was acquired, and then 10 nM A9CEQ7 was added. A spectrum was recorded every two minutes at room temperature over the course of one hour. All NMR spectra were collected on an Agilent 600MHz spectrometer.

3.2.8 Metal Testing

A continuous polarimetric assay was used to determine the effect of added divalent cations on A9CEQ7’s activity. The change in optical rotation was monitored using a polarimeter (Jasco P-1010) and a mercury line filter (405 nm). A 1300 μL reaction containing 100 nM
enzyme, 50 mM Na MES buffer, 2 mM D-galactaro-1,5-lactone, and 2 mM of Ca$^{2+}$, Co$^{2+}$, Mg$^{2+}$, Mn$^{2+}$, Ni$^{2+}$, or Zn$^{2+}$ at pD 6.4 in D$_2$O was monitored for 10 minutes at room temperature.

3.2.9 Kinetic Assay by Polarimetry

A continuous polarimetric assay was used to determine the kinetic parameters of A9CEQ7 using a Jasco P-1010 as described above. A 1300 μL reaction containing 100 nM enzyme and 50 mM sodium phosphate buffer at pD 6.4 in D$_2$O was monitored for one minute at room temperature with varying substrate concentrations (0.24 mM – 5.85 mM). The rates of the uncatalyzed reaction and the reaction catalyzed by the N240D mutant enzyme were determined using the same reaction conditions and monitoring for an additional one minute and 30 seconds. All enzymes were exchanged into 50 mM sodium phosphate buffer at pD 7.7 in D$_2$O prior to kinetic assays being performed. Data were fit to the Michaelis-Menten equation in Sigma Plot using the Enzyme Kinetics Module.

3.2.10 Crystallization and Structure Determination (Dr. Matt Vetting, AECOM)

A9CEQ7 was crystallized by sitting drop vapor diffusion using 96-well Intelliplates (Douglas Instruments). Crystallization drops were assembled by combining 0.5 μl of protein (10 mg/mL in 20 mM Tris-HCl pH 7.9, 0.15 M NaCl, 5 mM MgCl$_2$) with 0.5 μl of the reservoir (70 μl of 0.1 M Na-HEPES pH 7.5, 20% (w/v) PEG 4000, 10% (w/v) isopropanol). Crystals grew as intertwined rods over 7 days with dimensions of 50 x 50 x 200 μm. A single crystal was extracted, transferred to the reservoir condition with 20% (w/v) isopropanol, and then vitrified by plunging in liquid nitrogen. Data were collected at the Advanced Photon Source (Beamline
31-ID of Argonne National Laboratories) using a wavelength of 0.9788 Å and a Rayonix 225 HE detector (Rayonix, Llc). Data were integrated in MOSFLM [11] and scaled using SCALA [12]. An ensemble model was generated from a structurally unique set of structures available from IPR006992 (PDB IDs: 2ffi, 4do7, 4i6k, 4dia, 4di8) and used in molecular replacement. The program PHASER [13] within PHENIX [14] located the three molecules within the asymmetric unit and the final model was built with iterative cycles of fitting to electron density maps within the molecular graphics program COOT [15], followed by refinement within PHENIX. The quality and stereochemistry of the final structure were verified using MOLPROBITY [16] (refinement statistics in Table 3.2).

Table 3.2 Data Collection and Refinement Statistics.¹

<table>
<thead>
<tr>
<th>DATASET STATISTICS</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Space Group</td>
<td>P2₁</td>
</tr>
<tr>
<td>Unit Cell (Å, °)</td>
<td>a=53.2, b=77.6, c=99.7, β=95.5</td>
</tr>
<tr>
<td>Resolution (Å)</td>
<td>31.6-1.6 (1.69-1.60)</td>
</tr>
<tr>
<td>Completeness (%)</td>
<td>97.1 (100.0)</td>
</tr>
<tr>
<td>Redundancy</td>
<td>3.6 (3.5)</td>
</tr>
<tr>
<td>Mean(l)/sd(l)</td>
<td>8.5 (2.0)</td>
</tr>
<tr>
<td>Rsym</td>
<td>0.095 (0.541)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>STRUCTURE STATISTICS</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Resolution (Å)</td>
<td>31.6-1.6 (1.62-1.60)</td>
</tr>
<tr>
<td>Unique reflections</td>
<td>103114 (3564)</td>
</tr>
<tr>
<td>Rcryst (%)</td>
<td>14.7 (21.8)</td>
</tr>
<tr>
<td>Rfree (%, 5% of data)</td>
<td>18.1 (23.4)</td>
</tr>
</tbody>
</table>
Table 3.2 (cont.)

<table>
<thead>
<tr>
<th>Residues / Waters / Atoms total</th>
<th>842 / 1073 / 8205</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average B-factor (Å²)</td>
<td></td>
</tr>
<tr>
<td>Protein/Waters</td>
<td>17.4 / 30.3</td>
</tr>
<tr>
<td>RMSD</td>
<td></td>
</tr>
<tr>
<td>Bond Lengths (Å) / Angles (degrees)</td>
<td>0.009 / 1.276</td>
</tr>
</tbody>
</table>

MOLPROBITY STATISTICS

| Ramachandran Favored / Outliers (%) | 98.6 (0.0) |
| Clashscore b | 1.78 (99th pctl.) |
| Overall score b | 0.94 (100th pctl.) |

^a Statistics in parenthesis are for the highest resolution bin
^b Scores are ranked according to structures of similar resolution as formulated in MOLPROBITY

3.3 Results and Conclusion

*A. tumefaciens* strain C58 can use D-galA as a sole source of carbon. Figure 3.8 shows the genome neighborhood and the fold change in gene expression levels for cells grown on D-galA relative to cells grown on D-glucose. On the basis of the RNA-Seq results, Atu3138 is up-regulated 46-fold (top 1% of all genes in the genome). These results suggest that Atu3138 is part of a pathway for assimilating D-galA.
Figure 3.8 Numbered gene cluster from Figure 3.3 containing the putative lactonase (Atu3138). Note the transporters (9-11) are upregulated over 100 fold when A. tumefaciens is grown on D-galacturonate as a sole source of carbon. Atu3138 and Atu3138 (Gci) are up 46 and 65 fold respectively.

D-Galactaro-1,5-lactone was prepared in D_{2}O via bromine oxidation of D-galA [9]. This material was indistinguishable by ^1H-NMR spectroscopy from the product obtained by the NAD^+ dependent Udh (Atu3143). The synthetic δ-lactone is stable indefinitely at pD 4.8 and -80 °C.

The progress of the A9CEQ7 catalyzed reaction was monitored by ^1H-NMR spectroscopy (Figure 3.9). The ^1H-NMR spectrum of the synthetic 1,5-lactone at pD 6.4 is shown in Figure
3.9A; the red bars indicate the 1,5-lactone, the green bars the 1,4-lactone, and the blue bars the residual D-galA from the synthesis. The spectra in Figure 3.9B (recorded at two minute intervals) show the progress of the reaction. As the reaction proceeds, the resonances associated with the 1,5-lactone decrease in intensity as the resonances associated with the 1,4-lactone increase in intensity. After one hour (Figure 3.9C), the only remaining resonances (in addition to those associated with the residual D-galA) are those associated with the 1,4-lactone. No meso-galactarate is detected by hydrolysis of either lactone.

Figure 3.9 ^1H-NMR spectra of the isomerase substrate and product. A) Synthetic 1,5-lactone immediately after the pH was adjusted from 4.8 to 6.4. B) Time course of the reaction. C) Reaction after one hour. Resonances are color-coded to match the peaks associated with the structures in Figure 3.2.
The reaction progress was monitored with a polarimeter. In Figure 3.10, the black line shows the change in optical rotation during the non-enzymatic isomerization at pH 6.4; the red line shows the change in optical rotation during the reaction in the presence of A9CEQ7. The kinetic parameters were measured for the A9CEQ7 catalyzed reaction: $k_{\text{cat}}$, 440 s$^{-1}$; $k_{\text{cat}}/K_m$, 8.3 x 10$^4$ M$^{-1}$ s$^{-1}$. The rate enhancement is 6.8 x 10$^5$. The rate of the non-enzymatic reaction is dependent on pH, with pH 6.4 providing sufficient kinetic stability of the 1,5-lactone to allow confident rate measurements.

![Figure 3.10 Polarimetric profiles of 5.5 mM D-galactaro-1,5-lactone without enzyme (black diamonds), with A9CEQ7 (red squares), and with Gci (blue circles). Profile of 5.5 mM D-galactaro-1,4-lactone with Gci (green triangles).](image)

We also investigated the specificity of Gci for the 1,4- and 1,5-lactones. In Figure 3.10, the green line shows the progress of the reaction when Gci is added to 1,4-lactone produced in the absence of A9CEQ7; the blue line shows the progress of the reaction when Gci is added to a reaction mixture containing the 1,5-lactone. These reactions reach the same final optical
rotation. The rate constants for the non-enzymatic isomerization of the 1,5-lactone to the 1,4-lactone (black line) and the Gci catalyzed production of KDG in the absence of A9CEQ7 (blue line) are the same, establishing that Gci does not catalyze the ring opening of the 1,5-lactone.

On the basis of these results, we assign the D-galactarolactone isomerase (GLI) function to A9CEQ7. Although several lactonases have been identified in the AHS, this is the first lactone isomerase reaction.

The structure of GLI was determined by Dr. Matt Vetting (Albert Einstein College of Medicine) in the absence of a ligand to 1.6 Å by molecular replacement using an ensemble model of structural homologues from IPR006992. GLI has a typical amidohydrolase fold with a distorted (β/α)₈-TIM barrel. The closest structural homologue as determined by PDBFold [17] is LigI from Sphingomonas paucimobilis (PDB ID 4DI8) with an RMSD of 2.10 Å and a sequence identity of 27% over 253 Cαs. Despite low sequence identity and high RMSD values, LigI and GLI have very similar core TIM-barrel structures (Figure 3.11A). As expected by the differences in their substrates, LigI’s a flat planar dicarboxylate lactone and GLI’s a monocarboxylate sugar lactone, most of LigI and GLI’s differences are localized to the loops at the N terminal end of the barrel that play a role in substrate recognition. For example, Arg 130 in LigI, which coordinates the distal carboxylate of its substrate, is a proline in GLI (Pro 123).
Figure 3.11 Structure of A9CEQ7. A) Distorted (β/α)$_8$-TIM barrel. A9CEQ7 in cyan with supposed catalytic resides in yellow. LigI (4D8L) in magenta. B) Superimposition of catalytic residues with those of LigI and the 1,5-lactone with PDC. Note an asparagine (N240) has replaced an aspartate involved in activating a water at the end of the 8th β-strand.

Unlike many members of the amidohydrolase superfamily, LigI does not require a divalent metal for its lactonase activity. The presence of Ca$^{2+}$, Co$^{2+}$, Mg$^{2+}$, Mn$^{2+}$, Ni$^{2+}$, or Zn$^{2+}$ did not increase the rate of the GLI reaction (data not shown). These results suggest GLI also does not require a metal for catalytic activity.

In LigI, Asp 248 is the general base that activates a water molecule for nucleophilic attack on the lactone carbonyl group [8]. A superposition of the active sites of GLI and the PDC liganded D248A mutant of LigI at pH 8.5 is shown in Figure 3.11B. The active sites are essentially identical, with the notable exception that in GLI, Asn 240 is the structural homologue of Asp
248 in LigI. The N240D substitution produced a 28 fold decrease in activity (k_{cat}, 16 s^{-1}; k_{cat}/K_m, 3.6 \times 10^3 \text{M}^{-1} \text{s}^{-1}); no hydrolysis of the 1,5- or 1,4-lactone was observed in the presence of the D240N mutant.

Thus, the lactone isomerization reaction catalyzed by GLI can be explained by the lack of a general base for lactone hydrolysis. Inspection of the active site of GLI does not reveal either a general base that would activate the 4-OH group of the 1,5-lactone for intramolecular attack on the lactone carbonyl group or a nucleophile that would allow the formation of an acyl enzyme intermediate that would partition between the 1,5- and 1,4-lactones. Therefore, in analogy to the proposed mechanism for the non-enzymatic reaction, we assume the modest rate enhancement results from preferential binding of the boat conformer of the 1,5-lactone to enforce proximity of the 4 OH group and the lactone carbonyl group (Figure 3.12). The ability of GLI to catalyze lactone isomerization instead of hydrolysis can be explained by the absence of the general basic catalysis used by LigI.

This work combined SSNs, genomic context, and pathway reconstruction along with in vitro biochemistry and gene expression analysis to predict and confirm an in vitro activity for a novel D-galactarolactone isomerase, that catalyzes the isomerization of D-galactaro-1,5-lactone to D-galactaro-1,4-lactone; the gene that encodes this protein is up-regulated 46-fold when A.
*Agrobacterium tumefaciens* is grown on D-galA compared to growth on D-glucose. This result taken along with the in vitro evidence suggests that the physiological role of this enzyme is in the D-galA assimilation pathway. The generation and testing of knock out mutants plus complementation assays are needed to further confirm a physiological role for this novel activity and are the part of future work on this project.

In the next chapter the use of a combination of transporter proteins, regulon reconstruction, pathway reconstruction, and *in vitro* biochemistry will be described, in an approach that extends exploration of hexuronate (e.g., D-galA) assimilation to other classes of Proteobacteria and reveals a novel pathway variant of hexuronate metabolism.

3.4 References


CHAPTER 4: TRAP SBP INITIATED REGULON AND PATHWAY RECONSTRUCTION REVEALS NOVEL COMPONENTS OF D-GALA ASSIMILATION PATHWAYS IN SEVERAL CLASSES OF PROTEOBACTERIA


4.1 Introduction

In previous work as part of the EFI, the binding specificities of a diverse set of bacterial tripartite ATP-independent periplasmic (TRAP) transporter solute binding proteins (SPBs) (Pfam family PF03480, SBP_bac_7) were investigated by screening them against a library of 189 compounds using differential scanning fluorimetry (DSF). The goal of the screening was to identify the initial metabolite in a pathway to constrain the chemical space for possible substrates of downstream enzymes [1, 2]. In that work a large, isofunctional TRAP SBP cluster was identified whose purified members hit on either D-glcA and/or D-galA in the DSF assay with change in melting temperatures ($\Delta T_m$) greater than 7°C. The sequences in the cluster are greater than 30% identical, and those sequences make up greater than 20% of the TRAP SBP family (Figure 4.1) [2].
Figure 4.1 A) Diagram of three well known types of transporters. From left to right: a primary, active transporter, the ATP-binding cassette transporters (ABC transporters); they use the energy from ATP hydrolysis to transport solutes across the cell membrane using 3 components, the transmembrane, the solute binding, and the nucleotide binding component; a tripartite ATP-independent periplasmic (TRAP) transporters; they do not use ATP hydrolysis and thus do not have a nucleotide binding component but do have transmembrane and solute binding components; finally a transporter with only a transmembrane permease component. B) From left to right, then top to bottom: schematic of prototypical bacterial cell depicting transport as the first step prior to entry of a metabolite into a gene cluster containing enzymes of unknown function and/or specificity. TRAP SBP protein SSN showing a sample of targets selected for screening – a diverse set of targets was selected. Image of sample data from differential scanning fluorimetry (DSF) experiment. Table showing types of compounds in the DSF screen. C) Zoomed in view of TRAP SBP SSN with cluster of interest highlighted in yellow. Targets from this cluster hit on D-glcA and/or D-galA.
D-GlcA and D-galA are highly abundant sources of carbon found in plants, animals, and bacteria with a myriad of uses in medicine and industry. D-GlcA is common in the glycosaminoglycan (GAG) chains of proteoglycans in animals as well as in glucuronan and xanthan. D-GalA is the primary monomer of pectin, a polymer found in plant cell walls [3, 4].

Many bacteria can assimilate D-galA using one of several known metabolic pathways [5] such as the oxidative pathway described in Chapter 3 (Figure 4.2) used for D-galA, whose first step is transport of the compound into the cytoplasm. D-GlcA and D-galA, are taken up by *Escherichia coli* via secondary transporters from the major facilitator superfamily (Pfam clan CL0015) and the gluconate permease family (PF02447, GntP_permease) [6, 7]; however, the mechanisms of their uptake in other bacteria were mostly unknown. In the oxidative pathway for D-galA, after import, an NAD\(^+\)-dependent uronate dehydrogenase (Udh) oxidizes the β-pyranose form of D-galA to D-galactaro-1,5-lactone (δ-lactone). Next, a lactone isomerase (Gli) catalyzes the isomerization of D-galactaro-1,5-lactone to D-galactaro-1,4-lactone (γ-lactone). After this isomerization, Gci catalyzes the ring opening reaction of the γ-lactone to produce KDG that undergoes dehydration and decarboxylation by KdgD to form α-ketoglutarate semialdehyde, which is oxidized by KgsD to α-ketoglutarate, an intermediate in the citric acid cycle. Although Udh and Gci also accept glucarolactones as substrates (glucaro- and galactarolactone are C4 epimers), to our knowledge, direct evidence linking this pathway to D-glC\(\)A assimilation has not been reported [9, 29-32]; thus, we do not know if the substrate promiscuity observed for the TRAP SBPs, Udh, and Gci is a result of the structural similarity between D-glC\(\)A and D-galA, or if the pathway is responsible for assimilating both compounds. The fidelity of the pathway will be the topic of future work. Our previous work (Chapter 3) on D-
galA degradation pathway in *A. tumefaciens* in combination with the discovery of a new hexuronate transport mechanism (TRAP SBPs) led us to the focus of this work. We reasoned that the presence of previously unknown D-galA transport systems may indicate the presence of unknown regulators and enzymes involved in the assimilation of D-galA.

**Figure 4.2** Oxidative pathway leading into the TCA cycle described in the literature for the catabolism of D-galA. GudD and GarD are dehydratases that feed D-glucarate and meso-galactarate directly into the downstream portion of the pathway.

In this work, we used comparative genomics to reconstruct regulons and pathways using the TRAP SBPs that hit on D-glcA and D-galA or just D-galA as starting points. For example, an analysis of the TRAP SBP encoded by Bpro_3107 from *Polaromonas* sp. JS666 (Figure 4.3) is consistent with these TRAP SBPs belonging to predicted regulons and/or operons that encode enzymes orthologous to those that function together in the known oxidative pathway for microbial metabolism of D-galA. The predicted regulons are conserved in *Ralstonia pickettii* 12J; since sugar acid metabolism by *R. pickettii* 12J was already under investigation in another project in our lab, it was chosen as the host strain in this study. Further analysis identified a novel family of regulators from the gluconate repressor family (PF00392, GntR) and two novel
families from the lysine regulator family (PF03466, LysR_substrate). These regulators and their nucleotide binding sites are conserved across several classes of Proteobacteria (α, β, and γ). In cases where the regulons did not encode enzymes orthologous to those that function together in the known oxidative pathway for microbial metabolism of D-galA, namely δ-lactone isomerase (Gli) and γ-lactone cycloisomerase (Gci) [8, 9], further analysis revealed that members of these regulator families regulate the transcription of operons that encode a previously uncharacterized subfamily of lactone hydrolases from the SMP-30/gluconolactonase/LRE-like region family (PF08450, SGL), a member of the beta propeller clan (CL0186), that catalyze the direct conversion of the δ-lactone product of the Udh reaction to the ring-opened diacid, meso-galactarate, in an alternate route for its assimilation (It also converts D-glucara-1,5-lactone to D-glucarate). The reconstructed regulons also led us to another previously unknown enzyme from the aldose 1-epimerase family (PF01263, Aldose_epim) that catalyzes the mutarotation of the α- and β-anomers of D-glcA and D-galA (but not of D-glucose nor D-galactose), presumably after transport of the monosaccharide or post hydrolysis by a uronidase (e.g., glucuronidase, pectinase) in the cytosol upstream of the uronate dehydrogenase. Finally, a previously unknown family of tripartite tricarboxylate transporter family receptor (PF03401, TctC) proteins for binding and facilitating the import of the diacid, D-glucarate, was also encoded within these novel regulons. This work extends exploration of D-galA assimilation across several classes of Proteobacteria, and it also extends the reach of the EFI beyond strictly enzymes to both regulators and transporters.
Figure 4.3 A) *Polaromonas* sp. JS666 TRAP SBP from yellow cluster hit on D-glcA and D-galA in the DSF screen. B) Result of the regulon and pathway reconstruction. Phylogenetic tree of closely related Betaproteobacteria. Loci from each genome are depicted as part of the labels at the tips of the tree. *Polaromonas* sp. JS666 is represented at the bottom of the tree. The TRAP SBP Bpro_3107 is identified in red. Orange circles represent transcription factor binding sites conserved across these organisms. A reconstructed pathway is at the bottom of the image. Note the presence of UxuL (putative novel lactone hydrolase) upstream of Bpro_3107. In the absence of Gli and Gci (lactone isomerase and cycloisomerase), UxuL would linearize the δ-lactone to meso-galactarate.
4.2 Materials and Methods

4.2.1 Bioinformatics Methods for Metabolic Regulon and Pathway Reconstruction (Dr. Dmitry Rodionov, SB MRI)

The analyzed proteobacterial genomes, including 3 α-Proteobacteria, 12 β-Proteobacteria, and 19 γ-Proteobacteria were downloaded from the MicrobesOnline genome database [10]. Locus tag gene identifiers are used throughout. Orthologues between proteins from different taxonomic groups were defined as bidirectional best hits with a 30% identity threshold using the Smith-Waterman algorithm implemented in the GenomeExplorer program [11]. In dubious cases orthologues were confirmed by construction of phylogenetic trees and comparative analysis of gene neighborhoods using the MicrobesOnline tree browse tool [10]. Functional gene assignments and metabolic subsystem analysis were performed using the SEED annotation/analysis tool, which combines protein similarity search, positional gene clustering, and phylogenetic profiling of genes [12]. In addition, distant homology to characterized proteins in the Swiss-Prot protein database [13] and the InterPro [14] and PFAM [15] protein databases were used to verify protein functional and structural annotations. Multiple sequence alignments were constructed by MUSCLE [16]. Phylogenetic trees were built using a maximum likelihood algorithm implemented in the proml tool from the PHYLIP Package [17]. Trees were visualized using Dendroscope [18]. For genomic reconstruction of novel GguR, GudR and GulR regulons, we used the comparative genomics approach based on identification of candidate regulator-binding sites in closely related bacterial genomes reviewed in Rodionov 2007 [19]. First, we revealed orthologues of each regulator and analyzed the genomic context of their genes to reveal co-localizations with other catabolic pathway genes. Phylogenetic analysis of
each regulator family revealed several major groups of orthologues. For each group of orthologous proteins, we identified their putative DNA binding motifs using a combination of the phylogenetic footprinting approach and the motif-discovery tools implemented in the RegPredict Web server (http://regpredict.lbl.gov/) [20]. We collected training sets of orthologous upstream gene regions for each prospective regulator-controlled operon that were determined via the genome context analysis of the respective regulator-encoding genes. The upstream regions were analyzed using a DNA motif recognition program (the “Discover Profile” procedure implemented in RegPredict) to identify conserved palindromic DNA motifs. After construction of a positional-weight matrix for each identified DNA motif, the studied genomes were scanned with recognition matrices to determine additional candidate binding sites and finally reconstruct the respective regulons using the RegPredict tool. The scores of sites were calculated as a sum of nucleotide weights for each position. Sequence logos were built using the WebLogo package [21]. Reconstructed regulons are represented in the RegPrecise database (http://regprecise.lbl.gov/) [22].

4.2.2 Bacterial Strains and Reagents

*E. coli* and *R. pickettii* 12J strains were used in this study. *E. coli* strains XL1-Blue and BL21(DE3) were used for gene cloning and protein overexpression, respectively. *E. coli* strains were maintained in LB or on LB agar plates; *R. pickettii* 12J was maintained in NB (nutrient broth) or on NB agar plates. Ampicillin (50-100 μg/mL) was used as appropriate. For growth experiments and RNA isolation, M9 minimal medium was used, and D-glucose or D-galA was supplemented as carbon sources. Cell growth was monitored at 600 nm using a BioScreen C or
a Spectronic 20 Genesys depending on the experiment. Reagents were purchased from Sigma-Aldrich unless otherwise noted.

4.2.3 Cloning, Expression, and Protein Purification of Lactone Hydrolase and TctC Solute Binding Protein (Dr. Salehe Ghasempur, UIUC; Nawar Al-Obaidi, AECOM)

The lactone hydrolase, Rpic_4446 (UniProt AC B2UIY8, Pfam family SGL), and Bpro_3101 (Q128M7, TctC) were cloned, expressed, and purified as previously described [2]. Briefly, genes of interest were PCR amplified from genomic DNA using KOD Hot Start DNA Polymerase, with the resultant amplified fragment ligated into the N-terminal, TEV cleavable, 6X-His-tag vector, pNIC28-Bsa4 [23], by ligation independent cloning [24]. Vectors containing the genes of interest were transformed into BL21-CodonPlus(DE3)-RIL (Stratagene), and proteins were expressed using ZYP-5052 auto induction media in an LEX48 airlift fermenter with overnight growth at 22°C for 16-22 h. Proteins were purified by HisTrap Ni-NTA chromatography with the eluted target injected onto an inline HiLoad S200 16/60 pg (GE Healthcare) gel filtration column equilibrated with 20 mM HEPES (pH 7.5), 150 mM NaCl, 5% glycerol, and 5 mM DTT. Elutions were analyzed by SDS PAGE, with homogenous fractions concentrated by centrifugal ultracentrifugation to 10-20 mg mL⁻¹, snap frozen in liquid nitrogen, and stored at -80°C.

4.2.4 Cloning, Expression, and Protein Purification of the Mutarotase (Dr. Salehe Ghasempur, UIUC)

The gene encoding the mutarotase, Rpic_0950 (B2U9A3, Aldose_epim), was PCR amplified from R.ickettii 12J genomic DNA. PCR forward and reverse primers were synthesized by Integrated DNA Technologies (IDT) containing NdeI and XhoI restriction sites,
respectively (forward primer sequence: 5’-GACTGACCATATGCACCTGAACGTAG-3’; reverse primer sequence 5’-GTGCGGGGCGTACTGAGTTC-3’). The 100 µl PCR reaction contained 1 ng DNA, 20 µl 5X Phusion HF buffer, 0.2 mM dNTP mixture, 1 µl of NEB Phusion enzyme, 40 pmols of each primer, and used a PTC-200 Gradient thermocycler. The digested PCR products were ligated into the pET-15b vector linearized with the same restriction enzymes. The ligation product was transformed into XL1-Blue electrocompetent cells.

Rpic_0950 was overexpressed in BL21(DE3) cells transformed with the expression construct at 37°C in the presence of 100 µg/mL ampicillin until reaching an OD$_{600}$ of 0.5, at which point the cells were induced with 1 mM IPTG. Growth was continued overnight. The cells were harvested by centrifugation (15 min at 4500 x g), resuspended in binding buffer (5 mM imidazole, 0.5 M NaCl, 5 mM MgCl$_2$, 20 mM TRIS pH 7.9), lysed by sonication, and clarified by centrifugation. The lysate containing the His-tagged protein was loaded onto a 5 mL HisTrap FF Crude column (GE Healthcare) equilibrated with binding buffer and eluted with a linear 80 mL gradient from 0% to 100% of elution buffer (1 M imidazole, 0.5 M NaCl, 20 mM Tris-HCl pH 7.9).

4.2.5 Cloning, Expression, and Protein Purification of the Regulators (Dr. Irinia Rodionova, SB MRI)

Bpro_3110 (GguR) and Bpro_3418 (GudR) from *Polaromonas* sp. JS666 and Rpic_0945 (GguR) and Rpic_4453 (GulR) from *R. pickettii* 12J were amplified from genomic DNA using the polymerase chain reaction. Reaction mixtures contained 5 µL of 10X Pfx amplification buffer (Invitrogen), 0.3 mM dNTPs, 1 mM MgSO$_4$, primers at 0.3 µM each (Table 4.1), 1 unit of Pfx DNA polymerase, and 50 ng of genomic DNA in a total volume of 50 µL. Reactions were performed
using a PTC-200 gradient cycler (MJ Research) with the following cycling profile: 94°C for 5 min; followed by 35 cycles of 94°C for 15 s, 60°C for 30 s, and 68°C for 1 min and 30 s; followed by a final extension step at 68°C for 10 min. Reactions were run in duplicate and then pooled. Amplicons were cloned into pET-15b (N-terminal fusion 6X-His-tag) vectors using conventional cloning methods (PCR amplification followed by cloning into multiple cloning restriction sites in the vectors).

<table>
<thead>
<tr>
<th>Name</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rpic_0945_Ndel_F</td>
<td>GAGACCATATGCCCACCCGCCAGTTC</td>
</tr>
<tr>
<td>Rpic_0945_BamHI_R</td>
<td>CGTACACACATGGATCCGTATTGGCCGG</td>
</tr>
<tr>
<td>Rpic_4453_Ndel_F</td>
<td>CCGTTTTTGAACATATGTITGAACTCAGCACAATCC</td>
</tr>
<tr>
<td>Rpic_4453_BamHI_R</td>
<td>CAAGTGCCGATCCCGGTAGTTTCAG</td>
</tr>
<tr>
<td>Bpro_3110_Ndel_F</td>
<td>CATTGCTTGTCAACGGTTTTGCATATGACCCTCAGGAAAACC</td>
</tr>
<tr>
<td>Bpro_3110_BamHI_R</td>
<td>GCAAACAGAGGTTCACGGGATCCAGTCAGGCCCTGGTTC</td>
</tr>
<tr>
<td>Bpro_3418_Ndel_F</td>
<td>CGTCGATAGAATCGGCATATGACCACGGAGGAC</td>
</tr>
<tr>
<td>Bpro_3418_BamHI_R</td>
<td>GTGGTGCGGATCCAAATTGCCGGGTCACTC</td>
</tr>
</tbody>
</table>

Table 4.1 Primers used in this study.

Recombinant GguR and GudR regulator proteins were overexpressed as N-terminal fusion proteins with a 6X-His-tag in BL21(DE3) cells. Cells were grown in LB media to an OD$_{600}$ equal to 0.8 at 37°C, induced with 0.2 mM IPTG, and harvested after 12 h shaking at 20°C. Protein purification was performed using the rapid Ni-NTA agarose minicolumn protocol as previously described [25]. Briefly, harvested cells were resuspended in 20 mM HEPES buffer, pH 7.0, containing 100 mM NaCl, 0.03% Brij-35 detergent, and 2 mM β-mercaptoethanol.
supplemented with 2 mM phenylmethylsulfonyl fluoride and a protease inhibitor cocktail. Lysozyme was added to 1 mg/mL, and the cells were lysed by freezing-thawing followed by sonication. After centrifugation at 18,000 rpm, Tris-HCl buffer (pH 8.0) was added to the supernatant to a final concentration of 50 mM, and the mixture was loaded onto a Ni-NTA agarose column (0.2 mL). After washing with the starting buffer containing 1 M NaCl and 0.3% Brij-35 detergent, bound proteins were eluted with 0.3 mL of the starting buffer containing 250 mM imidazole. Protein size, expression level, distribution between soluble and insoluble forms, and extent of purification were monitored by SDS-PAGE.

4.2.6 Preparation of Lactones

The δ-lactones of D-glcA and D-galA were prepared from their sodium salts using the procedure described by Isbell and Frush [26] with the following modifications: at ½ the scale in D₂O, 2X bromine, a reaction time of 22 minutes and 30 seconds, and separation in a separatory funnel at 4 °C. Final products were stored at -80 °C (Figure 4.4). D-glucaro-1,4-lactone was purchased from Toronto Research Chemicals Inc. D-Galactaro-1,4-lactone was formed by incubating the δ-lactone overnight at room temperature.
Figure 4.4 D-glucaro-1,5-lactone was chemically prepared to determine the specificity of the lactone hydrolase. Top spectrum is D-glcA. Middle spectrum is the reaction mixture after bromine oxidation to yield D-glucaro-1,5-lactone. Bottom spectrum is a spectrum of the commercially available, D-glucaro-1,4-lactone.

4.2.7 DNA Binding Assays (Dr. Irina Rodionova, SB MRI)

The interaction of the purified recombinant GguR and GudR regulator proteins with their cognate DNA-binding sites in *R. picketti* 12J and *Polaromonas* sp. JS666 was assessed using two techniques: electrophoretic mobility shift assay (EMSA) and fluorescence polarization assay (FPA). The oligonucleotides containing the predicted binding sites were synthesized by Integrated DNA Technologies (Table 4.2). The double stranded DNA fragments were obtained by annealing synthesized complementary oligonucleotides at a 1:10 ratio of 5’-labeled with 6-carboxyfluorescein (for FPA) or biotin (for EMSA) to unlabeled complementary oligonucleotides.
Table 4.2 Oligonucleotides used in the electrophoretic mobility shift assay and the fluorescence polarization assay.

Using the FPA assay, we tested two GguR proteins (Rpic_0945 and Bpro_3110) and their cognate DNA binding sites upstream of the Rpic_0946 and Bpro_3109 genes. The 6-carboxyfluorescein-labeled 30-bp DNA fragments (10 nM) were incubated with increasing concentrations of the purified proteins (50–1500 nM) in a total volume of 100 μL of the binding buffer containing 100 mM TRIS (pH 8), 50 mM NaCl, 10 mM MgCl2, 2mM DTT, 5% glycerol, and 0.5 mM EDTA at 25°C for 1 h. Poly(dI-dC) was added to the reaction mixture as a nonspecific competitor DNA at 1 μg to suppress nonspecific binding. The fluorescence-labeled DNA was detected with the FLA-5100 fluorescent image analyzer (Fujifilm). To identify effectors of GguR, additional FPA experiments were performed to test the effect of D-glcA, D-galA, D-gluconate, meso-galactarate and 5-keto-4-deoxy-D-gluconate at concentrations of 2 mM. For effector titration, the protein and DNA fragment were incubated with increasing concentrations of 5-keto-4-deoxy-D-gluconate in the incubation mixture (0.15 to 2.5 mM).
Using the EMSA assay, we tested the GguR (Rpic_0945) and GudR (Bpro_3418) regulators and their cognate DNA sites. The biotin-labeled DNA fragments (0.5 nM) were incubated with increasing concentrations of the recombinant purified proteins in a total volume of 20 μL. The binding buffer contained 50 mM TRris-HCl (pH 8.0), 0.15 M NaCl, 5 mM MgCl₂, 1 mM DDT, 0.05% NP-40, 2.5% glycerol, and 1 μg herring sperm DNA. After 25 min of incubation at 37°C, the reaction mixtures were separated by electrophoresis on a 1.5% agarose gel (30 min, 90 V, room temperature). The DNA was transferred by electrophoresis onto a Hybond-N+ membrane and fixed by UV cross-linking. The biotin-labeled DNA was detected with the LightShift chemiluminescent EMSA kit (Thermo Fisher Scientific Inc, Rockford, IL, USA).

**4.2.8 Differential Scanning Fluorimetry (Nawar Al-Obaidi, AECOM)**

DSF was performed utilizing an Applied Biosystems 7900HT Fast-Realtime PCR system with excitation at 490 nm and emission at 530 nm. Reaction mixtures (20 μl final volume) contained 10 μM protein, 1 mM ligand, 5X Sypro Orange (5000X stock, Invitrogen) in 100 mM HEPES (pH 7.5), 150 mM NaCl. Samples were heated from 22 to 99°C at a rate of 3°C min⁻¹ with each sample in duplicate. Bpro_3101 (TctC) was screened against a 189 compound library. Tₘₕ (midpoint of unfolding) were calculated from fitting the melting curve to a Boltzmann equation. The average of eight control wells (no ligands) was used to calculate the change in Tₘ (ΔTₘ).

**4.2.9 Lactone Hydrolase Activity Screening by ¹H-NMR Spectroscopy**

To determine if Rpic_4446 (UniProt AC B2UIY8, Pfam family SGL) catalyzed the conversion of D-glucaro-1,5-lactone and/or D-galactaro-1,5-lactone to D-glucarate and/or
meso-galactarate, spectra of the reaction mixtures after addition of enzyme were recorded by $^{1}$H-NMR. A 650 μL reaction contained 2.5 mM of either δ-lactone, 1 μL enzyme stock (2 μM), 50 mM sodium phosphate buffer, and 2 mM MgCl$_2$ at pD 6.4 in D$_2$O. All NMR spectra were collected on an Agilent 600MHz spectrometer.

4.2.10 Metal Testing

A continuous polarimetric assay was used to determine the effect of added divalent cations on the activity of B2UIY8. The change in optical rotation was monitored using a polarimeter (Jasco P-1010) and a mercury line filter (405 nm). A 1300 μL reaction containing 100 nM enzyme, 50 mM MES buffer, 2 mM D-galactaro-1,5-lactone, and 2 mM of Ca$^{2+}$, Co$^{2+}$, Mg$^{2+}$, Mn$^{2+}$, Ni$^{2+}$, or Zn$^{2+}$ at pD 6.4 in D$_2$O was monitored for 10 minutes at room temperature.

4.2.11 Lactone Hydrolase Kinetic Assay

A continuous polarimetric assay was used to determine the kinetic parameters of the lactone hydrolase as described above. A 1300 μL reaction containing 100 nM enzyme, 50 mM MES buffer, and 2 mM MnCl$_2$ at pD 6.4 in D$_2$O was monitored for one minute at room temperature with varying substrate concentrations (0.24 mM – 5.85 mM). The rate of the uncatalyzed reaction was determined using the same reaction conditions and monitoring for an additional one minute and 30 seconds. Data were fit to the Michaelis-Menten equation in Sigma Plot using the Enzyme Kinetics Module.

4.2.12 Saturation Difference $^{1}$H-NMR Spectroscopy (Dr. Salehe Ghasempur, UIUC)
To determine if the aldose 1-epimerase family protein is active on either D-glcA or D-galA, saturation difference $^1$H-NMR (SD-NMR) spectroscopy was employed. The enzyme was exchanged into 20 mM sodium phosphate buffer (pD 8.0) in D$_2$O. Reactions (800 μL) containing 20 mM sodium phosphate buffer (pD 8.0), 2 mM D-glcA or D-galA, and 100 mM NaCl were performed in D$_2$O and 5 μL enzyme was added immediately before spectra were recorded. Samples were submitted to the Carl R. Woese Institute for Genomic Biology NMR facility for experimental set up and data collection. All NMR spectra were collected on an Agilent 600 MHz DDR2 spectrometer with OneNMR probe. A spectrum was obtained by saturating the $^1$H-NMR peak associated with the proton attached to C1 of the α-anomer of both hexuronates. The second spectrum was recorded with a 0.12ppm shift in saturation frequency. The difference in the first and second spectra was collected in the absence and presence of enzyme [27, 28].

4.2.13 Gene Expression Analysis (Dr. Salehe Ghasempur, UIUC)

*R. pickettii* 12J was streaked onto NB agar plates and incubated at 30°C overnight. A single colony was picked and used to inoculate a 3 mL liquid culture of nutrient broth and incubated in a tube roller at 30°C overnight. Cells were rinsed at room temperature 3 times with 1X PBS, used to inoculate a 3 mL culture of M9 minimal medium supplemented with 20 mM D-glucose and incubated in a tube roller at 30°C until reaching an OD$_{600}$ equal to 0.8. Cells were rinsed at room temperature 3 times with 1X PBS and used to inoculate 3 mL cultures of M9 minimal medium supplemented with 20 D-glucose or D-galA and incubated in a tube roller at 30°C until they reached an OD$_{600}$ equal to 0.4. RNAProtect Bacteria Reagent (6 mL) from Qiagen was added to each culture, and the mixtures were vortexed then incubated at room
temperature for 5 min. Next, the mixtures were centrifuged at 5000 x g for 10 min, the supernatants were discarded, and RNA was isolated from the cells using the RNeasy Mini Kit (Qiagen) according to the manufacturer’s instructions. RNA preps were treated with RNase-free DNase (Qiagen) to remove genomic DNA. RNA preps were deemed clear of any contaminating genomic DNA when PCRs using RNA preps as template with each primer pair failed to yield a band on an agarose gel. The cDNA was reverse transcribed using ProtoScript First Strand cDNA Synthesis Kit (New England BioLabs) according to the manufacturer’s instructions. The qPCR reaction was set up using FastStart Universal SYBR Green Master (ROX) (Roche) and performed on a LightCycler 480 II (Roche). Fold changes in gene expression were calculated using the formula $2^{(\Delta \Delta C_p \text{ sample} - \Delta C_p \text{ control})}$.

4.3 Results and Discussion

4.3.1 Regulon and Metabolic Pathway Reconstruction

In silico regulon and pathway reconstruction using the TRAP SBPs that hit on D-galA or D-glcA and D-galA revealed three new regulons named (in this study): GguR, GudR, and GulR. Using the TRAP SBP from Polaromonas sp. JS666, Bpro_3107, as the starting point, putative DNA transcription factor binding motifs were identified in Polaromonas sp. JS666: GguR and GulR. The DNA motif for the GguR site was conserved in R. pickettii 12J (Figure 4.5). Pathway reconstruction with proteins encoded in the TRAP SBP containing gene clusters is consistent with these gene clusters encoding enzymes of the known microbial oxidative pathway for D-gala metabolism. In the case where the regulons did not encode δ-lactone isomerase (Gli) and γ-lactone cycloisomerase (Gci) from the known pathway, a previously uncharacterized
subfamily of lactone hydrolases from the SMP-30/gluconolactonase/LRE-like region family (PF08450, SGL), a member of the beta propeller clan (CL0186), was identified. The reconstructed regulons also led us to another previously unknown enzyme from the aldose 1-epimerase family (PF01263, Aldose_epim) that we reasoned likely catalyzed the mutarotation of the α- and β-anomers of D-glcA and D-galA after import or after hydrolysis by a uronidase. Finally, a previously unknown subfamily of tripartite tricarboxylate transporter family receptor (PF03401, TctC) proteins for binding and facilitating the import of a diacid sugar was also encoded within these novel regulons.
Figure 4.5 Transcription factor DNA binding sites conserved across two families of Betaproteobacteria. A) Regulon and pathway reconstruction of members of the Comamondaceae family (Betaproteobacteria Burkholderiales) reveals GguR and GudR regulons for hexuronate catabolism. Note the TctC family member of unknown function (sky blue box) – diacid sugar import candidate. B) Regulon and pathway reconstruction for members of the Ralstoniaceae family (Betaproteobacteria Burkholderiales). GguR DNA binding site is conserved. R. pickettii 12J encodes an additional, different regulon, GulR. Note UxuL (yellow box in pathways) for converting the lactones to the diacid sugars.
4.3.2 Growth and Gene Expression Studies

To determine if *R. pickettii* 12J could use D-galA as a carbon source, cells were incubated at 30°C in the presence of the sugar, and growth was monitored overnight. *R. pickettii* 12J reached comparable growth yields on D-galA compared to growth on 20 mM D-glucose (Figure 4.7).

![GguR regulon](image1.png)
![GulR regulon](image2.png)

<table>
<thead>
<tr>
<th>Locus Tag</th>
<th>Gene Name</th>
<th>Description/Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rpic_0945</td>
<td><em>gguR</em></td>
<td>GntR family regulator</td>
</tr>
<tr>
<td>Rpic_0946</td>
<td><em>gudD</em></td>
<td>D-Glucarate dehydratase</td>
</tr>
<tr>
<td>Rpic_0947</td>
<td><em>exuT</em></td>
<td>MFS hexuronate transporter</td>
</tr>
<tr>
<td>Rpic_0948</td>
<td><em>agl</em></td>
<td>α-Glucuronidase</td>
</tr>
<tr>
<td>Rpic_0949</td>
<td><em>omp</em></td>
<td>Outer membrane porin</td>
</tr>
<tr>
<td>Rpic_0950</td>
<td><em>exuM</em></td>
<td>Hexuronate mutarotase</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Locus Tag</th>
<th>Gene Name</th>
<th>Description/Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rpic_4453</td>
<td><em>gulR</em></td>
<td>LysR family regulator</td>
</tr>
<tr>
<td>Rpic_4452</td>
<td><em>kdgD</em></td>
<td>5-Keto-4-deoxy-D-glucarate dehydratase (decarboxylating)</td>
</tr>
<tr>
<td>Rpci_4451</td>
<td><em>kgsD</em></td>
<td>α-Ketoglutarate semialdehyde dehydrogenase</td>
</tr>
<tr>
<td>Rpic_4450</td>
<td><em>gudP</em></td>
<td>Hexarate permease</td>
</tr>
<tr>
<td>Rpic_4449</td>
<td><em>gudD2</em></td>
<td>D-Glucarate dehydratase</td>
</tr>
<tr>
<td>Rpic_4448</td>
<td><em>garD</em></td>
<td>Galactarate dehydratase</td>
</tr>
<tr>
<td>Rpic_4447</td>
<td><em>udh</em></td>
<td>Hexuronate dehydrogenase</td>
</tr>
<tr>
<td>Rpic_4446</td>
<td><em>uxuL</em></td>
<td>Lactonase</td>
</tr>
<tr>
<td>Rpic_4445</td>
<td><em>exuT2</em></td>
<td>Hexuronate transporter</td>
</tr>
</tbody>
</table>

Figure 4.6 Genome context of genes/genomic organization. Zoomed in view of the genome neighborhoods in *R. pickettii* 12J gene clusters.
To determine if the lactone hydrolase and the mutarotase were induced when *R. pickettii* 12J was grown on D-galA, transcript levels for those genes were compared to transcript levels of those genes when the cells were grown on D-glucose. The lactone hydrolase was induced 512-fold ($2^9$ or a 9 cycle difference between control and experimental by qPCR) suggesting its involvement in the pathway to degrade D-galA, but the mutarotase was not (Figure 4.8). The mutarotase is in a gene cluster with a putative alpha-glucuronidase, so it is possible that the mutarotase plays a role in equilibrating the anomers of D-glCtA because it is likely that only one anomer is released as product from the reaction catalyzed by the glucuronidase. Future work growing the cells on D-glcA will help to clarify this observation.
Figure 4.8 Gene expression results of the exuM gene (mutarotase) and uxuL gene (lactone hydrolase) from cells grown on D-galA as sole carbon source. ExuM is in gene cluster A: GguR regulon. UxuL is in gene cluster B: GulR regulon.

4.3.3 Lactone Hydrolase Activity by $^1$H-NMR Spectroscopy

When Rpic_4446 (UniProt AC B2UIY8, Pfam family SGL) is incubated with either D-glucaro-1,5-lactone or D-galactaro-1,5-lactone, D-glucarate or meso-galactarate is formed, respectively. Figure 4.9 shows the NMR spectrum of the reaction mixture before and after the addition of enzyme. The top spectrum is of a mixture in which the predominant species is D-galactaro-1,5-lactone; minor species include D-galA leftover from the chemical preparation that had not been converted to the lactone form and D-galactaro-1,4-lactone that forms from the uncatalyzed, intramolecular isomerization of D-galactaro-1,5-lactone. The middle and bottom spectra are of meso-galactarate. The numbering of the compounds on the right refers to the
hydrogen atoms (not shown) attached the numbered carbon. These results suggest that the enzyme is converting the lactone to the linear form of the sugar. A comparable reaction occurs when the enzyme is incubated with D-glucara-1,5-lactone; the conversion to D-glucarate is observed (data not shown).

![Figure 4.9](image)

Figure 4.9 A) Galactaro-1,5-lactone. B) In the presence of putative lactone hydrolase. C) Standard for meso-galactarate. Without the presence of enzyme, D-galactaro-1,5-lactone would slowly convert to D-galactaro-1,4-lactone [9]. Only in the presence of enzyme is the presence of meso-galactarate observed.

4.3.4 Metal Testing

To determine to requirement of the enzyme for a divalent metal cation, a continuous polarimetric assay was used and reactions were carried out in the presence of the following
metals: $\text{Ca}^{2+}$, $\text{Co}^{2+}$, $\text{Mg}^{2+}$, $\text{Mn}^{2+}$, $\text{Ni}^{2+}$, or $\text{Zn}^{2+}$. Zinc was the best activator for the enzyme (Figure 4.10).

Figure 4.10 Progress curves of the conversion of the lactone to the linear meso-galactarate in the presence of different metals. Reaction progress was monitored using a polarimeter. Zinc was the best activator for this enzyme.

4.3.5 Lactone Hydrolase Kinetic Assay

A continuous polarimetric assay was used to determine the kinetic parameters of the lactone hydrolase. Figure 4.11 shows the progress of the uncatalyzed isomerization of D-galactaro-1,5-lactone to D-galactaro-1,4-lactone in the absence of enzyme (black line). The red line shows the progress of the reaction in the presence of enzyme; D-galactaro-1,5-lactone is converted to the optically inactive, meso-galactarate (red line goes to zero). These data further
support Rpic\textunderscore4446 as a lactone hydrolase. A similar procedure was used to determine the kinetic constants for this enzyme with both hexuronates (Table 4.3). The $K_M$s for both substrates are comparable (around 1 mM) and the $k_{cat}/K_M$ is at the expected values for members of this family.

Figure 4.11 Reaction progress curves measured using a polarimeter. In the absence of enzyme (black line), D-galactaro-1,5-lactone isomerizes to D-galactaro-1,4-lactone. In the presence of enzyme, the optically inactive meso-galactarate is formed (red line).
Table 4.3 Kinetic constants for Rpic_4446. The $K_M$s for both substrates are comparable (around 1 mM) and the $k_{cat}/K_M$ is at the expected values for members of this family.

<table>
<thead>
<tr>
<th>Substrate</th>
<th>$k_{cat}$ (s$^{-1}$)</th>
<th>$K_M$ (mM)</th>
<th>$k_{cat}/K_M$ (M$^{-1}$ s$^{-1}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glucaro-1,5-lactone</td>
<td>310 ± 30</td>
<td>0.88 ± 0.3</td>
<td>3.5 x 10$^5$</td>
</tr>
<tr>
<td>Galactaro-1,5-lactone</td>
<td>720 ± 60</td>
<td>1.1 ± 0.3</td>
<td>6.5 x 10$^5$</td>
</tr>
</tbody>
</table>

Inactive on Glucono- and Galactonolactones

4.3.6 Saturation Difference $^1$H-NMR Spectroscopy

The ability of ExuM to isomerize the alpha and beta anomers of D-glcA and D-galA was determined using saturation difference $^1$H-NMR (Figure 4.12). Data are shown for D-glcA. ExuM interconverted the alpha and beta anomers of D-glcA and D-galA but not D-glucose nor D-galactose. However, the mutarotase was not induced when cells were grown on D-galA as a sole carbon source. The next steps to addressing the physiological role for the mutarotase is to compare the kinetic constants for this enzyme with D-glcA and D-galA as substrates as well as check the gene expression levels when cells are grown on D-glcA.
Figure 4.12 Left panel, top spectrum: resonances associated with the C1 hydrogens in the equilibrated mixture of alpha and beta anomers of D-glca. Left panel, middle spectrum: the signal associated with the alpha anomer was saturated, and then the saturated spectrum was subtracted from left, top spectrum such that the only visible peak is that of the alpha anomer. In the presence of enzyme (left panel, bottom) the resonances associated with the C1 hydrogen of the beta anomer appear because the mutarotation in the presence of enzyme is faster than the relaxation time for the $^1$H nucleus.

4.3.7 Differential Scanning Fluorimetry of TctC (Nawar Al-Obaidi, Albert Einstein College of Medicine)

Bpro_3101 (TctC, 510340) was screened against a 189 compound library containing a large number of sugar compounds including all variants of 3-6 carbon sugar acids (aldonic, aldaric, and uronic acids). The only ligand which yielded significant stabilization to Bpro_3101 was glucarate ($\Delta T_m = 6.1^\circ C$), while an alternate putative ligand, meso-galactarate did not stabilize Bpro_3101 ($\Delta T_m = 0.25^\circ C$). The DSF results suggest that an alternative entry point into
the uronic acid utilization pathway for *Polaromonas* sp. JS666 and related organisms would be through the uptake of D-glucarate by the TctC solute binding protein Bpro_3101 and its related TctC membrane component, though the membrane component is not co-located in *Polaromonas* sp. JS666.

Figure 4.13 Heat stabilization of Bpro_3101 by D-Glucarate (red, $\Delta T_m = 6.1^\circ C$) and not meso-galactarate (blue, $\Delta T_m = 0.25^\circ C$) versus a control reaction (black). Relative fluorescence units (RFUs).

How D-glcA and D-galA were taken up by bacteria other than *E. coli* was largely unknown before our previous investigation of the binding specificities of a diverse set of TRAP SBPs. In this work we used information regarding TRAP SBP ligand specificity as a starting point to interrogate the downstream metabolic pathways for assimilating D-galA in several classes of Proteobacteria. *In silico* regulon and pathway reconstruction revealed a number of putative pathway novelties including new regulator and enzyme subfamilies as well as a new family of diacid sugar transporters (TctC).
This work extends exploration of D-galA assimilation across several classes of Proteobacteria and the reach of the EFI beyond strict enzymes to both regulators and transporters, and demonstrates the utility of the strategies and approaches (e.g., using SBPs to identify starting metabolites in a pathway, SSNs) being developed by the EFI.

4.4 References


CHAPTER 5: CONCLUSION

5.1 Introduction

In the past, sequencing of a gene was directly linked to the characterization of its product. However, over the past 20 years since the first complete bacterial genome was sequenced, functional characterization of a gene product has not benefited from the same improvement in scale as has the determination of a gene’s DNA sequence [1, 2]. Thus the scientific community is inundated with millions of gene sequences, and their subsequent translated protein sequences, that have not been functionally characterized. To realize the full potential of this information, efforts are underway to increase the rate of functional annotation.

One approach with a proven track record for aiding large scale functional annotation is the use of SSNs. SSNs provide a way to visualize relationships between sequence and function across a group of protein sequences. They also aid in identifying those sequences within a protein family whose functions are unknown, and those sequences serve as a good starting point for further investigation. SSNs also enable the ability to transfer function annotations to closely related sequences. Thus SSNs enable a researcher to interrogate sequence space and to generate hypotheses (e.g. what substrates to test, unknown clusters) [3, 4].

5.2 Findings of This Work

This work has enabled those lacking the bioinformatics knowledge and resources (e.g. Unix/Linux, computer cluster) to generate SSNs for their favorite protein family, easily via the
web, through the use of the Enzyme Function Initiative-Enzyme Similarity Tool [4]. For the 14,000 Pfam protein families, a user can generate a network for those families with up to 100,000 sequences (all but 121 families) [5]. Additionally, this work lead to the eventual development of a library of precomputed networks, thus eliminating the time a user would wait while their network was being generated (http://efi.igb.illinois.edu/est-precompute/). SSNs provide the option to go beyond a conventional BLAST list output to a graph that can be visualized in an interactive environment in the software Cytoscape [6].

The use of an SSN generated for a segment of the mechanistically diverse, enolase superfamily (Pfam families PF02746, PF01188, PF13378) enabled us quickly to identify sequence clusters of unknown function as well as to identify the number of members (12) from this superfamily encoded by the genome of the soil bacterium, A. tumefaciens. On the basis of sequence similarity to superfamily members known to dehydrate sugar acids (aldonic, aldaric, and uronic acids), we selected 7 proteins for cloning, purification, and screening against a library of 77 acid sugars. Two of the enzymes were active on substrates in our library screen (1 μM enzyme, 1 mM substrate, overnight at 30°C): best hits for Atu2811 were L-fuconate, 6-deoxy-L-talonate, L-galactonate, L-talonate, and D-ribonate; best hit for Atu4196 was meso-galactarate. Atu3139, Atu0270, Atu4120, Atu1406, Atu5458 did not show complete conversion of any of the substrates in the library.

Atu2811 dehydrated L-fuconate with the following kinetic constants: $k_{cat}$ 1.5 s$^{-1}$, $K_M$ 0.27 mM, and $k_{cat}/K_M$ 5.6 x 10$^3$. Since Atu2811 was part of a cluster of known function, this enzyme was not characterized further. Atu4196 dehydrated meso-galactarate with the following kinetic constants: $k_{cat}$ 0.12 s$^{-1}$, $K_M$ 0.80 mM, and $k_{cat}/K_M$ 1.5 x 10$^3$. Because this enzyme represents a
new family in the enolase superfamily, it was the focus of further characterized by another graduate student [7].

Although we were able to identify activities for two of the seven targets we screened, we did not identify activities for the remaining five. This highlights two limitations of this approach: A) putative orthologues were not investigated and B) the true substrate may be missing from the library. These unknowns are good candidates for future work to try other approaches being developed by the EFI (e.g. large scale gene neighborhood analysis, *in silico* ligand docking to multiple enzymes in the gene neighborhood, screening of knockout mutants for growth deficiencies, use of protein production pipeline to target multiple enzymes from each unknown cluster). This work provides an example for the use of protein SSNs in a shotgun approach to discover enzymes of unknown function in an enzyme family.

Target Atu3139 from *A. tumefaciens* was shown by another group to dehydrate D-galactaro-1,4-lactone to 2-keto-3-deoxy-L-threo-hexarate, a substrate not in our screening assay [8]. Thus we were drawn to the genome neighborhood of Atu3139 because it encoded a member of the enolase superfamily that catalyzed dehydration of a unique substrate. We noticed in the gene neighborhood a member of the amidohydrolase superfamily of unknown function adjacent to Atu3139, and a pathway to which Atu3139 belonged that was missing an enzyme. Thus we used a combination of genomic context, pathway reconstruction, biochemistry, and gene expression analysis to assign the isomerization reaction of D-galactaro-1,5-lactone to D-galactaro-1,4-lactone ($k_{cat}$, $440$ s$^{-1}$; $k_{cat}/K_m$, $8.3 \times 10^4$ M$^{-1}$ s$^{-1}$) to the amidohydrolase of unknown function involved in the oxidative pathway used by *A. tumefaciens* in converting D-galA to central metabolic intermediates [9].
Finally, we used a combination of transporter solute binding proteins, regulon reconstruction, pathway reconstruction, and in vitro biochemistry to interrogate D-galA assimilation in other Proteobacteria. This work lead to the identification of new family of lactone hydrolases from the beta propeller clan that ring opens D-galactaro-1,5-lactone to form meso-galactarate ($k_{\text{cat}}, 720 \text{ s}^{-1}; k_{\text{cat}}/K_m, 6.5 \times 10^5 \text{ M}^{-1} \text{ s}^{-1}$) in a novel pathway variant for assimilating D-galA.

In addition to testing and helping to further develop approaches to large scale functional assignment, these two projects further our understanding about the enzymes microbes use to assimilate D-galA. This information is useful because a number of research groups are interested in engineering microbes to convert D-galA to biofuels and chemical products. D-GalA is the constituent monomer of pectin that constitutes a large percentage of sugar beet pulp, citrus peels, and apple pomace, components of waste streams that often end up in landfills [10, 11]. This work filled in a gap in the known oxidative pathway for assimilating D-galA and identified a variant of the oxidative pathway. Without knowing the enzymes that catalyze the isomerization and lactonization steps in the pathway, could result, for instance, in the buildup of toxic intermediates in engineered strains that lack enzymes catalyzing for these steps. Also, S. cerevisiae is the preferred host to engineer for performing this conversion; however, it does not metabolize D-galA naturally. One limitation to engineering this host is finding enzymes that are soluble in its cytoplasm [11]. So this work that identified different bacteria that contain pathways to assimilate D-galA, helps provide strain engineers with more potential candidates for soluble enzymes.
5.3 Future Lines of Investigation

Future lines of investigation seek to address weaknesses and limitations to the functional annotation problem:

Metabolites – Current metabolite libraries are insufficient in enabling an investigator to interrogate unknown metabolism for two reasons: A) overrepresentation of metabolites with known metabolism and B) missing unknown metabolites. Future work would be to assemble focused libraries of metabolites that lack known pathways for these metabolites, discover and characterize unknown metabolites, and continue to expand the use of virtual libraries to represent those metabolites that are currently too difficult and costly to isolate (e.g. complex hetero-oligosaccharides, oligopeptides).

Proteins – Those protein sequences that do not belong to a protein family and those protein families to which no known function has been assigned (DUFs – domains of unknown function) constitute the most intractable parts of the protein universe. Future work would be to identify the most likely catalytic candidates from this group of sequences and to develop systematic, large scale experimental approaches to interrogate the most unknown territories of the protein universe.

Defining family boundaries – Currently there is no automated way to identify family (evolutionarily related proteins that catalyze the same reaction on the same substrate) boundaries within a superfamily of enzymes on a global scale (i.e. no single alignment score, percent identity, and/or E-value is adequate to use across all superfamilies to segregate protein sequences into isofunctional groups). Thus, identifying an alignment score at which to separate sequences in an SSN is subjective and a result of trial and
error on the part of the experimentalist. To take much of the guess work out of this process, future work in this area would combine the use of SSNs, sequence alignments, phylogenetic trees, and genome neighborhood networks in combination with experimental evidence to develop an algorithm to automate the process of defining family boundaries. The algorithm would be tested against a set of highly characterized protein families. One interesting possibility would be to use SSNs and similarities in genome neighborhoods to determine at what alignment score to output a network. Briefly, genome neighborhood networks would be generated from a sampling of SSNs (e.g. 100 SSNs clustered at 1% to 100% identity at 1% intervals). One would then determine the similarity in genome neighborhoods for each sequence in a cluster of an SSN for all clusters and for all 100 SSNs. The alignment score that yields the most similar genome neighborhoods would presumably be the alignment score at which the SSN clusters are isofunctional. One could even imagine using such an approach to further segregate one cluster in an SSN but leave the others, in a case where not all clusters in an SSN segregate into isofunctional clusters at the same alignment score (e.g. network of multiple Pfam families from the same Pfam clan). Thus, using this approach one may be able to calculate an optimal alignment score at which the genome neighborhoods for the queries are most similar.

Pipeline automation and miniaturization – The EFI has benefited from a cloning and protein production pipeline that makes use of robotics and automation. Future work would be to determine what other areas of the EFI pipeline could benefit from the use of robotics and automation. For instance, the construction of gene knockouts, which seems
fundamental to the functional discovery pipeline, takes considerable time and effort. Using robotics to generate knockout libraries for those organisms that are the most used could free up investigators. For those areas where automation is already being used, the focus would be to see if those processes could be miniaturized (e.g. microfluidics, lab-on-a-chip, and scale down reaction chemistries to reduce reagents).

Solving the functional annotation problem is still in its infancy, and the problem is too big to not require an interdisciplinary approach. This thesis describes a number of combined bioinformatics and experimental approaches that were used to identify previously unknown catalytic activities. Using these methods should increase the rate of assigning functions to enzymes discovered in genome projects. Only when the rate of assigning function, approaches the rate of genome sequencing is the problem solved and the true potential of sequence databases realized.

5.4 References


