Analysis of the Nucleophilic Attack 6-bladed β-Propeller Superfamily Provides Insight into the Evolution of Function in Strictosidine Synthase-like Proteins

Permalink
https://escholarship.org/uc/item/9mc6418q

Author
Hicks, Michael Aaron

Publication Date
2011

Peer reviewed|Thesis/dissertation
Analysis of the Nucleophilic Attack 6-bladed β-Propeller Superfamily Provides Insight into the Evolution of Function in Strictosidine Synthase-like Proteins

by

Michael Aaron Hicks

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Pharmaceutical Sciences and Pharmacogenomics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO
For my teachers, past and present, who have helped instill in me a love of learning that continues to this day, for my colleagues and friends, who make the ups and downs of science (and life) worth it, and for my family, whose love and support have allowed me to pursue my dreams and strive to do great things.
Acknowledgments

Many people need to be acknowledged for their contributions to this work. First and foremost, I thank my Thesis Committee Chair Professor Patricia C. Babbitt, who allowed me to pursue a project of my own creation and who helped me down a path to becoming a true scientist. I thank Alan Barber, a fellow graduate student in the Pharmaceutical Sciences and Pharmacogenomics program and Babbitt lab member, without whom none of the large-scale computational analyses would be possible. I thank also Professor Sarah O’Connor, whose lab helped provide both expertise in strictosidine synthase proteins and biochemical validation of one of our computational predictions. I thank the members of the Babbitt lab, especially those in the wet lab: Dr. Ray Nagatani, Jr., Grant Shackelford, Professor Margaret Glasner, Dr. Sunil Ojha, and Stephanie Lucas. Though the experimental system proved more difficult to work with than I could have imagined a priori, the wet lab group always provided the support and ideas I needed to tackle these tough issues. To the rest of the Babbitt lab, but especially Florian Lauck, Dr. Alex Schnoes, Dr. Daniel Almonacid, Dr. Susan Mashiyama, Dr. Shoshana Brown, Dr. Ben Polacco, Dr. Elaine Meng and Alan Barber, thank you for all you’ve done in helping me to develop many of the ideas presented in this work and for creating a great working environment.

During my time at UCSF, I was fortunate enough to participate in the Students Modeling a Research Topic (SMART) program through the Science and Health Education Partnership (SEP), where I worked with junior and senior high school students at Galileo
High School in a year-long collaboration involving my thesis research. I would like to acknowledge the SMART program advisors, Sabine Jeske and Ben Koo, the high school teacher with whom I worked, Stan DeBella and his students. Having the ability to teach students about my project and having them, in turn, go to a conference to teach the scientific community about what they learned gave me great satisfaction and provided me a much needed perspective about how far I have come in science.

I also thank my graduate program, especially Professor Frank Szoka and Professor Deanna Kroetz, and the program administrator, Debbie Acoba. Finally, I thank the University of California, San Francisco for helping to cultivate one of the best places possible both to learn and conduct scientific research.

Note to the Reader

A modified version of Chapter 2 has been submitted for publication in an upcoming book entitled Proteins Families: Relating Protein Sequence, Structure and Function edited by Professor Christine Orengo and Dr. Alex Bateman. The manuscript is titled “The Nucleophilic Attack 6-bladed β-Propeller (N6P) Superfamily” by Michael A. Hicks, Alan E. Barber II, and Patricia C. Babbitt. The introduction to the publication, written by Patricia C. Babbitt, has been removed from this chapter, and small modifications have been made throughout the chapter for clarity and consistency with the rest of the dissertation. Alan E. Barber II helped generate the similarity networks and created a large
scale alignment in PROMALS\textsuperscript{1}. Michael A. Hicks performed the analysis and wrote the manuscript with Patricia C. Babbitt, who supervised and directed the work.

A modified version of Chapter 3 has been published in the journal *Proteins: Structure, Function and Bioinformatics* and is available online here:

http://onlinelibrary.wiley.com/doi/10.1002/prot.23135/abstract. The article is titled “The Evolution of Function in Strictosidine Synthase-like Proteins” by Michael A. Hicks, Alan E. Barber II, Lesley-Ann Giddings, Jenna Caldwell, Sarah E. O’Connor, and Patricia C Babbitt. Lesley-Ann Giddings, Jenna Caldwell, and Sarah E. O’Connor did the wet lab experiments, created the corresponding figures, wrote the experimental portion of the methods section, and contributed a few sentences detailing the experimental results. Alan E. Barber II helped create the similarity networks presented in the chapter. Michael A. Hicks performed the computational analysis and wrote the manuscript with Patricia C. Babbitt, who supervised and directed the work.
Proteins function in many biochemical processes essential to life. Our ability to understand how these functions relate to both the sequence and structure of proteins are paramount to understanding the underpinnings of myriad diseases and indeed life itself. Unfortunately, our understanding of this relationship can, at times, be quite limited. Subtle variations in sequence and structure can lead to quite different function. However, computational methods, paired with biochemical characterization, can help better define these complex relationships.

One approach to achieve this goal is through an analysis of a protein’s superfamily, defined as a group of evolutionarily related proteins which conserve some aspect of function, even though the overall functions of individual proteins may be quite distinct. By comparing evolutionarily related proteins, new insights into the evolution of function, and by proxy, the relationship between a proteins’ sequence, structure, and function can be made.

Using a variety of computational tools, large-scale analyses of protein sequences and structures of the six-bladed β-propeller fold class were conducted. These studies identified more than 2500 sequences belonging to a superfamily of enzymes termed the Nucleophilic Attack 6-bladed β-Propeller (N6P) superfamily, in which an apparent
mechanistic similarity involving nucleophilic attack appeared to be conserved. This superfamily was further classified into three subgroups: the arylesterase-like, the senescence marker protein-30/gluconolactonase/luciferin-regenerating enzyme-like (SGL), and the strictosidine synthase-like (SSL) subgroups. The first two subgroups had previously been shown to catalyze hydrolytic reactions whereas the SSL subgroup had only been shown to catalyze a condensation reaction.

Interestingly, most SSL proteins differed from known strictosidine synthase (SS) enzymes in that they contained metal-coordinating active site ligands common to the rest of the superfamily but which were missing in SS. This insight allowed the identification of nearly 500 sequences annotated as “strictosidine synthase” as functionally misannotated proteins. Furthermore, function predictions based on the superfamily context were made for several hundred of these SSL proteins, and in one instance was biochemically confirmed by a collaborating laboratory. Unfortunately, many SSL proteins proved difficult to work with in the wet lab and as such, many more questions remain regarding the reaction specificity of these proteins.
Table of Contents

PREFACE

Copyright ........................................................................................................................................ ii
Dedication ...................................................................................................................................... iii
Acknowledgements ................................................................................................................... iv
Note to the Reader ..................................................................................................................... v
Abstract ......................................................................................................................................... vii
Table of Contents ...................................................................................................................... ix
List of Tables .................................................................................................................................... xiii
List of Figures .................................................................................................................................... xiv

CHAPTER 1: An Introduction for a General Audience................................................................. 1

1.1 Proteins – Sequence, Structure and Function .................................................................... 2
1.2 The Superfamily/Subgroup/Family Hierarchy ................................................................. 3
1.3 The Value of Classification ............................................................................................... 5

INTRODUCTION TO CHAPTER 2 .......................................................................................... 7

CHAPTER 2: Sequence/Structure/Function Relationships in the Nucleophilic Attack 6-bladed β-Propeller (N6P) Superfamily .............................................................................. 8

2.1 Classification of Proteins in the 6-Bladed β-Propeller Fold Class and the N6P Superfamily ......................................................................................................................... 8
2.2 Definition of the N6P Superfamily ..................................................................................... 12
2.3 What We Know and Don't Know about the Enzymes of the N6P Superfamily .....14

I. The Senescence marker protein-30/Gluconolactonase/Luciferin-regenerating enzyme-like (SGL) Subgroup ..................................................................................14

II. The Arylesterase-like Subgroup ...........................................................................17

III. The Strictosidine Synthase-like (SSL) Subgroup ..............................................19

2.4 Functional Predictions and Prediction of Misannotation of SSL Subgroup Enzymes ......................................................................................................................21

2.5 Do the Strictosidine Synthases Really Belong to the N6P Superfamily? .............24

2.6 Re-examination of the Boundaries of the N6P Superfamily in the Context of other β-Propeller fold Proteins .........................................................................................28

2.7 Using the Superfamily Context to Select Protein Targets for Experimental Characterization ..................................................................................................................33

2.8 Conclusion .............................................................................................................36

2.9 Access to Data from this Work ............................................................................37

2.10 Acknowledgements ..............................................................................................38

INTRODUCTION TO CHAPTER 3 ..................................................................................39

CHAPTER 3: The Evolution of Function in Strictosidine Synthase-like Proteins....40

3.1 Abstract ..................................................................................................................40

3.2 Introduction .............................................................................................................42

3.3 Materials and Methods ..........................................................................................48

I. Data Set Sources and Curation ............................................................................48

II. Generating Sequence Similarity Networks .........................................................49
III. Structure-guided Sequence Alignments .................................................................50
IV. Gene Context Analysis ............................................................................................51
V. Phylogenetic Tree .....................................................................................................51
VI. Structure-Function Linkage Database ...................................................................52
VII. General Methods and Analytical Techniques ......................................................52
VIII. Cloning and Protein Expression .........................................................................53
IX. Protein Purification .................................................................................................54
X. Hydrolase Activity with p-Nitrophenyl Acetate ......................................................55
XI. Pictet-Spenglerase Activity ....................................................................................56

3.4 Results and Discussion ..........................................................................................56

I. Similarities and Differences among SSL, SGL, and Arylesterase-like Subgroup Proteins are Complicated and Suggest that Most SSL Proteins do not Catalyze the SS Reaction ..................................................................................................................................................................................57

II. SS may have Evolved from an Ancestor with Metal-Coordinating Active Site Residues ..............................................................................................................................................................................................................................................67

III. Biological Information Provides Clues about Functional Properties of Some SSL Subgroup Proteins ..............................................................................................................................................................................................................................................70

IV. Comparison of Liganded SGL and SSL Subgroup Structures Reveals Intriguing Similarities in Active Site Mechanisms Despite the Differences between their Divergent Reactions and Active Sites ..............................................................................................................................................................................................................................................72

3.5 Summary ..................................................................................................................76

3.6 Access to Data from this Work ...............................................................................77

3.7 Acknowledgements ..................................................................................................77
CHAPTER 4: Experimental Progress in the Expression and Purification of
Strictosidine Synthase-like Proteins ...............................................................78

CONCLUDING REMARKS ............................................................................87

BIBLIOGRAPHY ..........................................................................................89

APPENDIX ..................................................................................................96

A1. Supplementary Material from Chapter 3: The Evolution of Function in
Strictosidine Synthase-like Proteins ..............................................................96

   I. Supplementary Figures from Chapter 3 ......................................................96

   II. Supplementary Tables from Chapter 3 .....................................................98

A2. Screenshots of N6P SF Data in the Structure-Function Linkage Database
(SFLD) ........................................................................................................99

PUBLISHING AGREEMENT .....................................................................103
List of Tables

Table 3.1. Experimentally Characterized Activities of Selected N6P SF Enzymes........44

Table 3.S1. Comparable Positions of Reactive Groups and Ligands in Active Sites for SS
and DFPase Mechanisms ........................................................................................................98

Table 3.S2. Pairwise Structural Comparisons for Proteins Described in the Text ..........98
List of Figures

Figure 2.1. Structure-Based Similarity Network of Crystal Structures Sharing the Six-Bladed β-Propeller Fold..........................................................10
Figure 2.2. Examples of Chemical Reactions Catalyzed by the N6P SF .................14
Figure 2.3. Sequence Similarity Network for the SGL Subgroup Downloaded from the Current SFLD ..................................................................................15
Figure 2.4. Sequence Similarity Network for the Arylesterase-like Subgroup Downloaded from the SFLD .................................................................18
Figure 2.5. Sequence Similarity Network for the Strictosidine Synthase-like (SSL) Subgroup Downloaded from the SFLD ..............................................20
Figure 2.6. Sequence Similarity Network Showing Conservation of Metal Coordinating Active Site Residues for 2567 Proteins of the N6P SF .........................................................23
Figure 2.7. Active Site Superposition of (a) SGL and Arylesterase-like Subgroup Proteins and (b) SGL and Arylesterase-like Subgroup Proteins with SS ...........................................25
Figure 2.8. A Related Catalytic Strategy Unites the SS Enzymes with the Rest of the N6P SF ........................................................................................................27
Figure 2.9. Sequence Similarity Network of those Six-Bladed β-Propeller Proteins from the Largest Cluster of the Structure Similarity Network in Figure 2.2 ..................31
Figure 2.10. Sequence Similarity Network of 393 Proteins of the Strictosidine Synthase-like (SSL) Subgroup Downloaded from the SFLD, Shown as in Figure 2.5, but Edges are Drawn if the Pairwise Similarity Score between Two Nodes is less than $1 \times 10^{-80}$ (More Stringent Cutoff) ................................................................................35
Figure 2.11. Gene Context for an Uncharacterized SGL Subgroup Member from Burkholderia sp. H160 as Seen on the Integrated Microbial Genomes (IMG) System.

Figure 3.1. Examples of Chemical Reactions Catalyzed by the N6P SF.

Figure 3.2. Sequence Similarity Network for the N6P SF.

Figure 3.3. Active Site Superposition of (a) SGL and Arylesterase-like Subgroup Proteins and (b) SGL and Arylesterase-like Subgroup Proteins with SS.

Figure 3.4. Sequence Similarity Network Showing Conservation of Metal Binding Residues for the 516 SSL Subgroup Sequences Generated from All-by-all Pairwise Comparisons.

Figure 3.5. Active Site Motif Alignments of Nine SSL Subgroup Proteins with Structures from the Arylesterase-like, SGL and SSL subgroups.

Figure 3.6. Bayesian Phylogenetic Tree of a Representative Subset of SSL Subgroup.

Figure 3.7. Sequence Similarity Network Showing Types of Life for the 516 SSL Subgroup Sequences Generated from All-by-all Pairwise Comparisons.

Figure 3.8. Gene Neighborhood for SSL Proteins in Gram Negative Bacteria as Seen on the Integrated Microbial Genomes (IMG) System.

Figure 3.9. A Related Catalytic Strategy Unites the SS Enzymes with the Rest of the N6P SF.

Figure 3.S1. (a) The Active Site of PON1 and (b) The Activation of Water by the His-His Dyad and the Nucleophilic Attack on an Ester Carbonyl.

Figure 3.S2. V. vinifera SSL (gi|147772032) Exhibited Hydrolase Activity with the Model Substrate pNPAc Compared to C. roseus SS, and Eluent from Cultures Expressing the Empty Vectors pET32b and pET28a.
Figure 3.S3. LC-MS Data Show V. vinifera SSL (gi|147772032) Does Not Catalyze the SS Reaction

Figure 4.1. Expression Analysis and Purification Attempt of Sequence 6 (gi|15596490; from Pseudomonas aeruginosa PAO1 and Annotated as ‘hypothetical protein PA1293’ in Genbank) in the pHis-Gβ1 Construct Carried out with Stephanie Lucas on February 12 and 13, 2009

Figure 4.2. Expression Analysis and Purification Attempt of Sequence 6 Truncated (gi|15596490; from Pseudomonas aeruginosa PAO1 and Annotated as ‘hypothetical protein PA1293’ in Genbank with a Predicted Signal Peptide Removed) in the pHis-Gβ1 from April 24, 2009

Figure 4.3. Human APMAP Expression and Purification Experiment in Pichia pastoris from March 11 to March 19, 2010

Figure S1. Screenshot of the N6P Superfamily Page in the SFLD

Figure S2. Screenshot of the N6P SF Alignment

Figure S3. Screenshot of the SS Family Page
Chapter 1: An Introduction for a General Audience
“Biology will relate every human gene to the genes of other animals and bacteria, to this great chain of being.”

Walter Gilbert

1.1 Proteins – Sequence, Structure, and Function

Proteins, encoded by genes, are essential to life. They function in a variety of biological roles including the production of energy for biochemical processes, the defense against foreign invaders by acting as antibodies, and the interpretation of the outside world through the process of signal transduction. The study of proteins is therefore essential to our understanding of the underpinnings of life. There are a variety of ways to study proteins, but the focus of this dissertation is on using information gained from genome sequencing projects to learn about how a protein’s function relates to its amino acid sequence and three-dimensional structure. In this light, Walter Gilbert’s quote is quite relevant, as the essence of the methodology used in this line of research relies on detecting similarities between genes (or in this case, their protein products) across all types of life.

One of the biggest challenges we face is an incomplete understanding of the relationship between a protein’s sequence, its structure, and its function. At the most fundamental level, proteins are composed of amino acids, which can be represented as a string of letters known as an amino acid sequence. These sequences can be derived from the genes encoded on the DNA sequence of a given organism. These amino acid sequences fold into a three-dimensional structure which helps determine its biological function.
However, the relationship between sequence, structure and function is not straightforward. Subtle variations in sequence and structure can lead to quite different functions. Conversely, very diverse amino acid sequences and three-dimensional structures can sometimes lead to the same function.

With the elucidation of thousands of genomes, a plethora of information is now available. Similar genes (and their corresponding protein products) from one organism can be compared to the same genes (and the same protein products) from many other organisms. Similarity networks provide a means by which to obtain a global view of these relationships. In these networks, each node represents a protein and each edge connecting two nodes represents some measure of similarity above a statistically significant threshold (for a more detailed discussion about networks and their applications, see Ref. 3). By using this information as a starting point, new hypotheses regarding the intricate relationship between protein sequence, structure and function can be made. Our ability to decipher this relationship is paramount to unlocking the potential of the Genomic Age.

1.2 The Superfamily, Subgroup, Family Hierarchy

One way to describe the relationship between similar proteins from many different organisms is through the use of a hierarchical classification scheme, grouping the most similar proteins at the bottom and the most diverse, but still evolutionarily related, at the top. In this section, I describe three classification levels for proteins: the family level, the subgroup level, and the superfamily level.
We define an enzyme family as a group of evolutionarily related enzymes that catalyze the same overall chemical transformation. For example, the strictosidine synthases, as the name indicates, all catalyze the synthesis of strictosidine from tryptamine and secologanin. As will be described in detail later, they do so through the use of a conserved piece of structure; namely, a glutamic acid found in all strictosidine synthases.\(^4\)

A subgroup is defined more broadly and can take into account additional information specific to a given biological system. It can contain one or more related enzyme families, which may or may not catalyze similar reactions. In the case of the strictosidine synthases, the family belongs to a larger subgroup of strictosidine synthase-like enzymes that do not conserve the glutamic acid found at the family level in the strictosidine synthases, but they do share a high level of sequence similarity with each other. This strictosidine synthase-like subgroup will be discussed in detail in Chapter 3.

The broadest relationship between evolutionarily related enzymes, as described in this dissertation, is the superfamily level. A superfamily may have multiple subgroups, each of which may have multiple families. Members of a superfamily can catalyze quite different overall reactions, but each enzyme catalyzes a similar partial chemical transformation that unifies these related proteins into a single superfamily. For the Nucleophilic Attack 6-bladed β-propeller (N6P) superfamily, to which strictosidine synthases belong, the common partial reaction is a nucleophilic attack, which will be discussed in detail in Chapter 3; however, it should be noted that despite this common
partial reaction, the overall reactions in different subgroups within this N6P superfamily are quite distinct.

### 1.3 The Value of Classification

The value of this hierarchical classification scheme is manifold. For example, this information provides defined relationships between related proteins, and gives experimentalists a “gold-standard” set of data from which to generate hypotheses regarding the functions of uncharacterized proteins. For the N6P superfamily, these data have been put in the Structure-Function Linkage Database (SFLD) (http://sfld.rbvi.ucsf.edu//; Appendix A2), giving users access to a variety of useful resources, including evidence codes, which let users know the amount of confidence they should have in a given protein’s function annotation. Additionally, this high quality data set seeks to correct function misannotation problems, which can lead experimentalists down the wrong research path, wasting time and money (for a detailed analysis of the misannotation issue, see Ref. 5).

Additionally, this organizational framework provides important information for enzyme engineers trying to generate new and useful functions (for recent reviews on protein engineering and the use of sequence-structure-function information, see Refs. 6,7). By defining an enzyme’s function as a series of chemical steps rather than an overall chemical transformation, we are able to link a given chemical step to an evolutionarily conserved piece of structure. When another protein sharing the same fold and the same
piece of evolutionarily conserved structure catalyzes a different overall reaction, it suggests that this reaction may use a similar partial chemical transformation to the first. If this is true, we learn something about how Nature has evolved different reactions using the same fundamental partial reaction. It also suggests something about the potential overall chemical transformations a given protein fold may “know” how to do; that is, a partial chemical reaction may be “hard-wired” into a given protein scaffold. Enzyme engineers can use this information to create enzymes with new functions. For example, a single substitution in two different members of the enolase superfamily (L-Ala-D/L-Glu epimerase (AEE) and muconate lactonizing enzyme (MLE)) permitted the catalysis of a third superfamily reaction (the o-succinyl-benzoate synthase (OSBS) reaction); prior to the substitution, neither protein could catalyze the OSBS reaction. However, our understanding of Nature’s design rules is limited in many cases, providing a barrier to engineering efforts. As we progress in our understanding, we may be able to use this information to help in biofuel and pharmaceutical production, as well as bioremediation technology (for a review of current progress in these fields, see Ref. 9).
Introduction to Chapter 2

This chapter introduces the Nucleophilic Attack 6-bladed β-Propeller (N6P) Superfamily and discusses in detail how this group of 6-Bladed β-Propeller proteins is distinguished from others in the fold class. Additionally, the three subgroups of the superfamily are introduced, and the challenges of rationalizing the inclusion of strictosidine synthases in the superfamily context are detailed. Finally, as introduced in Chapter 1.3, the use of the superfamily context to select protein targets for experimental characterization is established.

This chapter is derived from a recently submitted manuscript for publication in an upcoming book entitled Proteins Families: Relating Protein Sequence, Structure and Function edited by Professor Christine Orengo and Dr. Alex Bateman. The submitted manuscript is titled “The Nucleophilic Attack 6-bladed β-Propeller (N6P) Superfamily” by Michael A. Hicks, Alan E. Barber II, and Patricia C. Babbitt. The introduction to the submitted manuscript, written by Patricia C. Babbitt, has been removed from this chapter, and small modifications have been made throughout the chapter for clarity and consistency with the rest of the dissertation. Alan E. Barber II helped generate the similarity networks and created a large scale alignment in PROMALS\(^1\). Michael A. Hicks performed the analysis and wrote the manuscript with Patricia C. Babbitt, who supervised and directed the work.
Chapter 2: Sequence/Structure/Function
Relationships in the Nucleophilic Attack 6-bladed
β-Propeller (N6P) Superfamily
2.1 Classification of Proteins in the 6-Bladed $\beta$-Propeller Fold Class and the N6P Superfamily

In this section, we use major protein classification resources (CATH$^{10}$, SCOP$^{11}$ and Pfam$^{12}$) as a framework to provide a context for our definition of the N6P SF within the larger scope of the 6-bladed $\beta$-propeller fold.

Examination of similarity relationships between a set of 41 non-redundant crystal structures of six-bladed $\beta$-propeller fold proteins (Figure 2.1), identified one major cluster as well as several smaller clusters of related proteins. Mapping of functional information designating major functional classes (defined as Pfam families$^{12}$) onto this structure similarity network shows that structures of known functional groups cluster together. Additionally, this view allows us to begin to see connections between proteins of different functions. In the major cluster of 18 protein structures, 8 families are represented: glucose-sorbosone dehydrogenase (GSDH), SMP-30/Gluconolacanase/LRE-like region (SGL), arylesterase, strictosidine synthase, major royal jelly protein (MRJP), NHL, PD40, and low-density lipoprotein repeat class B (Ldl_recept_b).
Figure 2.1. Structure-Based Similarity Network of Crystal Structures Sharing the Six-Bladed β-Propeller Fold. Forty-one non-redundant structures were compared pairwise using the algorithm TM-align. The structural representatives of the N6P SF are boxed. Structures are colored by Pfam family as shown. Each node represents a structure and is labeled with its PDB ID. Edges are drawn between nodes only if the alignment score is >0.7, a score considered statistically significant by the authors of TM-align. For this network, the median alignment length is 233 residues and the median RMSD is 2.75 Å.

Classification of this fold class and of the proteins of the N6P SF are generally similar among the three comprehensive structure and sequence resources (SCOP, CATH, and Pfam) and are generally consistent with those shown in Figure 2.1. Of the structures represented in this Figure, 61% are classified in SCOP v1.75 and 78% in CATH v3.4.0.
Among these, several belong to an identical homologous superfamily (SF) in CATH: the TolB, C-terminal domain SF (2.120.10.30). These include proteins from the SGL (1pjx.pdb, 2ghs.pdb, 2p4o.pdb), arylesterase (1v04.pdb), MRJP (2qe8.pdb), NHL (1rwi.pdb, 1q7f.pdb), PD40 (2ojh.pdb) and LDL-recept_b (1npe.pdb, 1ijq.pdb) Pfam families. CATH also includes proteins not in this central cluster as belonging to the same homologous SF, notably two additional PD40 proteins (1crz.pdb, 2hqs.pdb), an additional LDL-recept_b (1n7d.pdb), and a GSDH protein (1cru.pdb) that does not connect into our central cluster at this threshold (though three other pdb structures that are not currently classified in CATH do connect). Other proteins in the network are classified into different homologous SFs in CATH: the neuraminidases and sialidases (2.120.10.10) and the phytases (2.120.10.20). Strictosidine synthase (str_synth) and the six-bladed β-propeller domain of the folate receptor (Folate_rec) are not yet classified in CATH.

The classification of SFs by SCOP is mostly in agreement with CATH, though there are key differences. Like CATH, SCOP classifies the SGL (2dg1.pdb, 1pjx.pdb, 2ghs.pdb, 2p4o.pdb) and arylesterase (1v04.pdb) structures in the same SF (b.68.6). However, SCOP differs from CATH in that NHL (1rwi.pdb, 1q7f.pdb) and Ldl_recept_b (1npe.pdb, 1ijq.pdb) proteins belong to two distinct SFs (b.68.9 and b.68.5, respectively).

The definition of the N6P SF that is used in the SFLD differs from the SF definition used by SCOP, CATH and Pfam principally because the SFLD adds criteria associated with conserved catalytic strategies to refine further definitions at the SF level. In some SFs,
such strategies can be associated with specific features conserved in members’ active sites as well. Our definition of the N6P SF is most similar to the SCOP definition, but includes strictosidine synthase (neither CATH nor SCOP has yet classified the strictosidine synthase structures (2fp8.pdb in the network)) and the subgroup we describe as the strictosidine synthase-like (SSL) proteins. Consistent with our definition, in this Chapter we evaluate the boundaries of the N6P SF with respect to its nearest neighbors in sequence, structure, and features of the chemical strategy by which catalysis is enabled. Below, the N6P SF and its constituent subgroups are defined as they are classified in the SFLD.

2.2 Definition of the N6P Superfamily

The 2,500 members of the N6P SF can be distinguished from the other proteins of the six-bladed β-propeller fold class (~36,000 total members for the families listed in Figure 2.1 from Pfam 25.0) by sequence and structure similarity and by similarity in the fundamental chemical strategy by which catalysis is enabled. A small number of proteins within the SF, members of the SFLD-defined strictosidine synthase (SS) family, differ in additional features that distinguish them from the rest of the SF as well. These are described in detail below. First, proteins in the three subgroups that comprise the N6P SF share higher overall sequence similarity with each other than they do with any other protein sharing the same fold. Second, despite the quite different overall reactions represented by the few members of the SF that have been biochemically characterized, all of the members of the N6P SF can be united by a common catalytic step i.e., nucleophilic
attack on an electrophilic substrate (Figure 2.2). Finally, with the exception of the small number of enzymes that include those biochemically characterized (or predicted) to catalyze the SS reaction, nearly all of the members of the SF appear to be metal dependent, sharing in their active site architectures four highly conserved metal binding residues that coordinate to a divalent metal. These conserved metal binding ligands can be identified in multiple sequence alignments representing the huge majority of the 2500 members of the N6P SF (not shown). Although other members of the fold class are metal dependent, the particular metal dependent binding modality found in the N6P SF distinguishes it from all of these other proteins.

The members of the N6P SF are represented by three highly diverse subgroups, designated by the different colors within the rounded rectangle shown in Figure 2.1. Each of these subgroups was previously identified as part of the same SF by sequence similarity and by one of the known reactions its members catalyze\textsuperscript{14}, some of which are depicted in Figure 2.2. A more detailed description of structure-function relationships among these subgroups is given below.
Figure 2.2. Examples of Chemical Reactions Catalyzed by the N6P SF. Red asterisk indicates electrophilic atom that is attacked. (a) SS reaction. (b) Examples of lactonase, esterase and phosphotriesterase reactions catalyzed by members of the SGL and arylesterase-like subgroups. Figure and Legend are adapted from 15.

2.3 What We Know and Don't Know about the Enzymes of the N6P Superfamily

I. The Senescence marker protein-30/Gluconolactonase/Luciferin-regenerating enzyme-like (SGL) Subgroup

The Senescence marker protein-30/Gluconolactonase/Luciferin-regenerating enzyme-like (SGL) subgroup includes about 1500 members (Figure 2.3), several of which are known to catalyze lactonase, esterase and organophosphatase activities (Figure 2.2b).

Senescence marker protein-30 (SMP-30), also known as regucalcin, functions as a lactonase in the biosynthesis of L-ascorbic acid in non-primate mammals16 and has been shown to hydrolyze toxic organophosphates in mouse liver17,18. Luciferin-regenerating enzyme (LRE), as the name suggests, catalyzes the regeneration of luciferin in fireflies
through hydrolysis of oxyluciferin to 2-cyano-6-hydroxybenzothiazole, which then is condensed with D-cysteine\(^1\). Additional biochemically characterized proteins from this subgroup include drug-responsive protein 35 (Drp35) which functions in the resistance to antibiotics by \textit{Staphylococcus aureus}\(^{20,21}\) and diisopropylfluorophosphatase (DFPase)\(^{22}\), whose native activity is unknown, but which can break down organophosphates, including nerve toxins such as soman\(^{23}\).

Figure 2.3. Sequence Similarity Network for the SGL Subgroup Downloaded from the Current SFLD. The 1449 proteins of the SGL subgroup of the N6P SF are shown.
Sequence similarity was computed from pairwise BLAST comparisons of all of the members of the subgroup using the SFLD as a background model. Nodes are colored by type of life as designated in the key. Small circular nodes represent biochemically uncharacterized proteins; large diamond-shaped nodes with white outlines indicate biochemically characterized proteins. Edges are drawn if the pairwise similarity score between two nodes is less than $1 \times 10^{-60}$. Arrow indicates the Burholderia sp. H160 discussed in the text.

A conserved set of metal-coordinating ligands is present in the active site of all structurally characterized proteins of the SGL subgroup. Detailed mechanistic studies done on the DFPase enzyme have implicated these residues as critical for its organophosphatase activity$^{24-27}$. In the proposed DFPase mechanism, an aspartic acid coordinated to a divalent metal performs a direct nucleophilic attack on the phosphorous atom of the phosphoryl group of DFP forming a phosphoenzyme intermediate. Water then completes the hydrolytic reaction through regeneration of the aspartic acid, as demonstrated by $\text{H}_2^{18}\text{O}$ incorporation experiments$^{24}$.

Mutation studies of Drp35 suggest the structurally equivalent metal coordinating ligands are necessary to catalyze the lactonase reaction, although the identity of one of the ligands differs from that of DFPase (Asn120 in DFPase) and is in a structurally equivalent position to Asp138 in Drp35. In this case, it is proposed that Asp138 activates a water molecule, leading to a nucleophilic attack on the carbon atom of the carbonyl group of the lactone. The oxygen atom next to the scissile bond is proposed to be protonated by the other metal-coordinated aspartate$^{21}$.

For other enzymes in this subgroup, the details of their reaction mechanisms are unclear. Though the metal-coordinating ligands likely play a role, additional residues may act as
specificity determinants. Furthermore, the vast majority of proteins in the SGL subgroup are biochemically uncharacterized. Gene context analysis of SGL members from bacteria suggests several proteins likely play roles in sugar metabolism and transport (Hicks MA, Barber AE II, unpublished), though their reaction specificities are also currently unknown.

II. The Arylesterase-like Subgroup

The arylesterase-like subgroup has one defined family in the SFLD: the paraoxonases (which include PON1, PON2 and PON3), with the reaction specificities of the rest of the subgroup classified as unknowns (Figure 2.4). Sequences from this family share about 60% identity and conserve a set of seven residues thought to be important in catalysis. Five coordinate to the catalytic calcium ion (four of which are conserved at the SF level) and the other two, a His-His dyad, appear to be critical for their lactonase and esterase activities. These proteins differ both in their expression patterns and the degree to which they catalyze their respective reactions. In humans, PON1 expression is found in the liver, PON3 is found in both the liver and kidney, and PON2 is found in most tissues, including the heart, liver, kidney, lung, small intestine, spleen, stomach, placenta and testis (for a review, see Ref. 28). PON1 catalyzes lactonase, esterase and organophosphatase reactions though lactonase appears to be its native activity29-31. PON2 and PON3 also appear to be lactonases yet differ from PON1 in that they have very limited arylesterase activity and almost no organophosphatase activity32. Additionally, high density lipoprotein-associated PON1, the best studied member, exerts its anti-
atherogenic properties through inhibition of cell-mediated oxidation of low density lipoproteins and stimulation of cholesterol efflux, likely via its lactonase activity. Furthermore, paraoxonases play a role in acylhomoserine lactone-dependent quorum-sensing systems, helping to prevent microbial infection through their lactonase activity. For example, in *Pseudomonas aeruginosa*, PON1 in serum has been shown to prevent the formation of bacterial biofilm in vitro. Interestingly, PON1-deficient mice show enhanced resistance to *P. aeruginosa* infection, likely due to increased expression of PON2 and PON3.

Figure 2.4. Sequence Similarity Network for the Arylesterase-like Subgroup Downloaded from the SFLD. The 123 proteins of the arylesterase-like subgroup of the N6P SF are shown. Sequence similarity was computed from pairwise BLAST comparisons of all of the members of the subgroup using the SFLD as a background model. Nodes are colored by type of life as designated in the key. Small circular nodes represent biochemically uncharacterized proteins; large diamond-shaped nodes with white outlines indicate
biochemically characterized proteins. Edges are drawn if the pairwise similarity score between two nodes is less than $1 \times 10^{-30}$.

Structural and mutational analyses indicate that the lactonase and esterase reactions catalyzed by the PON enzymes are likely due to a proton-shuttle mechanism involving the activation of a metal-coordinated water by a His-His dyad, leading to a nucleophilic attack on the carbon atom of a carbonyl group$^{35}$. It is believed that the organophosphatase reaction catalyzed by PON1 may be performed in an analogous fashion to DFPase of the SGL subgroup, which involves a direct nucleophilic attack by a metal-coordinated aspartic acid$^{24}$.

As the vast majority of proteins in the arylesterase-like subgroup remain uncharacterized, both biochemically and structurally, other functions likely remain to be discovered.

### III. The Strictosidine Synthase-like (SSL) Subgroup

The SSL subgroup represents a group of nearly 400 proteins and is named for the small number of proteins within the subgroup that have been shown to catalyze the strictosidine synthase (SS) reaction (Figure 2.5). Strictosidine is the precursor to the monoterpenoid indole alkaloid biosynthesis pathway, which produces about 2000 known compounds, several of which are used in the treatment of cancer, malaria, hypertension, schizophrenia

---

$^{1}$Much of the discussion and some of the Figures in the following sections describing structure-function relationships in the SSL subgroup is based on recent work published elsewhere$^{15}$ and are detailed in Chapter 3.
and arrhythmic heart disorders\textsuperscript{36}. The huge majority of the rest of the subgroup, comprising ~400 proteins, are uncharacterized.

Figure 2.5. Sequence Similarity Network for the Strictosidine Synthase-like (SSL) Subgroup. The 393 proteins of the SSL subgroup of the N6P SF are shown. Sequence similarity was computed from pairwise BLAST comparisons of all of the members of the subgroup using the SFLD as a background model. Nodes are colored by type of life as denoted in the key. Small circular nodes represent biochemically uncharacterized proteins; large diamond-shaped nodes with white outlines indicate biochemically characterized proteins. Edges are drawn if the pairwise similarity score between two nodes is less than $1 \times 10^{-50}$. Strictosidine synthases are marked by the oval. Arrows indicate the \textit{Vitis vinifera} SSL and \textit{Homo sapiens} APMAP proteins discussed in the text.
2.4 Functional Predictions and Prediction of Misannotation of SSL Subgroup

Enzymes

SS catalyzes a Pictet-Spengler reaction resulting in the condensation of tryptamine and secologanin to form the strictosidine product. This bi-substrate condensation reaction is quite different from the other known hydrolytic reactions catalyzed by the SGL and arylesterase-like subgroups of the SF (Figure 2.2). And although the condensation reaction catalyzed by SS requires a nucleophilic attack on an electrophilic substrate, which is the fundamental chemistry common to the N6P SF, it differs from most of the other SF members in that it is metal independent and lacks all four metal coordinating ligands common to both the SGL and arylesterase-like subgroups. Thus, the mechanism by which SS catalyzes this fundamental step is quite distinct from the reactions of the SGL and arylesterase-like subgroups. Instead of metal-assisted catalysis, SS uses a functionally required glutamate, Glu309, which activates tryptamine by abstracting a proton, allowing for a nucleophilic attack on a carbon atom of the aldehyde of a second substrate, secologanin, forming an iminium species. The indole then attacks the iminium moiety to form the strictosidine product.

Consistent with these functional and mechanistic differences, both a multiple alignment of the SSL subgroup proteins (Chapter 3) and the sequence similarity network provided in Figure 2.6 indicates that most SSL proteins appear to have the same metal-coordinating ligands common to the SGL and arylesterase-like subgroups. Taken together with the known mechanism of SS, these observations suggest three important
conclusions. First, despite their putative annotations as SSs, these uncharacterized SSL proteins are unlikely to catalyze the SS reaction, as SS proceeds through a metal-independent mechanism. Moreover, steric clashes with metal ligands would likely make productive binding of tryptamine and secologanin impossible in these SSL subgroup proteins. Second, SSL enzymes are more likely to catalyze hydrolytic reactions similar to those of the other two subgroups. Support for this idea is provided by initial experimental studies of a human protein from the SSL subgroup, Adipocyte Plasma Membrane-Associated Protein (APMAP). Those results suggest that APMAP exhibits at least a low level of esterase activity using both beta-naphthyl acetate and phenyl acetate as substrates. Like most other SSL proteins, APMAP possesses all four metal-coordinating ligands (Figure 2.6). Though no additional mutational or mechanistic data for this enzyme is presently available, we predict that these residues play an essential role in catalyzing these esterase reactions, based on analogy to the reaction mechanisms of characterized SGL subgroup members. Third, as most of the proteins in the SSL subgroup have been annotated as “strictosidine synthase” or "strictosidine synthase-like" in public sequence databases such as GenBank, we can conclude that the annotation of nearly 400 proteins in GenBank as “strictosidine synthase” or "strictosidine synthase-like" is misleading to the point that they could be considered to be misannotated.
Figure 2.6. Sequence Similarity Network Showing Conservation of Metal Coordinating Active Site Residues for 2567 Proteins of the N6P SF. As depicted in the key, nodes are colored by the number of metal coordinating active site residues predicted using the multiple sequence alignment described in Figure 2.9. Edges are drawn if the pairwise similarity score has an E-value < 1 x 10^{-10} using UniRef100 as the background model. At this threshold, the median alignment length is 238 residues and the median percent identity is 30.0%. Subgroups of the N6P SF are circled or boxed and labeled. The SS proteins are indicated by a white arrow. (The outlier to the far left of the SSL subgroup is an artifact of the organic layout from the Cytoscape program.[39])

While these observations confirm the classification of the arylesterase-like, SGL, and the huge majority of SSL subgroup proteins as members of a single SF, they also raise a question whether, despite their overall sequence and structural similarity to the rest of the proteins of the N6P SF, the metal independent SSs should be included in the SF as well.
We argue that the preponderance of evidence supports the inclusion of the SSs in the N6P SF, as discussed below.

2.5 Do theStrictosidine Synthases Really Belong to the N6P Superfamily?

As is clear from our analysis of the SSL subgroup, complex variations in sequence among seemingly related proteins can complicate function annotation transfer from biochemically characterized to uncharacterized proteins, resulting in misannotation of function for many members of this subgroup. Moreover, defining what to include in a functionally diverse SF is challenging, as exemplified especially for the SS enzymes.

With respect to the first criteria defined by the SFLD for inclusion of proteins in the N6P SF (Section 2.1), the SS proteins can be classified as members of the N6P SF based on their overall similarities in sequence and structure to other SF members. For the second, similarity in the fundamental catalytic strategy by which catalysis is enabled, the SSs also fit as members of the N6P SF. As shown in Figure 2.2, like the reactions of other functionally characterized members of the SF, the SS reaction involves a nucleophilic attack on an electrophilic substrate, albeit using a different mechanism than the other known SF members. At first pass, the SSs appear to fail the last required criterion for inclusion in the SF, however, i.e., similarity in active site architecture or functionally important residues. Figure 2.7, which compares the active site of SS with those of metal dependent enzymes from the SGL and arylesterase-like subgroups, illustrates how very different the active sites of the SSs are relative to the rest of the SF. It is these active site
differences, taken together with the dissimilarity between the condensation reaction catalyzed by SS and the hydrolytic reactions of the rest of the SF, that make it difficult to rationalize the inclusion of the SSs as members of the N6P SF.

Figure 2.7. Active Site Superposition of (a) SGL (Drp35, pdb_id: 2dg1; red) and Arylesterase-like (PON1, pdb_id: 1v04; cyan) Subgroup Proteins. Conserved metal coordinating residues are labeled and colored by element; gray: carbon, blue: nitrogen, red: oxygen. (b) The same superposition as shown in (a), but with the metals removed for clarity and SS (pdb_id: 2fpb; green) added. Four SS residues that superimpose with alpha carbon positions for metal coordinating residues in Drp35 and PON1 are labeled black. The glutamate (Glu309) required for SS activity in 2fpb is also labeled black. These five SS residues are colored by element; green: carbon, blue: nitrogen, red: oxygen. Figure and Legend are adapted from Ref. 15.
Further comparison of the SSs with other N6P SF members moderates this view, however, leading us to classify the SSs as members of the N6P SF after all. First, a phylogenetic analysis of the SSL subgroup (Chapter 3) suggests that SS likely arose from a metal-dependent ancestor. Second, as shown in Figure 2.8a, comparison of a liganded structure from the SGL subgroup (DFPase) to liganded SS proteins shows that both enzymes initiate nucleophilic attack from the same position in the active site. These active site similarities allow us to speculate that SS evolved to bind tryptamine in the region of the active site occupied by two of the four metal binding ligands present in other SF members such as DFPase and leads to a hypothesis for a mechanistic role of tryptamine that is analogous to the roles of the metals in DFPase. The mechanistic diagrams shown on the right of Figure 2.8 describe this mechanistic hypothesis. In DFPase (Figure 2.8b), the reaction is initiated by a metal assisted nucleophilic attack of the oxygen of Asp229 of DFPase on the phosphorus of substrate, DFP. We speculate that the analogy in SS might involve substrate-assisted catalysis in which the amine group of tryptamine is deprotonated by Glu309. This is then followed by the subsequent attack of the amine on the aldehyde moiety of secologanin, (Figure 2.8c). Mutagenesis of Glu309 has previously established that it is required for SS activity.
Figure 2.8. A Related Catalytic Strategy Unites the SS Enzymes with the Rest of the N6P SF. Left panels: active site representations; right panels: diagrams depicting steps in the catalytic mechanisms as described in the text. (a) Superposition of the active site of DFPase (pdb_id: 2gvv; red backbone, SGL subgroup) with dicyclopentyl phosphoramidate inhibitor bound (orange; only the phosphorous of the inhibitor is depicted) (as seen in (b)) and of SS (green backbone) with tryptamine bound (pdb_id: ...
2fpb; pink) and with secologanin bound (pdb id: 2fpc; gray, only the aldehyde carbon is depicted) (as seen in (c)). Atoms and ligands are depicted as in (b) and (c), with the metal from (b) removed for clarity. (b) The left panel shows the active site of DFPase as in (a) except that the coordination of the metal to the metal binding ligands is shown. The nucleophilic oxygen of the catalytic Asp229, and the electrophilic phosphorous atom of the inhibitor are depicted as ball and stick. Two water molecules and the phosphoryl oxygen coordinated to the metal have been removed for clarity. The right panel depicts the nucleophilic attack of the oxygen of Asp229 on the phosphorous of the substrate DFP. (c) The left panel shows the superposition of SS with tryptamine bound and with secologanin as in (a). The acidic oxygen atom of the catalytic Glu309, the reactive nitrogen atom of tryptamine, and the electrophilic carbon of secologanin are shown as ball and stick. The right panel shows the deprotonation of the amine group of tryptamine by Glu309 and the subsequent nucleophilic attack of the amine on the aldehyde of secologanin. Figure and Legend are adapted from Ref. 15.

2.6 Re-examination of the Boundaries of the N6P SF in the Context of Other β-Propeller fold Proteins

The inclusion of SS in the N6P SF raises new questions for understanding the relationship of this SF to other β-propeller proteins. In the context of the N6P SF, we were able to rationalize the loss of metal coordinating residues in the mechanism and active site of SSs, allowing us to classify these proteins as members of the SF. But this interpretation is complicated in the context of the larger fold class which shows that other potentially homologous members of the fold class also lack conservation of metal binding ligands typical of the N6P SF. Thus, this larger view forces us to re-evaluate our initial criteria for excluding from the SF other proteins in the fold class based on their lack of metal binding ligands (Figure 2.9). We also note that different databases classify relationships among especially highly divergent groups of proteins differently. For example, we noted earlier that CATH classifies protein domains from the SGL subgroup of the N6P SF together with members of the Pfam families MRJP, NHL, PD40, and
LDL-recept_b, and a GSDH protein in a single structural SF they designate as the TolB, C-terminal domain SF. This differs from the N6P SF definition currently used by the SFLD. We argue, however, that neither definition should be considered as more "correct" than the other. Rather, each merely reflects the different criteria each resource has chosen to use in the classification of sequence, structural, and functional relationships.
Figure 2.9. Sequence Similarity Network of those Six-Bladed β-Propeller Proteins from the Largest Cluster of the Structure Similarity Network in Figure 2.2. Each node (designated as a “representative node”) signifies a group of similar proteins, ranging from 1-200 proteins, where each protein shares at least 40% identity with at least one other protein in the group, as calculated using CD-HIT-40. 4454 representative nodes are shown (relative node sizes not designated), representing a total of 17,281 sequences. Sequence similarity was computed from all-by-all pairwise BLAST comparisons among the sequences in and between each representative node. Edges are drawn between representative nodes only if the E-value of the best alignment score of a protein represented by that node is at least 1 x 10^-7, using as a background model UniRef100. At this threshold, the median alignment length is 211 residues and the median percent identity is 29.8%. As shown in the key, coloring is by number of conserved metal coordinating residues in the active sites of the majority of proteins represented by that representative node, as determined from a multiple sequence alignment of the N6P SF generated using the PROMALS program (alignment not shown). Large representative nodes with a light colored outline indicate that the representative node contains at least one structure. The shape of each representative node designates its Pfam assignment as determined by its best hit to Pfam models for members of the N6P SF only if at least one protein of the representative node hit can be assigned to that model. If the proteins represented by a representative node match to multiple Pfam models, then the majority dictates the shape. (Note that the group of red colored representative nodes isolated from the majority of the SGL subgroup in the Figure is still considered members of the subgroup. Their apparent separation results from the inadequacy of this 2D depiction to provide an accurate visual representation of high-dimensional connections.)

From the perspective of the SFLD classification, these relationships become more complicated than for classifications based primarily on sequence and structural similarity as the SFLD attempts to address how functional features are best included in the formal classification of functionally diverse enzyme SFs. Here, we examine further whether other proteins sharing the six-bladed β-propeller fold that also catalyze hydrolytic reactions should be considered as members in the SF. These include phytases, which catalyze the hydrolysis of phytic acid to phosphate, and neuraminidases, which hydrolyze glycosidic linkages of neuraminic acids, neither of which connect to the main cluster of proteins in the structure-based similarity network (Figure 2.1).
To evaluate whether phytases are good candidates for inclusion in the N6P SF, a superposition of phytase with DFPase was examined (not shown). Although this superposition revealed a couple of calcium-coordinating ligands in the phytase active site that are somewhat similar in position to metal binding ligands of DFPase, these side chains align poorly relative to alignments between structures classified as members of the SF. Overall, this alignment suggests a different binding modality relative to the N6P SF and suggests a distinct structure-function mapping in phytases. Superposition of this phytase structure with that of SS is not any better (also not shown), further arguing against the inclusion of the phytases as members of the N6P SF. Structural alignments between neuraminidases and N6P SF members are also poor and suggest these proteins are too dissimilar to include them in the SF. Outside of the families that catalyze hydrolytic reactions, including the NHL, PD40, GSDH and MRJP Pfam families, a further examination of those structures most similar to the N6P SF in Figure 2.1 suggests that these structural relationships are too weak to justify inclusion in the N6P SF (not shown).

Interestingly, we find strong conservation patterns between SGL subgroup proteins and seven-bladed β-propeller proteins (virginiamycin B lyases), which may indicate that our initial search for SF members, limited to using the canonically defined 6-bladed propeller fold, was insufficient. In β-propeller proteins, blade count may not be highly relevant in defining characteristics of a SF, as sequence similarity between proteins of different blade counts can be more significant than that within proteins sharing the same blade count\textsuperscript{42}. For example, sequences of the sialidases, which share the six-bladed β-propeller
fold with the N6P SF, are more distant from the SGL subgroup proteins than the SGL subgroup is to the 7-bladed virginiamycin B lyases. However, given the current lack of functional information for the SGL subgroup, a strong link between the N6P SF and the virginiamycin B lyases cannot be established.

2.7 Using the Superfamily Context to Select Protein Targets for Experimental Characterization

Sequence similarity networks that map characterized proteins together with unknowns allow identification of clusters of unknowns that have not yet been characterized, enabling a strategy for choosing proteins for characterization that could be especially informative for functional inference. For example, sequence similarity networks can help identify “functional boundaries” within a given SF and provide an estimation of how many functions remain to be discovered. Addressing this problem is a significant challenge for functional inference as clustering by sequence and structure alone frequently does not track with functional properties (see Refs. 5,44 for some examples). Additionally, knowledge gained from the SF perspective can suggest fundamental catalytic capabilities and constraints that, along with gene context and other orthogonal information, can be used to improve predictions of the functions of unknowns.

An example in the N6P SF is provided by the SSL subgroup protein from Vitis vinifera. The sequence similarity network for this subgroup (Figure 2.5) shows that this uncharacterized protein shares a high degree of overall sequence similarity with known
SSs. As reported previously\textsuperscript{15}, a structure-guided sequence alignment representing structures from all three N6P subgroups, several other sequences that sample the diversity of the SSL subgroup, and the \textit{Vitis vinifera} protein, revealed active site properties in the latter protein that appeared "transitional" between the true SSs and majority of the SSL subgroup proteins that conserve all four of the canonical metal binding ligands. The \textit{Vitis vinifera} protein conserves three of the four metal-coordinating ligands, with the missing ligand substituted by a glycine, a residue it shares in common with all known SSs. As predicted by the presence of these metal binding ligands especially, subsequent biochemical characterization of this protein demonstrated that it does not catalyze the SS reaction as a measurable level but instead exhibits a low level of esterase activity using \textit{p}-nitrophenyl acetate as a substrate. This result suggests properties of a functional boundary between SSs and SSL proteins and provides insight about possible paths by which SSs may have evolved within the background of the hydrolytic and metal-dependent SSL proteins.

Using a more stringent E-value cutoff than used for the SSL subgroup network presented in Figure 2.3, the similarity network in Figure 2.10 shows multiple discrete clusters, one of which cleanly separates all of the SSs into a single cluster. Although insufficient information currently exists to infer how many of the other clusters shown in this network are isofunctional, we hypothesize from this network view that at least several different functional families may remain to be identified in the subgroup. Further comparison of sequence profiles representing these clusters may be useful for hypothesizing additional
functional boundaries within the subgroup and for choosing sequences for structural characterization that would likely be useful in answering these questions.

Figure 2.10. Sequence Similarity Network of 393 Proteins of the Strictosidine Synthase-like (SSL) Subgroup Downloaded from the SFLD, Shown as in Figure 2.5, but Edges are Drawn if the Pairwise Similarity Score between Two Nodes is less than $1 \times 10^{-80}$ (More Stringent Cutoff). Sequence similarity was computed from pairwise BLAST comparisons of all of the members of the subgroup using the SFLD as a background model. Nodes are colored as designated in the key. Small circular nodes represent biochemically uncharacterized proteins; large diamond-shaped nodes indicate biochemically characterized proteins. Biochemically characterized SS are marked by the oval.

Like the SSL subgroup, the SGL subgroup, which comprises the vast majority of the N6P SF, is also poorly defined in terms of functional families and identification of the reactions they catalyze. However, knowing some types of reactions the characterized members of this subgroup catalyze (lactonase, esterase, organophosphatase) provides
hints about specific reactions that may be represented. When evaluated together with orthogonal information such as well-defined gene context, new targets for biochemical and structural characterization can be identified. For example, a potentially attractive target for characterization is the predicted SGL protein encoded in the genome of Burkholderia sp. H160 (indicated by the arrow in Figure 2.3), for which the gene context suggests a role in a D-xylose uptake and metabolism pathway (Figure 2.11). Putting this together with its likely function as a hydrolase, we speculate that this protein may catalyze the breakdown of xylono-1,4 lactone to produce D-xylonate.

Figure 2.11. Gene Context for an Uncharacterized SGL Subgroup Member from Burkholderia sp. H160 as Seen on the Integrated Microbial Genomes (IMG) System (http://img.jgi.doe.gov/cgi-bin/w/main.cgi). Genes are colored by predicted function: Red (labeled F) – Xylono-1,4-lactonase (SGL member). To the left, light blue (labeled E) – D-xylose dehydrogenase; white (labeled D) – glycoside hydrolase family 3 domain protein; teal (labeled C) – ABC transporter related; olive green (labeled B) – ABC transporter membrane protein 2, carbohydrate uptake transporter-1 family; light green (labeled A) – ABC transporter membrane protein 1, carbohydrate uptake transporter-1 family.

2.8 Conclusion

This Chapter describes links between structural and functional properties of the proteins of the N6P SF and places their relationships in the larger context of the six-bladed β-propeller fold proteins. Further dissection of differences among the proteins of the three
subgroups of the SF, the arylesterase-like, SGL and SSL subgroups, reveals clues about the functions of uncharacterized proteins in each, allowing classification of all of the SF members in the SFLD at least at the subgroup level. The limits described here for the prediction and annotation of reaction specificity families also addresses the difficulty of predicting functional boundaries and illustrates some important challenges for discriminating functions based on sequence or structural similarities, especially for functionally diverse enzyme SFs. These issues complicate our understanding of sequence, structure and function relationships in such systems and illustrate some of the significant challenges for achieving high quality annotations in the absence of experimental validation.

2.9 Access to Data from this Work

The N6P SF has been added to the Structure-Function Linkage Database (SFLD) and curated into subgroups and families where possible, based on available high quality experimental information. SFs are automatically updated on a regular basis and new sequences and structures are classified according to the SFLD schema using automated protocols to the extent possible, then further checked by human curators. These sequence and structure data are freely available at http://sfld.rbvi.ucsf.edu, along with full-length alignments of representative sequences identifying key conserved amino acids. Interactive versions of SF, subgroup, and family networks are also available for download and can be viewed using the freely available Cytoscape program which can be downloaded from http://www.cytoscape.org/.
2.10 Acknowledgements

This work was supported by NIH R01 GM60595 and U54 GM093342 to PCB. AEB was supported by the ARCS Foundation and the PhRMA Foundation Predoctoral Informatics fellowship. Molecular graphics images were produced using the UCSF Chimera package from the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (supported by NIH P41 RR001081). We thank Jimin Pei and Nick Grishin for their help in generating the PROMALS3D alignment.
Introduction to Chapter 3

This chapter describes in detail the strictosidine synthase-like (SSL) subgroup of the Nucleophilic Attack 6-bladed β-Propeller (N6P) superfamily. Additionally, based on both bioinformatic and biochemical analyses, we suggest most SSL proteins do not catalyze the strictosidine synthase reaction, but instead catalyze hydrolytic reactions. We propose biological roles for several bacterial SSL proteins and discuss the possibility of strictosidine synthase (SS) evolving from an ancestor with metal-coordinating active site residues.

A modified version of Chapter 3 has been published in the journal *Proteins: Structure, Function and Bioinformatics* and is available online here:

http://onlinelibrary.wiley.com/doi/10.1002/prot.23135/abstract. The article is titled “The Evolution of Function in Strictosidine Synthase-like Proteins” by Michael A. Hicks, Alan E. Barber II, Lesley-Ann Giddings, Jenna Caldwell, Sarah E. O’Connor, and Patricia C Babbitt. Lesley-Ann Giddings, Jenna Caldwell, and Sarah E. O’Connor performed the wet lab experiments, created the associated figures, wrote the experimental portion of the methods section, and contributed a few sentences detailing the experimental results. Alan E. Barber II helped create the similarity networks presented in the chapter. Michael A. Hicks performed the computational analysis and wrote the manuscript with Patricia C. Babbitt, who supervised and directed the work.
Chapter 3: The Evolution of Function in Strictosidine Synthase-like Proteins
3.1 Abstract

The exponential growth of sequence data provides abundant information for the discovery of new enzyme reactions. Correctly annotating the functions of highly diverse proteins can be difficult, however, hindering use of this information. Global analysis of large superfamilies of related proteins is a powerful strategy for understanding the evolution of reactions by identifying catalytic commonalities and differences in reaction and substrate specificity, even when only a few members have been biochemically or structurally characterized.

A comparison of >2500 sequences sharing the six-bladed β-propeller fold establishes sequence, structural and functional links among the three subgroups of the functionally diverse N6P superfamily: the arylesterase-like and senescence marker protein-30 gluconolactonase/luciferin-regenerating enzyme-like (SGL) subgroups, representing enzymes that catalyze lactonase and related hydrolytic reactions, and the so-called “strictosidine synthase-like” (SSL) subgroup. Metal-coordinating residues were identified as broadly conserved in the active sites of all three subgroups except for a few proteins from the SSL subgroup, which have been experimentally determined to catalyze the quite different strictosidine synthase (SS) reaction, a metal-independent condensation reaction. Despite these differences, comparison of conserved catalytic features of the arylesterase-like and SGL enzymes with the SSs identified similar structural and mechanistic attributes between the hydrolytic reactions catalyzed by the former and the condensation reaction catalyzed by SS. The results also suggest that despite their annotations, the great
majority of these >500 SSL sequences do not catalyze the SS reaction; rather, they likely catalyze hydrolytic reactions typical of the other two subgroups instead. This prediction was confirmed experimentally for one of these proteins.

### 3.2 Introduction

The number of protein sequences available in public databases such as UniProtKB/TrEMBL is now well over 16 million and continues to rise at exponential rates while our ability to functionally characterize these proteins is limited to just a small fraction of these. Computational approaches that can improve our ability to predict and study their molecular functions are therefore critical for leveraging useful information from genome sequencing projects. Yet, the application of these methods to many superfamilies (SFs) can be confounded by complicated patterns of variation across diverse but related protein sequences. Thus, even as an initial step for understanding their functions, annotation transfer from characterized to uncharacterized proteins based on simple similarity metrics can be insufficient to achieve high confidence prediction of their reaction specificities.

One approach to address the widening gap between experimentally characterized and uncharacterized proteins is a global comparison of sequence, structural, and functional features of evolutionarily related proteins to identify common catalytic properties as well as features that distinguish them. Such studies of a number of functionally diverse enzyme SFs have contributed to general hypotheses regarding nature’s evolutionary
strategies in the diversification of protein function (see Refs. 44,47-49 for some reviews) and provided useful roadmaps to interpreting their structure-function relationships.

This paper describes a functionally diverse enzyme SF distinguished both by its large size (>2500 sequences) and lack of experimental characterization for all but a tiny handful of its members. Previously described14 as comprising enzymes that catalyze several different chemical reactions, including strictosidine synthase (SS), paraoxonase (PON), and lactonohydrolase (Table 3.1), all of these enzymes belong to the 6-bladed β-propeller fold class and share a common catalytic feature involving nucleophilic attack on an electrophilic substrate. Termed here the Nucleophilic Attack 6-bladed β-Propeller (N6P) SF, its members can be clustered into three subgroups based on their sequence similarities. We define a subgroup as sharing similarity in sequence and function within a SF; each subgroup in turn may be comprised of multiple families, each of which catalyzes a distinct overall reaction.
<table>
<thead>
<tr>
<th>Protein Accession ID</th>
<th>Subgroup</th>
<th>Lactonase</th>
<th>Esterase</th>
<th>Organophosphatase</th>
<th>SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rauvolfia serpentina strictosidine synthase (pdb_id: 2fpb, 2fpc, 2fp8, 2pf9, 2vaq, 2v91)</td>
<td>SSL</td>
<td>?(a)</td>
<td>?</td>
<td>?</td>
<td>Yes(b) 50</td>
</tr>
<tr>
<td>Catharanthus roseus strictosidine synthase (GI:18222)</td>
<td>SSL</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>Yes 51</td>
</tr>
<tr>
<td>Ophiorrhiza pumila strictosidine synthase (GI:13928598)</td>
<td>SSL</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>Yes 52</td>
</tr>
<tr>
<td>Rauvolfia marnii strictosidine synthase (GI: 21097)</td>
<td>SSL</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>Yes (Predicted) (c) 53</td>
</tr>
<tr>
<td>Rauvolfia verticillata strictosidine synthase (GI: 118076220)</td>
<td>SSL</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>Yes (Predicted) (c) 54</td>
</tr>
<tr>
<td>Ophiorrhiza japonica strictosidine synthase (GI: 193792547)</td>
<td>SSL</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>Yes (Predicted) (c) 55</td>
</tr>
<tr>
<td>Homo sapiens adipocyte plasma membrane-associated protein (APMAP) (GI: 24308201)</td>
<td>SSL</td>
<td>?</td>
<td>Yes 37</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Homo sapiens paraoxonase 1 (PON1) (GI: 130675)</td>
<td>Arylesterase-like</td>
<td>Yes 30</td>
<td>Yes 30</td>
<td>Yes 30</td>
<td>?</td>
</tr>
<tr>
<td>Homo sapiens paraoxonase 2 (PON2) (GI: 6174935)</td>
<td>Arylesterase-like</td>
<td>Yes 30</td>
<td>Yes 30</td>
<td>No 30</td>
<td>?</td>
</tr>
<tr>
<td>Homo sapiens paraoxonase 3 (PON3) (GI: 29788996)</td>
<td>Arylesterase-like</td>
<td>Yes 30</td>
<td>Yes 30</td>
<td>Yes 30</td>
<td>?</td>
</tr>
<tr>
<td>PON1 G2E6 mutant (pdb_id: 1v04)</td>
<td>Arylesterase-like</td>
<td>Yes 31</td>
<td>Yes 31</td>
<td>Yes 31</td>
<td>?</td>
</tr>
<tr>
<td>Loligo vulgaris diisopropyl fluorophosphatase (DFPase) (pdb_id: 1e1a, 1pjx, 2ggv, 2ggv, 3byc)</td>
<td>SGL</td>
<td>No 24</td>
<td>?</td>
<td>Yes 24</td>
<td>?</td>
</tr>
<tr>
<td>Fusarium oxysporum lactonohydrolase (GI: 6448475)</td>
<td>SGL</td>
<td>Yes 30</td>
<td>?</td>
<td>No 24</td>
<td>?</td>
</tr>
<tr>
<td>Rattus norvegicus senescence marker protein-30 (regucalcin) (GI: 68067383)</td>
<td>SGL</td>
<td>Yes 36</td>
<td>Yes 17</td>
<td>Yes 17</td>
<td>?</td>
</tr>
<tr>
<td>Staphylococcus aureus drug responsive protein 35 (Drp35) (pdb_id: 2dg0, 2dg1, 2dso)</td>
<td>SGL</td>
<td>Yes 21</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Zymomonas mobilis gluconolactonase (GI: 48655)</td>
<td>SGL</td>
<td>Yes 57</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>

(a) No available literature reference showed that this activity was detected; (b) detected activity: "Yes", (c) Based on similarities to experimentally characterized SSs (see text): "Yes (Predicted)"; (d) no detectable activity: "No".
The strictosidine synthase-like (SSL) subgroup, the primary focus of this Chapter, currently includes three experimentally characterized SSs from *R. serpentina*[^58], *C. roseus*[^59], and *O. pumila*[^52] along with a much larger group of ~500 SSL sequences for which the reaction specificity has not been experimentally identified but that have been annotated in public databases as “putative strictosidine synthases” or “strictosidine synthase family proteins,” based on their similarities to the experimentally characterized SSs. In addition to these, proteins from *R. mannii*[^53] and *R. verticillata*[^54], which share 100% identity with the SS from *R. serpentina*, and a sixth protein from *O. japonica*[^55], which, based on the metabolic profiles of these species and presence of residues important to their SS activity, are predicted to catalyze the SS reaction as well. SS enzymes are produced by higher plants and catalyze the metal independent condensation of tryptamine and secologanin via a Pictet-Spengler reaction (Figure 3.1a) to generate strictosidine, a precursor molecule of the monoterpenoid indole alkaloid biosynthesis pathway. Starting with strictosidine, many pathways in different plants then lead to the production of about 2000 alkaloid compounds, several of which are useful in the treatment of a variety of human diseases, including cancer, malaria, and schizophrenia[^60]. Though in nature SS shows exquisite specificity for its substrates, rationally designed mutations in the enzyme’s active site permit modifications of substrate specificity and the formation of altered products[^61]. These “unnatural” products have been shown to be incorporated into the pathway, leading to potentially medically useful compounds[^62].

Many more SSs will likely be discovered as new plant genomes are solved, opening the possibility of additional “naturally” decorated variants that could enlarge the number of available precursors for new drugs in the strictosidine class. Sequences that have been
experimentally characterized as catalyzing the SS reaction are termed “true” SSs below to distinguish them from the uncharacterized SSL proteins, many of which we suggest catalyze some other reaction instead.

Figure 3.1. Examples of Chemical Reactions Catalyzed by the N6P SF. Red asterisk indicates electrophilic atom that is attacked. (a) SS reaction. (b) Examples of lactonase, esterase and phosphotriesterase reactions catalyzed by members of the SGL and arylesterase-like subgroups.

The other two subgroups in the N6P SF are termed herein the arylesterase-like and senescence marker protein-30/gluconolactonase/luciferin-regenerating enzyme (SGL) subgroups, based on the Pfam families of the same name\textsuperscript{12}. The arylesterase-like subgroup comprises ~200 proteins. Its characterized members include the human serum paraoxonases (the PON1, PON2, and PON3 families), which have been shown to have lactonase activity\textsuperscript{30}. PON1 proteins additionally catalyze ester bond cleavage and organophosphate degradation reactions\textsuperscript{29,31}, whereas PON2 and PON3 proteins have very limited arylesterase activity and virtually no organophosphatase activity\textsuperscript{32} (Table 3.1,
Figure 3.1b). *In vitro* studies have also shown that many of these proteins catalyze one or more of these reactions “promiscuously;” that is, at a significant rate enhancement but not at the level expected for native-like activity\(^{29,30,63}\). Like the arylesterase-like subgroup, the SGL subgroup, containing about 1800 members, catalyzes a similar set of chemical reactions. Examples include senescence marker protein-30, an enzyme involved in L-ascorbic acid biosynthesis in non-primate mammals\(^{16}\) that can also breakdown toxic organophosphates in mouse liver\(^{17}\), drug responsive protein-35 (Drp35), involved in the resistance to antibiotics by *Staphylococcus aureus*\(^{20,21}\), diisopropylfluorophosphatase (DFPase), which degrades organophosphates\(^{22}\) but for which the native activity remains unknown, and luciferin-regenerating enzyme, a protein that catalyzes luciferin-regeneration in fireflies\(^{19}\). Characterized proteins in the arylesterase-like and SGL subgroups are metal dependent and the great majority of sequences in these subgroups share a conserved set of active site residues that are involved in the coordination of a divalent metal ion. These have been implicated by mechanistic studies as being important for catalysis of these lactonase, esterase and organophosphatase activities\(^{21,24,27,64}\). Our sequence comparisons show that a similar pattern of metal coordinating ligands is conserved in the great majority of SSL sequences as well, including the SSL subgroup member human Adipocyte Plasma Membrane-Associated Protein (APMAP), which was recently determined to have some arylesterase activity\(^{37}\). The vast majority of proteins in both the arylesterase-like and SGL subgroups have not been biochemically characterized so that, like the SSL subgroup, their specific functions are not known.
In this paper, we describe a global computational comparison of the >2500 members of the N6P SF to identify their sequence, structural, and mechanistic links, which are then used to develop hypotheses about the functions of the many sequences of unknown function (“unknowns”) represented in the SSL subgroup. This global comparison provides a context for discriminating functional features and lays a foundation for prediction of reaction specificity of the unknowns in the N6P superfamily.

3.3 Materials and Methods

I. Data Set Sources and Curation

Full length protein sequences gathered from Pfam\textsuperscript{12} seed set families with near structural homology to any protein sharing the six-bladed $\beta$-propeller fold (strictosidine synthase: PF03088; Arylesterase: PF01731; SGL: PF08450; Folate_rec: PF03024; GSDH: PF07795; Ldl_recept_b: PF00058; MRJP: PF03022; NHL: PF01436; PD40: PF07676; Phytase: PF02333) were used as an initial starting point in identifying new homologous sequences. These data were combined into a “seed set” which was used as a query set for a series of BLAST\textsuperscript{65} searches against UniRef100\textsuperscript{46}. All hits from the initial BLAST search with E-values less than 1E-5 were kept and added. The set was filtered using HMMER 3.0 beta and its three defined filtering criteria\textsuperscript{66}. Any sequence that did not meet the three filtering criteria for any of the Pfam HMMs (ie: did not receive a HMMER score against any Pfam model) were dropped. This set was again used as a query set for BLAST using the same procedure from the first iteration. All sequences from the final set
were cropped to the sequence corresponding to the apparent six-bladed β-propeller domain. A curated sequence subset was aligned using PROMALS3D\(^1\). The clusters corresponding closely to the models for strictosidine synthase, SGL and Arylesterase contained conserved metal coordinating residues and were added to a “final SF set.”

Detectable activity, as defined in Table 3.1, is denoted as “yes” if a protein was reported in the literature to have hydrolytic or condensation activity above background for a given compound in a chemical class (ie: lactone, ester, organophosphate or tryptamine + secologanin for strictosidine synthase) by a spectrophotometric, HPLC, or cell-based assay.

**II. Generating Sequence Similarity Networks**

Sequence similarity networks of the SF and the SSL subgroup were generated using an altered version of a previously described methodology\(^3\) and visualized using the Cytoscape program\(^39\). Networks were generated in which nodes represent sequences and edges represent BLAST-based connections against the UniRef100 database\(^46\). An edge is drawn between two sequences only if the statistical significance of the similarity score between them is less than (better than) a defined E-value cutoff. The organic layout was used to generate the final graph.
III. Structure-guided Sequence Alignments

The best aligning chains of serum paraoxonase 1 from the arylerase-like subgroup (pdb_id: 1v04.pdb\textsuperscript{35}), \textit{Loligo vulgaris} ganglion diisopropyfluorophosphatase (pdb_id: 1pjx.pdb\textsuperscript{67}) and \textit{Staphylococcus aureus} drug responsive protein-35 (pdb_id: 2dg1.pdb\textsuperscript{21}) from the SGL subgroup, and \textit{Rauvolfia serpentina} strictosidine synthase (pdb_id: 2fpb.pdb\textsuperscript{36}) from the SSL subgroup were aligned using the Needleman-Wunsch algorithm\textsuperscript{68} as implemented in the Matchmaker program\textsuperscript{69} in Chimera\textsuperscript{70}. A companion program, Match -> Align, was used to generate a multiple sequence alignment based on the structure alignment. The sequence alignment was then refined by eye using the aligned structures as a guide.

In the case of 2gvv.pdb (DFPase with inhibitor bound), 2fpb.pdb (strictosidine synthase with tryptamine bound) and 2fpc.pdb (strictosidine synthase with secologanin bound), a structure-based sequence alignment was generated using the Matchmaker program\textsuperscript{69} in Chimera\textsuperscript{70}, as described above, without refinement by eye. The structure alignment was further refined by aligning the alpha carbons of the last three metal-coordinating residue positions. Distances for reactive group positions were then measured in Chimera\textsuperscript{70}. Sequence-based alignments of the SSL subgroup were generated for each phylogenetically-defined cluster using MUSCLE\textsuperscript{71}. For example, proteins in the plant only cluster were aligned to each other prior to generating a full subgroup alignment. Profile alignments, in which each cluster of aligned sequences was aligned with another cluster, were then created. Protein sequences from the arylerase-like and SGL
subgroups with associated structures were then aligned to the SSL subgroup alignment using MUSCLE. This overall alignment was refined by eye using the structure-based multiple sequence alignment as a guide.

IV. Gene Context Analysis

The amino acid sequence of a SSL gene fused to the transmembrane portion of an ABC transporter (gi|13471676) was used to identify other putative ABC transporter fusions by BLAST searches using the integrated microbial genomics system\textsuperscript{45}. The top eight non-redundant hits were selected based on their alignment length and gene neighborhoods evaluated.

V. Phylogenetic Tree

Proteins in the SSL subgroup alignment were filtered to 40\% identity using cd-hit\textsuperscript{54}, resulting in about 30 clusters. A single protein was selected from each cluster based on the median length of that cluster. Trees were constructed with MrBayes v3.1.2\textsuperscript{72,73} under the WAG amino acid substitution model\textsuperscript{74} using a gamma distribution to approximate rate variation among sites.
VI. Structure-Function Linkage Database

Data for the N6P SF has been added to the Structure-Function Linkage Database (SFLD) (http://sfld.rbvi.ucsf.edu)\(^75\) as described in the text.

VII. General Methods and Analytical Techniques

Secologanin was isolated as previously described\(^76\). All chemicals were purchased from Sigma Aldrich unless otherwise noted.

A Varian Cary 50 Bio UV/Visible Spectrophotometer equipped with a Cary 50 microplate plate reader was used to measure hydrolysis products in colorimetric assays. UPLC and MS analyses were performed in tandem on an Acquity Ultra Performance BEH C18 column with a 1.7 mm particle size, 2.1 x 100 mm dimension, which was coupled to a Micromass LCT Premier TOF Mass Spectrometer by Waters Corporation (Milford, MA) with an electrospray ionization source. Analytes were separated, using a 10-50% acetonitrile: water (0.1% formic acid) over 5 minutes and flow rate of 0.5 mL min\(^{-1}\). For MS analyses, the capillary and sample cone voltages were 3,000 V and 30 V, respectively. The source temperature was 100°C while the desolvation temperature was 300°C. The cone and desolvation gas flow rates were 60 L hr\(^{-1}\) and 800 L hr\(^{-1}\), respectively.
VIII. Cloning and Protein Expression

*C. roseus* SS in pET28a (+) and empty vector were transformed into *Escherichia coli* BL21 (DE3) cells for protein expression. The *V. vinifera* SSL gene (gi|147772032) was synthesized by CODA genomics (now Verdezyne; Carlsbad, CA) and NcoI and XhoI restriction sites were introduced by PCR for standard directional cloning into pET32b (+). The construct was then transformed into *E. coli* Rosetta DE3 cells for protein expression. PON1 (variant G2E6) in pET32b (+) was transformed into *E. coli* Origami B DE3 cells for protein expression. All liquid and solid media were supplemented with 1 mM CaCl₂.

Overnight cultures were grown at 37°C in sterile LB-broth containing 1 mM CaCl₂ and the appropriate antibiotic selection. Cultures of *C. roseus* SS (500 mL), containing 1 mM CaCl₂ and 50 µg/mL of kanamycin were inoculated with overnight cultures (1:100 dilution), and grown at 37°C until the optical density at 600 nm reached 0.5-0.75. After cultures were chilled at 4°C for 30 min, protein expression was induced with 1 mM isopropyl-β-D-1-thiogalactopyranoside (IPTG). Cells were harvested by centrifugation after 18 hours of protein expression at 18°C and stored at -80°C. Cultures of pET28a(+) (500 mL) containing 1 mM CaCl₂ and 50 µg of kanamycin were incubated with overnight cultures (1:100 dilution), and grown at 30°C until the optical density at 600 nm of 0.5-0.75 was reached. Protein expression was induced with 1 mM IPTG and after 4 hours the cells were harvested by centrifugation and stored at -80°C.
Cultures of *V. vinifera* SSL and the empty pET32b (+) vector (500 mL) containing 1 mM CaCl$_2$ and 100 µg/mL of ampicillin and 34 µg/mL of chloroamphenicol were inoculated with overnight culture (1:100 dilution), and grown at 37°C until the optical density at 600 nm of 0.5-0.75 was reached. Protein expression was induced with 1 mM IPTG and after 6 hours the cells (2 hours for empty vector) were harvested and stored at -80°C. PON1 was expressed as previously reported$^{77}$.

**IX. Protein Purification**

PON1 was purified as described previously$^{77}$. Cells expressing *C. roseus* SS, pET28a (+), pET32b (+), and the *V. vinifera* SSL were lysed by sonication in lysis buffer (pH 8) containing 50 mM HEPES, 1 mM CaCl$_2$, 300 mM NaCl, 10 mM imidazole, 10 % glycerol, 4 mg lysozyme, and 0.4 µg leupeptin and pepstatin protease inhibitors. The lysate was then incubated in 0.1% tergitol for 2.5 hours at 4°C. After centrifugation, the supernatant was incubated with 0.01% v/v pre-equilibrated Ni-NTA resin suspension for 1 hour before the flow-through was collected. The resin was washed with one column volume of lysis buffer and 2 column volumes of wash buffer containing 50 mM HEPES, 1 mM CaCl$_2$, 300 mM NaCl, 20 mM imidazole, 10 % glycerol, and 0.1% tergitol. The resin was then washed with increasing concentrations of imidazole and the histidine-tagged proteins were eluted in pH 8 buffer containing 50 mM HEPES, 1 mM CaCl$_2$, 300 mM NaCl, 250 mM imidazole, and 10 % glycerol. The eluent was concentrated in Amicon Ultra centrifugal filter units by Millipore (Billerica, MA) and buffer exchanged using 50 mM HEPES buffer containing 162 mM NaCl, 1 mM CaCl$_2$, and 10 % glycerol.
The final protein concentration was determined using a bichinchoninic acid assay by Pierce (Rockford, IL).

X. Hydrolase Activity with \( p \)-Nitrophenyl Acetate

A stock of 300 mM \( p \)-nitrophenyl acetate (pNPAc) was prepared in HPLC-grade methanol and diluted for enzyme assays. Colorimetric assays to detect the formation of \( p \)-nitrophenol at 405 nm were prepared in a MICROTEST 96 well plate from Becton Dickinson Labware (Franklin Lakes, NJ) with a final volume of 250 µL containing either 530 nM (\( C. \) roseus SS, pET28a (+)) or 53 nM (PON1, \( V. \) vinifera SSL, and pET32b (+)) of protein in 100 mM HEPES buffer containing 1 mM CaCl\(_2\) and 2.6 mM NaCl. After pre-equilibration at room temperature for 10 minutes, assays were initiated by the addition of pNPAc (0.5 mM, 1 mM, 1.6 mM, 2.1 mM, 2.6 mM, and 3.2 mM). Time-points were chosen such that the rate of product formation was linear, to ensure accurate measure of initial rates (between 3 and 30 minutes after initiation of the reaction). The measured pathlength of 0.79 cm and \( p \)-nitrophenol extinction coefficient of 18,000 M\(^{-1}\) cm\(^{-1}\) were used to convert the absorbance units of \( p \)-nitrophenol into concentrations by the Beer-Lambert law. Kinetic parameters were estimated from two kinetic trials by fitting the data to the Lineweaver-Burk plot since limited substrate solubility did not enable an acceptable fit to the Michaelis-Menten equation (maximal substrate concentration was less than the 2-3 x \( K_M \)).
XI. Pictet-Spenglerase Activity

To detect Pictet-Spenglerase activity, assays were prepared in a final volume of 100 uL containing 234 nM protein (C. roseus SS or V. vinifera SSL), and 200 µM tryptamine in 100 mM pH 7 phosphate buffer. Assays were initiated by the addition of 1.2 mM secologanin and incubated at 30°C overnight. Ten percent of the assay volume was quenched with HPLC-grade methanol, clarified by centrifugation for 5 minutes in a microcentrifuge, and analyzed by LC-MS using selected ion monitoring at the mass of the expected product (strictosidine, m/z 531).

3.4 Results and Discussion

In the sections below, sequence similarity networks \(^3\) generated from all-by-all pairwise comparisons of >2500 sequences are used to summarize relationships across the subgroups of the N6P SF and provide functional context for more detailed comparisons of their active sites and mechanisms. Functional and biological features mapped to the networks then enable visualization of functional trends across the SF.

Although the annotations in public databases for the unknowns in the SSL subgroup implicitly infer that they catalyze the SS reaction, our global analysis of sequence, structure and function relationships in the N6P SF suggests that they do not. In the first section, we show that although all of the SSL subgroup proteins, including the characterized SS enzymes, cluster closely together and are quite distinct from the
arylesterase-like and SGL subgroups, the active sites of the true SS enzymes are very different from the predicted active sites of the other SSL subgroup sequences. Moreover, sequence analysis of active site motifs of the great majority of these SSL proteins shows them to be more similar to those of the arylesterase-like and SGL subgroups than to the true SSs, suggesting that they are more likely to catalyze hydrolytic reactions common to those two subgroups instead. Confirmation of hydrolytic activity (and the lack of detectable SS activity) in one of the SSL proteins (gi|147772032 from Vitis vinifera), a SSL subgroup member that is most similar to true SSs, provides experimental support for this prediction and confirms the conclusion that the true SSs that catalyze the Pictet-Spengler reaction are outliers in this superfamily. In the next section, phylogenetic analysis is used to address the relationship of the SSs to the rest of the SSL subgroup, allowing us to suggest that the SS reaction may have evolved from a metal dependent ancestor. The third section provides additional evidence that the huge majority of the SSL unknowns are unlikely to catalyze the SS reaction and presents clues about some of their biological functions, including a role for some in ABC transport systems. In the final section, we describe further structural and mechanistic similarities between the outlier SS enzymes and rest of the SF that help rationalize the differences in their active sites and overall reactions and provide support for their inclusion in the N6P SF.

I. Similarities and Differences among SSL, SGL, and Arylesterase-like Subgroup Proteins are Complicated and Suggest that Most SSL Proteins do not Catalyze the SS Reaction
Figure 3.2. Sequence Similarity Network for the N6P SF. Each node represents one of the >2500 SF protein sequences; edges between nodes are drawn only if the similarity between a pair of sequences is better than an E-value threshold cutoff of 1E-10 (median alignment length = 238 residues; median percent identity of pairwise comparisons = 30.0%). The network is visualized using the organic layout in Cytoscape. While the lengths of connecting edges tend to correlate with the relative dissimilarities of each pair of sequences, these distances do not represent a quantitative correlation. Coloring is by subgroup; red: SGL, green: SSL, blue: arylesterase-like. Large diamond shaped nodes: proteins represented in Table 3.1.

The sequence similarity network comparing the proteins in the N6P SF shows that SSL proteins share a higher degree of similarity with SGL members than either subgroup does with any member of the arylesterase-like subgroup (Figure 3.2), even though comparison of their enzymatic functions suggests a closer relationship between the arylesterase-like and SGL subgroups than for either with the true SSs (Figure 3.1). The best connection seen between an arylesterase-like protein and any other subgroup is at an E-value of 1.2X10^{-8}, where a single edge appears between the arylesterase-like and SGL subgroups.
(data not shown). At an E-value threshold of 1X10^{-5}, multiple connections are seen between the arylesterase-like and SSL subgroups. No connections are seen between the SSL and arylesterase-like subgroups at the E-value threshold of 1X10^{-5}, which is the least significant E-value at which we feel connections can be considered as minimally confident. Consistent with the functional evidence, comparison of available structures for all three subgroups also shows that the active sites of these arylesterase-like and SGL proteins are similar to each other while the active site of the true SS structures is indeed highly divergent from the other two subgroups (Figure 3.3).

Using Drp35 (pdb_id: 2dg1; SGL subgroup) as a reference structure, PON1 (pdb_id: 1v04; arylesterase-like subgroup) aligns at 86 alpha carbon positions with an overall RMSD of 1.18 Å, indicating that the overall structures of these two representative N6P SF members are highly similar. The most striking feature in their superposed active sites is a conserved set of four residues coordinating to a divalent metal ion (Figure 3.3a); this metal-dependent active site architecture has been implicated in the phosphotriesterase mechanism of DFPase (SGL subgroup) and perhaps also in the phosphotriesterase mechanism of PON1 (arylesterase-like subgroup). In DFPase, the oxygen from an aspartate from blade 5 is thought to perform a direct nucleophilic attack on a phosphorous atom in a phosphoryl group of DFP, as demonstrated by an H$_2^{18}$O incorporation experiment in which the metal coordinating oxygen of the aspartic acid is replaced by an oxygen atom of a solvent water molecule. The other three residues involved in metal

---

1 Although additional residues may coordinate to the metal in some enzymes (such as a fifth residue seen in the PON1 structure), a mechanistic role has not yet been defined beyond the four residues detailed in our analysis. As such, our analysis is limited to speculating on the role of these four residues in the SF.
coordination appear to be necessary to maintain the electrostatic environment required for proper orientation of the substrate. The identity of a physiologically relevant divalent metal ion for many of these proteins is unclear; *in vitro* hydrolytic activity has been found using a variety of metals for proteins in the SGL subgroup, including Ca$^{2+}$, Mg$^{2+}$, Zn$^{2+}$, and Mn$^{2+}$ in human SMP-30; Mn$^{2+}$ and Zn$^{2+}$ for the lactonase activity and Mg$^{2+}$, Mn$^{2+}$, Co$^{2+}$ and Cd$^{2+}$ for the DFPase activity in mouse SMP-30; and Ca$^{2+}$ in Drp35.

Regardless of their exact identity, these metals appear necessary to polarize the sp2-hybrid bonds of substrates (such as carbonyl or phosphoryl groups), and to stabilize the resulting negative charge from transition states and other intermediates. In addition to the catalytic Ca$^{2+}$ ion required for lactonase and esterase reactions in PON1, a histidine-histidine dyad has been implicated in activation of a water molecule involved in nucleophilic attack on the carbonyl of the substrate (Figure 3.S1).\textsuperscript{35,79}
Figure 3.3. Active Site Superposition of (a) SGL (Drp35, pdb_id: 2dg1; red) and Arylesterase-like (PON1, pdb_id: 1v04; cyan) Subgroup Proteins. Conserved metal coordinating residues (see text) are colored by element; gray: carbon, blue: nitrogen, red: oxygen. (b) Superposition of Drp35, PON1 and SS (pdb_id: 2fpb; green). SS residues are labeled black that superimpose with alpha carbon positions for metal coordinating residues in Drp35 and PON1. The glutamate (Glu309) required for SS activity in 2fpb is also labeled black. These 5 residues are colored by element; green: carbon, blue: nitrogen, red: oxygen. The metals shown in (a) have been removed for clarity.

In contrast, the active site of the true SS from *Rauvolfia serpentina* lacks all four of these metal coordinating ligands (Figure 3.3b). Using Drp35 as a reference structure, SS aligns at 154 alpha carbon positions with an overall RMSD of 1.67 Å, suggesting significant overall structural similarity. However, while the secondary structure features of the SS
and Drp35 active sites overlay, SS lacks all four metal binding ligands, making the active site environment substantially different from those of Drp35 and PON1. This difference is consistent with a metal-independent mechanism in the Pictet-Spengler reaction that is thought to proceed using a catalytic glutamic acid from blade 6 (Glu309), which abstracts a proton from the amine group of the substrate tryptamine, allowing it to attack the aldehyde of secologanin. The indole then attacks the resulting iminium species to form the tetrahydro-β-carboline product, strictosidine⁴.

Figure 3.4. Sequence Similarity Network Showing Conservation of Metal Binding Residues for the 516 SSL Subgroup Sequences Generated from All-by-all Pairwise Comparisons. Nodes, edges, and layout are as in Figure 3.2 except that edges are drawn only for comparisons scoring better than an E-value threshold cutoff of 1E-50 (median
alignment length = 297 residues, median percent identity = 41%). The arc with black arrows indicates the region of the network in which most of the connections between SSL proteins and the SGL subgroup are found (Figure 3.2). Large diamond shaped nodes: enzymes that have been biochemically and/or structurally characterized. Large nodes labeled with arrows and gi numbers: proteins shown in the motif alignment in Figure 3.5. True SSs are labeled and circled, including both biochemically characterized and predicted SSs (see Table 3.1). Nodes are colored by number of metal-coordinating residues; all four present: red, 3 of 4: yellow, 2 of 4: green, 1 of 4: cyan, no metal-coordinating residues present: gray.

In an attempt to resolve the apparent contradiction in the relationships between sequence, structure and reaction specificity among the three subgroups, we examined the SSL subgroup in greater detail as shown in Figures 3.4 and 3.5. In the network shown in Figure 3.4, increasing the stringency at which edges are drawn to an E-value less than $1 \times 10^{-50}$ illustrates the similarity connections among the SSL subgroup proteins in more detail and shows how the network begins to “come apart” into smaller clusters representing higher levels of sequence similarity within each subset. From Figure 3.4, it is clear that the experimentally characterized true SSs lie at the periphery of the network, sharing similarity connections with only a subset of SSL subgroup proteins. Strikingly, the Figure also shows that with the notable exception of the true SSs, the vast majority of the SSL proteins also conserve the four metal binding ligands generally expected of SF members that catalyze hydrolytic reactions. Of those that do not (129 of the 516 sequences shown in Figure 3.4), about half are so diverse from other sequences in the set that their alignments in the active site region are difficult to confirm, making it difficult to evaluate whether or not they are indeed missing one of more metal binding ligands. Most of the others of these proteins have not been experimentally characterized. As a result, insufficient evidence is available to speculate further regarding their reaction specificities.
Figure 3.5. Active Site Motif Alignments of Nine SSL Subgroup Proteins with Structures from the Arylesterase-like (blue, pdb_id:1v04), SGL (red, pdb_id’s: 2dg1, 1pxj) and SSL subgroups (green, pdb_id: 2fbp). Numbering is from the SS structure, also highlighted in green. Highlighted in bold and yellow: four conserved metal coordinating active site residues that are found in the characterized arylesterase-like and SGL subgroup members and the majority of uncharacterized sequences in the SSL subgroup (indicated by gi numbers, Figure 3.4), but not in the true SS (row highlighted in green). Highlighted in bold and gray: catalytic glutamate required for SS activity. Note that in PON1 (1v04.pdb), a 5th residue (N255) is thought to coordinate to the calcium; this residue is conserved in many of the SSL proteins shown and may play a similar role.

Although no structures are yet available for any of these SSL subgroup unknowns, their sequence conservation of the four “canonical” metal binding ligands suggest that most are indeed metal dependent. Figure 3.5 provides a structure-guided sequence alignment comparing active site motifs from the structures in Figure 3.3 with nine SSL sequences (indicated by white arrows in Figure 3.4) that are divergent from the true SSs. As indicated in the Figure, these unknowns generally appear to conserve all the four metal binding residues and also likely lack the catalytic glutamate required for catalysis of the SS reaction, again suggesting they are more likely to catalyze a hydrolytic reaction than the SS reaction. The recent identification of a low level of arylesterase activity in human APMAP37, an SSL protein that appears to have all four metal coordinating residues (gi|24308201 in Figure 3.5) provides some initial experimental validation of this
prediction. However, more detailed analysis of many more of these proteins will be required to determine their reaction specificities or promiscuous capabilities for known reactions in the arylesterase-like and SGL subgroups. For example, we note that while SSL proteins do not appear to conserve the His-His dyad that PON1 requires for its lactonase/esterase activities (Figure 3.S1), other lactonase/esterase reactions in the characterized SGL enzymes also lack this feature.

The SSL sequences most similar to the true SSs share between 37 and 49% identity with them, sufficiently distant to suggest that they may well catalyze reactions other than SS. Importantly, they also appear to lack the catalytic glutamate required for SS activity (data not shown except for the proteins labeled by gi numbers in Figure 3.5), again supporting the hypothesis that they do not catalyze the SS reaction. Notably, several of these closest neighbors to the true SSs also lack one or more of the conserved metal binding ligands typical of the hydrolytic SF members. Because it appears in Figure 3.5 to clearly lack the metal binding ligand at position 210 and is also most similar to SS in the motif shown in the Figure that is associated with that position, we chose the protein from *Vitis vinifera*, gi|147772032, for experimental examination of its ability to catalyze either a hydrolytic or the SS reaction. Consistent with our hypotheses, it was found to hydrolyze the model substrate *p*-nitrophenyl acetate (pNPAc), albeit at a level lower than the activity reported for PON1 but still significantly higher than background hydrolysis observed with *C. roseus* SS (Figure 3.S2). Steady-state kinetic analyses of the *V. vinifera* protein assayed with pNPAc revealed that the protein had a $V_{\text{max}}$ of $220 \pm 140$ mol min$^{-1}$ mol$^{-1}$ and an estimated $K_M$ of $8 \pm 3$ mM. Interestingly, the *V. vinifera* protein does not appear to
catalyze the Pictet-Spengler condensation between tryptamine and secologanin (Figure 3.S3).

The low level of hydrolytic activity in this enzyme is not unexpected since its true substrate is not yet known, nor are the consequences of missing one of the metal binding ligands. This missing metal ligand, a highly conserved Asn that aligns in all of the presumed hydrolytic enzymes shown in Figure 3.5 (position 210), is replaced by a Gly in both the *Vitis vinifera* protein and SS. Although this asparagine has been shown to play a role in maintaining the electrostatic environment necessary for the degradation of toxic organophosphates by DFPase in the SGL subgroup, it might be less important for other hydrolytic reactions such as those catalyzed by some lactonases or esterases.

Alternatively, a subset of SSL proteins lacking one or more canonical catalytic residues may play a regulatory role in some as yet unknown cellular process (for examples in other systems see Ref. 80).

The differences between this *Vitis vinifera* protein and true SSs may also suggest variations that are “transitional” between the majority of SSL proteins predicted to be hydrolases and the true SSs. Since so few structures and little biochemical evidence are available for their functions or mechanisms for either the arylesterase-like or SGL subgroups and none for the SSL proteins (besides the true SSs), we cannot resolve the differences at this time. Further evaluation of this hypothesis will likely require both structural characterization and identification of the physiological substrate of the *Vitis vinifera* protein as well as additional uncharacterized proteins from the SSL subgroup.
II. SS may have Evolved from an Ancestor with Metal-Coordinating Active Site Residues

To gain additional clues regarding how the SS reaction could have evolved in the context of SSL subgroup ancestry, we constructed a phylogenetic tree for a representative subset of the subgroup (Figure 3.6). The clustering patterns of the leaves of the tree appear similar to the clustering patterns seen in the sequence similarity networks (Figure 3.4), with clusters generally correlating with type of life (Figure 3.7). Two distinct clades of plant proteins are present: one corresponding to the plant only cluster of the SSL subgroup network, and the other clade joined with bacterial proteins corresponding to the “mixed” cluster of the SSL subgroup in Figure 3.7. Additionally, separate clades for vertebrate and for invertebrate animal proteins can be distinguished; however, the interior node joining these two clades is not well-resolved.
Figure 3.6. Bayesian Phylogenetic Tree of a Representative Subset of SSL Subgroup. Thirty proteins for which no two proteins share greater than 40% identity were used. Coloring of leaves and branches is by type of life: plants, green; bacteria, blue; vertebrate animals, red; invertebrate animals, orange. Characterized SS (gi|193792547) is indicated by an asterisk. Branch confidence values: >0.95, filled circle; 0.70-0.94, open circle; 0.65-0.69, no circle.

Figure 3.7. Sequence Similarity Network Showing Types of Life for the 516 SSL Subgroup Sequences Generated from All-by-all Pairwise Comparisons. Nodes, edges, and layout are as in Figure 3.2 except that edges are drawn only for comparisons scoring better than an E-value threshold cutoff of 1E-50 (median alignment length = 297 residues, median percent identity = 41%). The arc with black arrows indicates the region of the network in which most of the connections between SSL proteins and the SGL subgroup are found (Figure 3.2). Large diamond shaped nodes: enzymes that have been biochemically and/or structurally characterized. Large nodes labeled with arrows and gi numbers: proteins shown in the motif alignment in Figure 3.5. True SSs are labeled and circled, including both biochemically characterized and predicted SSs (see Table 3.1). Coloring is by “type of life”; bacteria: blue, plants: green, archaea: magenta, vertebrate animals: red, invertebrate animals: orange, protozoa: yellow.
As noted earlier, gi|147772032 from *V. vinifera* may reflect features of an evolutionary transition between hydrolytic SSLs and SS in the ancestry of the subgroup, displaying characteristics of the metal-coordinated active site common to most of the SF, as well as characteristics more similar to those in the SS active site motif (Figure 3.5). An evaluation of the position of this sequence in the tree relative to characterized SS may provide clues about how the SS function could have arisen. It is therefore interesting that a characterized SS protein (gi|193792547) is the nearest neighbor of gi|147772032 in the tree. The next most interior node, gi|1754987 (not shown in Figure 3.5), also appears to share three of the four metal coordinating residues common to most of the SF members, with the Asn/Gly position (position 254 in Figure 3.5) substituted by an Ala. Though the interior node joining the next closest sequence (gi|225467502) with these three others is not well-resolved, the interior node distinguishing the plant only cluster from the rest of the SSL subgroup is resolved to a posterior probability of 1. All the proteins contained therein, with the sole exception of the true SS, conserve some or all four of the metal coordinating active site residues common to the rest of the SF, lending support to the idea that the SSs, found only in recently evolved higher plants, may have diverged from a metal-dependent ancestor.

It is also possible that there are more complex evolutionary origins to the functions represented in the SSL subgroup, and that SS and SSL proteins may be paralogous rather than orthologous. Sequencing of the genomes of additional higher plants to fill in the sequence links among these very diverse proteins, along with further experimental characterization of SSL proteins will be required to address this issue.
Besides information from homology, many other types of information have been used to infer functional properties of sequences discovered in genome projects (see \textsuperscript{81} for a recent review). Gene context suggests that several SSL proteins from gram negative bacteria may function in ABC transport, most likely involved in the uptake of carbohydrates (Figure 3.8). These proteins are sometimes fused to ABC transmembrane domains (such as gi|13471676 from \textit{Mesorhizobium loti MAFF303099}), but can also appear as independent domains proximal to genes encoding ABC transport machinery, such as an ATP-binding domain or substrate binding protein. Based on the sequence similarity of these ABC transporter components (not shown), we can hypothesize that these proteins are members of the carbohydrate uptake transporter-2 (CUT2) subfamily. Proteins in this subfamily are known to import monosaccharides such as ribose\textsuperscript{82} and xylose\textsuperscript{83} as well as ribonucleotides\textsuperscript{84}. The proximity of at least two of these proteins (gi|238761435 from \textit{Yersinia kristensenii ATCC 33638} and gi|258637446 from \textit{Pantoea sp. At-9b}) to genes encoding proteins similar to carboxymuconolactone decarboxylases suggest that these transporters likely import lactones and related compounds. Consistent with that interpretation, the presence of metal-coordinating residues in all of the SSL domains in this set of proteins suggests a potential enzymatic role as well, perhaps functioning as the first step in a secondary sugar metabolism pathway.
Figure 3.8. Gene Neighborhood for SSL Proteins in Gram Negative Bacteria as Seen on the Integrated Microbial Genomes (IMG) System. Colored genes (other than light-yellow) have orthologous components in another organism. Boxed: permease protein of sugar ABC transporters (light green, labeled A); binding protein components of sugar ABC transporter (cyan, labeled B); a second permease protein of sugar ABC transporter, but in this case fused to a SSL protein (red, labeled C); a second, independent SSL protein (light-pink or in some cases, light-yellow, when an orthologous component is not detected, labeled D); and an ATP binding protein (light blue, labeled E).

Other SSL proteins appear to be involved in immunity. For example, hemomucin, an innate immune receptor in flies, is represented in the cluster of proteins from invertebrates (Figure 3.7; orange). All of the proteins in this cluster have an SSL domain while only some have the additional mucin-type repeats described elsewhere. Interestingly, nearly all these proteins appear to have the metal coordinating residues associated with hydrolytic activity (Figure 3.4). Whether the presence of these residues...
confers hydrolytic activity and whether its potential unknown activity is required for signal transduction by the hemomucin receptor remain to be tested.

IV. Comparison of Liganded SGL and SSL Subgroup Structures Reveals Intriguing Similarities in Active Site Mechanisms Despite the Differences between their Divergent Reactions and Active Sites

Although all of the members of the N6P SF appear to share a general catalytic strategy involving nucleophilic attack on an electrophilic substrate, the very substantial differences in active site and overall chemical reaction between true SSs and the great majority of other SF members indicates that the true SSs are functional and structural outliers of the SF. It is difficult to rationalize these differences between the SSs and the rest of the SSL subgroup because we lack structural or mechanistic characterization of any SSL protein, or even their reaction specificities. However, since the great majority of SSL proteins share the four metal binding ligands typified by the other two subgroups, structural and mechanistic comparison of the active site of SS with PON1 and DFPase allows a first-pass speculation about how SS activity could have evolved from a metal dependent ancestor.

In contrast to the metal independent Pictet-Spengler reaction\textsuperscript{4} catalyzed by SS, conservation of metal ligands across the rest of the N6P SF appears to be a fundamental requirement for the activation of substrates for hydrolysis and the stabilization of intermediates and transition states. For its lactonase and esterase activities\textsuperscript{35}, PON1
requires these conserved metal coordinating residues as well as a fifth metal binding residue and a His-His dyad (Figure 3.S1). DFPase lacks both the His-His dyad and fifth metal binding residue and uses its four metal-coordinating residues directly in the phosphotriesterase mechanism. Superposition of the liganded DFPase and SS structures reveals similarity in the positions of two of these metal ligands in DFPase with that of bound tryptamine from SS (Figure 3.9). Remarkably, the position from which the direct nucleophilic attack that forms a phosphoenzyme intermediate in DFPase occurs is in a similar spatial orientation to that involved in the nucleophilic attack by tryptamine on the aldehyde of secologanin to form the carbinolamine intermediate in SS (Figure 3.9, Table 3.S1).
Figure 3.9. A Related Catalytic Strategy Unites the SS Enzymes with the Rest of the N6P SF. Left panels: active site representations; right panels: diagrams depicting steps in the catalytic mechanisms as described in Refs. 24 and 4. (a) The left panel shows the active site of DFPase (pdb id: 2gvv; red backbone, SGL subgroup) with dicyclopentyl phosphorimidate inhibitor bound (orange; only the phosphorous of the inhibitor is depicted). The nucleophilic oxygen of the catalytic Asp229, and the electrophilic
phosphorous atom of the inhibitor are depicted as ball and stick. Two water molecules and the phosphoryl oxygen coordinated to the metal have been removed for clarity. The right panel depicts the nucleophilic attack of the oxygen of Asp229 on the phosphorous of the substrate DFP. (b) The left panel shows the superposition of SS (green backbone) with tryptamine bound (pdb id: 2fpb; pink) and with secologanin bound (pdb id: 2fpc; gray, only the aldehyde carbon is depicted). The acidic oxygen atom of the catalytic Glu309, the reactive nitrogen atom of tryptamine, and the electrophilic carbon of secologanin are shown as ball and stick. The right panel shows the deprotonation of the amine group of tryptamine by Glu309 and the subsequent nucleophilic attack of the amine on the aldehyde of secologanin. (c) Superposition of the proteins shown in (a) and (b). Atoms and ligands are depicted as in (a) and (b), with the metal from (a) removed for clarity.

Based on this structural similarity and the known mechanisms of these enzymes, we speculate that the substitution of two conserved metal residues to glycines in SS (positions 210 and 254 in Figure 3.5), which includes the loss of the aspartate required in the phosphotriesterase mechanism in DFPase, may have created space for the binding of tryptamine in the SS active site. The eventual loss of the other, now unneeded, metal coordinating residues could then have occurred over the course of evolution. Consistent with this notion, the amino group in tryptamine seems to play a role in the mechanism of SS that is analogous to the role that the nucleophilic oxygen atom of the conserved aspartic acid plays in the phosphotriesterase mechanism of DFPase. That is, both reaction mechanisms involve a nucleophilic attack on a sp2-hybridized electrophilic atom from the same part of the active site, one using metal-assisted catalysis and the other using substrate-assisted catalysis.

The scenario is more complicated for PON1. Here, catalysis of lactonase/esterase reactions involves the use of a His-His dyad from a different side of the active site. As with this enzyme, many other variations in mechanism across the enzymes of the SF are
also likely and suggest that additional residues in SSL subgroup proteins could play critical roles in whatever specific function(s) they catalyze. Thus, characterization of some of these SSL unknowns will likely reveal other complex variations, allowing us to understand better how the N6P active site architecture has evolved to support a variety of different reactions.

3.5 Summary

In this article, we examine the largely uncharacterized SSL subgroup of the N6P SF in the context of the known structure-function relationships of the SF. The results of this global analysis lead to the prediction that the great majority of these so-called SSL proteins do not catalyze the SS reaction but rather catalyze hydrolytic reactions typical of the arylesterase-like and SGL subgroups instead. Experimental evidence for hydrolytic activity in two SSL subgroup enzymes, APMAP and the enzyme from Vitis vinifera, together with a phylogenetic analysis, suggests that the SS function could have arisen from an ancestor with a metal-coordinating active site. Based on domain organization, operon context and putative active site residues, we suggest that some of the SSL proteins may perform biological roles in bacterial ABC transporter systems. Finally, we demonstrate that despite the relative outlier status of the true SSs in reaction and active site architecture compared to other SF members, they share some similar structural features and have retained a common mechanistic strategy involving nucleophilic attack on an electrophilic substrate that supports their unification with the rest of the SF. Overlaid on this common mechanistic strategy, the very substantial differences between
the sequences and active site structures among different members of the N6P SF specify their very distinct overall reactions.

3.6 Access to Data from this Work

The data presented here for the SSL subgroup, including SSL and SS sequence and structure data and full-length alignments of representative sequences identifying key conserved amino acids, have been added to the Structure-Function Linkage Database (SFLD) (http://sfld.rbvi.ucsf.edu). Interactive versions of the networks, allowing users to examine the larger similarity context for specific proteins in the set are freely available for download.

3.7 Acknowledgements

This work was supported by NIH R01 (GM074820) to SEO and NIH R01 GM60595 and U54 GM093342 to PCB. AEB was supported by the ARCS Foundation and the PhRMA Foundation Predoctoral Informatics fellowship. LAG was supported by a National Science Foundation Predoctoral fellowship and the Ford Foundation Predoctoral fellowship. Molecular graphics images were produced using the UCSF Chimera package from the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (supported by NIH P41 RR001081). We also thank Peter Bernhardt for CrSTR-pET28a (+), Dan Tawfik for PON1-pET32b (+) constructs.
Chapter 4: Experimental Progress in the Expression and Purification of Strictosidine Synthase-like Proteins
In this chapter, three separate experiments that yielded promising, but ultimately intractable results for the expression and purification of SSL proteins are detailed, along with some concluding remarks about the project and potential future directions.

A promising experiment in the expression and purification of sequence 6 (gi|15596490; from *Pseudomonas aeruginosa PAO1* and annotated as ‘hypothetical protein PA1293’ in Genbank) in the pHis-Gβ1 construct was carried out with Stephanie Lucas on February 12 and 13, 2009 and is shown in Figure 4.1. Briefly, BL-21 (DE3) cells containing the construct were grown up in LB containing carbenicillin at a concentration of 100 µg/mL, expression was induced with 1 mM IPTG at an OD$_{600}$ between 0.6 and 0.8 and grown overnight in the cold room shaker. Cells were lysed by microfluidizer in a 50 mM sodium phosphate buffer containing 500 mM NaCl, pH 7.4, and eluted with an imidazole gradient from a Ni-NTA column using the FPLC. From Figure 4.1, soluble protein was produced, though more was insoluble than soluble. The majority of the protein appeared to bind to the column and eluted at the end of fraction A5. The protein was somewhat pure, but at low amounts and would likely require additional purification steps, in which more protein would be lost. A brute-force method, in which large volumes (likely several liters) would be needed, may result in a greater amount of protein than the yields seen here (estimated from Figure 4.1 as being on the order of tens of micrograms).
Figure 4.1. Expression Analysis and Purification Attempt of Sequence 6 (gi|15596490; from *Pseudomonas aeruginosa* PAO1 and Annotated as ‘hypothetical protein PA1293’ in Genbank) in the pHis-Gβ1 Construct Carried out with Stephanie Lucas on February 12 and 13, 2009. The top photo is a negative contrast image of a 10% Bis-Tris SDS-PAGE gel, stained with Invitrogen’s SimplyBlue SafeStain. Lanes from left to right: Lane 1. BioRad broad-range standard. Lane 2. The soluble portion of the cell lysate of the pHis-Gβ1 + sequence 6 construct. Lane 3. The insoluble portion of the cell lysate of the pHis-Gβ1 + sequence 6 construct. Lane 4. The flow-through from a nickel-NTA affinity column. Lane 5. Fraction A1. Lane 6. Fraction A4. Lane 7. Fraction A5. Lane 8. Fraction A6. Lane 9. Fraction A8. The bottom portion of the figure is a Western Blot of the same gel with equivalent samples loaded in the same lanes. Additionally, a positive control of a purified His-tagged menC is loaded in Lane 10. Underneath the photos is the 280 nm trace from the purification run.
Figure 4.2 shows an experiment from April 24, 2009 involving the same protein in Figure 4.1, with a few modifications to the protein and expression protocol. First, a predicted N-terminal signal peptide was removed from the DNA sequence encoding the protein, resulting in a construct termed pHis-Gβ1 + sequence 6 truncated. Additionally, the chaperone proteins GroEL/ES were co-expressed. Briefly, a cell line of BL-21(DE3) containing the plasmid encoding GroEL/ES was generated and made electrocompetent. The pHis-Gβ1 + sequence 6 truncated construct was transformed into this cell line. Expression conditions were the same as previously described, with the only modification being the addition of a solution of 0.2% L-arabinose to induce chaperon expression.

Purification was done on Ni-NTA agarose resin using four concentrations of imidazole. A casein-MgATP wash step (as described in Ref. 88) was used to wash off excess GroEL that was bound to the sequence 6 truncated protein (later, a milk-MgATP wash buffer was developed, making this protocol amenable to FPLC purifications). Unfortunately, yields of purified protein were quite small, as seen in the SDS-PAGE gel, and not all of the GroEL could be removed. Additionally, degradation products were seen in the Western Blot.
Figure 4.2. Expression Analysis and Purification Attempt of Sequence 6 Truncated (gi|15596490; from Pseudomonas aeruginosa PAO1 and Annotated as ‘hypothetical
protein PA1293' in Genbank with a Predicted Signal Peptide Removed) in the pHis-Gβ1 from April 24, 2009. The top photo is an image of a 10% Bis-Tris SDS-PAGE gel, stained with Invitrogen’s SimplyBlue SafeStain. Lanes from left to right: Lane 1. BioRad broad-range standard. Lane 2. The soluble portion of the cell lysate of the pHis-Gβ1 + sequence 6 construct. Lane 3. The insoluble portion of the cell lysate of the pHis-Gβ1 + sequence 6 construct. Lane 4. The flow-through from a nickel-NTA agarose resin. Lane 5. The elution fraction without a casein-MgATP wash. Lane 6. The flow-through using casein-MgATP wash. Lane 7. Elution with 50 mM imidazole. Lane 8. Elution with 75 mM imidazole. Lane 9. Elution with 100 mM imidazole. Lane 10. Elution with 520 mM imidazole. The bottom portion of the figure is a Western Blot of the same gel with equivalent samples loaded in the same lanes.

Expression in *Pichia pastoris* also showed some promising results, but unfortunately suffered from low protein yields as well. In Figure 4.3, an experiment trying to express human APMAP (gi|24308201 with the predicted N-terminal signal peptide removed, termed here ‘24308201 truncated’) was performed. On Thursday, March 11, 2010, a 5 mL culture of YPD containing 500 µg zeocin was inoculated with a single colony from a streaked plate of colony 7 of X33 cells containing the pPICZ-alpha + 24308201 truncated construct (located in a box in the -80°C freezer). After growth overnight at 30°C, 2 mL were used to inoculate 1 L of YPD containing 100 mg zeocin and grown at 30°C for 50.5 hours. Cells were spun down and resuspended in 350 mL BMM. Induction was maintained by adding 1.75 mL 100% methanol at 12, 23 and 35.3 hours. At 40 hours, the cells were spun down and the BMM media was concentrated using a 30kDa membrane. When the volume decreased to ~100 mL, an Amicon Ultra filter was used to concentrate the material further to ~15 mL. A dark precipitant was noticed almost right away after switching concentrators, suggesting that the protein may not have been very stable at that concentration. 7.5 mL of the sample was flash frozen in cryotubes containing a final glycerol concentration of 20%. The other 7.5 mL was dialyzed overnight (in 4 L of 50 mM Tris, pH 8.0 at 4°C) after diluting the sample to 40 mL in 50 mM Tris, pH 8.0 and
placing in 12-14 kDa molecular weight cutoff tubing. A Q-sepharose FF anion exchange column was used in an attempt to further purify the protein, but visualization of the protein post-purification proved difficult, even by Western Blot. The use of 25 mM octyl-thioglucoside (OTG) pre-dialysis resulted in no degradation products (Figure 4.3), though the lack of visualized degradation products may have been due to differences in the total amount of protein loaded on the SDS-PAGE.
Figure 4.3. Human AP MAP Expression and Purification Experiment in *Pichia pastoris* from March 11 to March 19, 2010. Top photos are of the same 10% Bis-Tris SDS-PAGE gel, stained with Invitrogen’s SimplyBlue SafeStain with the image on the right with the gamma setting adjusted. Lanes from left to right: Lane 1. BioRad broad-range standard. Lane 2. 1:5 dilution of frozen glycerol stock diluted in 50 mM Tris, pH 8.0, prior to
dialyzing in 4 L of 50 mM Tris, pH 8.0. Lane 3. Sample of same protein post-dialysis without OTG overnight at 4°C. Lane 4. Concentrated sample of lane 2. Lane 5. Sample of protein with 25 mM OTG. Lane 6: Sample of same protein post-dialysis with OTG overnight at 4°C. Lane 7. Concentrated sample of lane 5. Bottom photos are from Western Blots of the same gels with equivalent samples loaded in the same lanes. The one on the left was exposed for 1 minute, and the one on the right was exposed for 5 minutes.

The experimental tractability of this system remains a problem. The Vitis vinifera protein, gi|147772032, described in Chapter 3, provided limited success, but enough to give some experimental evidence that supported our predictions about functions of SSL proteins. One apparent trick was the use of detergent (0.1% tergitol in the case of gi|147772032), which, in the proteins I attempted to work with, was rarely used and dialyzed out whenever possible. Perhaps detergent should remain in all buffers to maintain protein stability (OTG addition resulted in no degradation products in Figure 4.3, though the lack of visualized degradation products may have been due to differences in the total amount of protein loaded on the SDS-PAGE). An alternative way to improve stability and solubility of these proteins may be through ancestral reconstruction (for a recent review, see Ref. 89). A third option could be purification from the organisms that express these proteins natively, though information is lacking regarding conditions such as where, when and how much of these proteins are expressed in their native context. Until these proteins can be isolated in sufficient quantities for biochemical analysis, little more can likely be said about the specific biochemical functions of these proteins.
Concluding Remarks

“Thus, science, in the very act of solving problems, creates more of them.”

~Abraham Flexner

In a purely intellectual sense, Flexner’s comment describes the very essence of science; as we seek answers to questions, we inevitably are led to new questions, needing new answers.

In the case of the strictosidine synthase-like proteins (and indeed, the rest of the N6P superfamily), many questions remain to be addressed. What are the native activities of these thousands of uncharacterized proteins? What is the biological role of the human adipocyte plasma membrane-associated protein? Can novel strictosidine synthase enzymes with greater substrate flexibility be engineered using strictosidine synthase-like proteins as a starting point? What function(s) do the predicted metal coordinating ligands of the hemomucin immune receptor (or the luciferin-regenerating enzyme) serve? What kinds of enzymatic activities remain to be discovered in the N6P SF and how did they evolve?

To answer these questions, greater functional characterization of the N6P SF will be required. From a biochemistry perspective, the most pressing concern is the current limitation in producing necessary amounts of protein for enzymatic characterization.
However, as protein expression and purification technologies continue to develop, this concern may diminish, ultimately leading to new answers and questions to be asked.
Bibliography

gluconolactonase in L-ascorbic acid biosynthesis, and its knockout mice are prone to scurvy. Proc Natl Acad Sci U S A 2006;103(15):5723-5728.


Appendix

A1. Supplementary Material from Chapter 3: The Evolution of Function in Strictosidine Synthase-like Proteins

I. Supplementary Figures from Chapter 3

Figure 3.S1. (a) The Active Site of PON1 (pdb_id: 1v04; blue backbone, arylesterase subgroup). Orientation of the active site is the same as in Figure 3.9. The phosphate coordinated to the metal has been removed for clarity. The water thought to be activated by His115 and perform the nucleophilic attack on lactone and ester compounds is displayed as a red ball. Calcium ion is a green sphere. The four conserved metal coordinating residues (discussed in the text) are colored by element; gray: carbon, blue: nitrogen, red: oxygen. The 5th metal coordinating residue (Asn270; see footnote in the text) and the His-His dyad are colored by element; light blue: carbon, blue: nitrogen, red: oxygen. (b) The activation of water by the His-His dyad and the nucleophilic attack on an ester carbonyl as described in Ref. 35.
Figure 3.S2. *V. vinifera* SSL (gi|147772032) Exhibited Hydrolase Activity with the Model Substrate pNPAc Compared to *C. roseus* SS, and Eluent from Cultures Expressing the Empty Vectors pET32b and pET28a. Hydrolysis activity of PON1 served as a positive control. The rate of background pNPAc hydrolysis has been subtracted from rates obtained for each concentration of substrate. The error bars represent the 95% confidence of the standard deviation from three experiments.

Figure 3.S3. LC-MS Data Show *V. vinifera* SSL (gi|147772032) Does Not Catalyze the SS Reaction. (a) Positive control demonstrating formation of strictosidine product at m/z 531 by a true SS from *C. roseus*. (b) No detectable strictosidine product formation by *V. vinifera* SSL under identical conditions (gi|147772032).
II. Supplementary Tables from Chapter 3

Table 3.S1. Comparable Positions of Reactive Groups and Ligands in Active Sites for SS and DFPase Mechanisms.

<table>
<thead>
<tr>
<th></th>
<th>Phosphorous of inhibitor of DFPase (pdb id: 2gvv)</th>
<th>Nitrogen of tryptamine in SS (pdb id: 2fpb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oxygen of Asp229 in DFPase</td>
<td>(i) 3.53 Å</td>
<td>(ii) 3.33 Å</td>
</tr>
<tr>
<td>(pdb id: 2gvv)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aldehyde carbon of secollogan in SS (pdb id: 2fpc)</td>
<td>(iii) 4.20 Å</td>
<td>(iv) 2.69 Å</td>
</tr>
</tbody>
</table>

Distances measured between reactive nucleophile and electrophile in (i) DFPase and (iv) SS, respectively. Distances measured between (ii) reactive nucleophiles in DFPase and SS and (iii) reactive electrophiles in DFPase and SS.

Table 3.S2. Pairwise Structural Comparisons for Proteins Described in the Text.

<table>
<thead>
<tr>
<th></th>
<th>DFPase (pdb id: 2gvv)</th>
<th>Drp35 (pdb id: 2dg1)</th>
<th>SS (pdb id: 2fpb)</th>
<th>SS (pdb id: 2fpc)</th>
<th>PON1 (pdb id: 1v04)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DFPase (pdb id: 2gvv)</td>
<td></td>
<td>1.10 Å RMSD over 145 alpha carbons</td>
<td>1.13 Å RMSD over 100 alpha carbons</td>
<td>1.14 Å RMSD over 100 alpha carbons</td>
<td>0.87 Å RMSD over 82 alpha carbons</td>
</tr>
<tr>
<td>Drp35 (pdb id: 2dg1)</td>
<td></td>
<td></td>
<td>1.15 Å RMSD over 116 alpha carbons</td>
<td>1.15 Å RMSD over 113 alpha carbons</td>
<td>0.95 Å RMSD over 90 alpha carbons</td>
</tr>
<tr>
<td>SS (pdb id: 2fpb)</td>
<td></td>
<td></td>
<td></td>
<td>0.38 Å RMSD over 301 alpha carbons</td>
<td>1.21 Å RMSD over 89 alpha carbons</td>
</tr>
<tr>
<td>SS (pdb id: 2fpc)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.26 Å RMSD over 105 alpha carbons</td>
</tr>
<tr>
<td>PON1 (pdb id: 1v04)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Similarities are reported as root mean square deviations (RMSD) over number of alpha carbons aligned. Proteins names are colored according to subgroup; red: SGL, green: SSL, blue: arylesterase. Calculations were done in Chimera\(^6\) using the MatchMaker program\(^7\) with default settings.
A2. Screenshots of N6P SF Data in the Structure-Function Linkage Database (SFLD)

Figure S1. Screenshot of the N6P Superfamily Page in the SFLD. As detailed in the text, the superfamily is divided into three subgroups each of which contains at least one family.
Figure S2. Screenshot of the N6P SF Alignment. Highlighted in white are key residues, which are listed in the table below. The table contains the function of these residues, an evidence code for their assignment, and a reference for the evidence code.
Figure S3. Screenshot of the SS Family Page. The active site image is of a strictosidine synthase structure. Clicking on the image will download a Chimera session file containing the strictosidine structure with the same attributes as shown in the figure. The
family reaction (strictosidine synthase) is shown at bottom. Clicking on the reaction will provide a list of partial reactions and give detailed information about what is known about the mechanism, along with evidence codes and references.
Publishing Agreement
It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

[Signature]
Author Signature

[Date]
Date