VARIABILITY OF SPECIFICITY DETERMINANTS IN THE O-SUCCINYLBENZOATE SYNTHASE FAMILY

A Thesis

by

CHENXI WANG

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Approved by:

Chair of Committee,	Margaret E. Glasner
Committee Members,	James C. Hu
	Steve Lockless
Head of Department,	Gregory Reinhart

December 2012

Major Subject: Biochemistry

ABSTRACT

Understanding how protein sequence, structure and function coevolve is at the core of functional genome annotation and protein engineering. The fundamental problem is to determine whether sequence variation contributes to functional differences or if it is a consequence of evolutionary divergence that is unrelated to functional specificity. To address this problem, we cannot merely analyze sequence variation between homologous proteins that have different functions. For comparison, we need to understand the factors that determine sequence variation in proteins that have the same function, such as a set of orthologous enzymes.

Here, we address this problem by analyzing the evolution of functionally important residues in the *o*-succinylbenzoate synthase (OSBS) family. The OSBS family consists of several hundred enzymes that catalyze a step in menaquinone (Vit. K2) synthesis. Based on phylogeny, the OSBS family can be divided into eight major subfamilies. We assayed wild-type OSBS enzyme activities. The results show that the enzymes from γ-Proteobacteria subfamily 1 and Bacteroidetes have relatively low values, the enzyme from Cyanobacteria subfamily 1 is intermediate, and the values for the proteins from the Actinobacteria and Firmicutes subfamilies are relatively high. We are using computational and experimental methods to identify functionally important amino acids in each subfamily. Our data suggest that each subfamily has a different set of functionally important residues, even though the enzymes catalyze the same reaction. These differences may have accumulated because different mutations were required in each subfamily to compensate for deleterious mutations or to adapt to changing environments. We assessed the roles of these amino acids in enzyme structure and function. Our method achieved 70% successful rate to identify positions that play important roles in one family but not another. The residues P119 and A329 play important role in *D. psychrophila* but not in *T.fusca* OSBS. We also observed two class switch mutations in *T.fusca*, P11 and P22. The mutations at these two position have a similar kinetic parameters as wild-type *D. psychrophila* OSBS.

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Dr. Glasner, and my committee members, Dr. Hu, Dr. Lockless, and previously Dr.Sze for their guidance and support throughout the course of this research.

Thanks also go to my friends and colleagues and the department faculty and staff for making my time at Texas A&M University a great experience. I also want to extend my gratitude to the Robert A. Welch Foundation and Norman Hackerman Advanced Research Program and to all the undergraduate students who were willing to participate in the study.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

CHAPTER I

INTRODUCTION AND LITERATURE REVIEW

Evolution of Protein Functions

Homologous proteins can have different functions, but some aspect of that function is typically conserved. In enzymes, this conserved feature is usually an aspect of catalysis, such as a partial chemical reaction or intermediate (*1-3*). For example, all proteins in the enolase superfamily use a set of conserved active site residues to catalyze a common partial reaction in which a base abstracts a proton alpha to a carboxylate to form a metal-stabilized enolate anion intermediate (Figure 1) (*4*). Using this conserved partial reaction, proteins in the enolase superfamily catalyze at least 20 chemically diverse reactions, including dehydration, racemization, and cycloisomerization (*4, 5*).

Specificity is determined by additional catalytic, ligand binding, and other residues (*4, 6-9*). New protein functions arise from divergence of these *specificity determinants* (the subset of functionally important residues that are responsible for conferring different functions on homologous proteins).

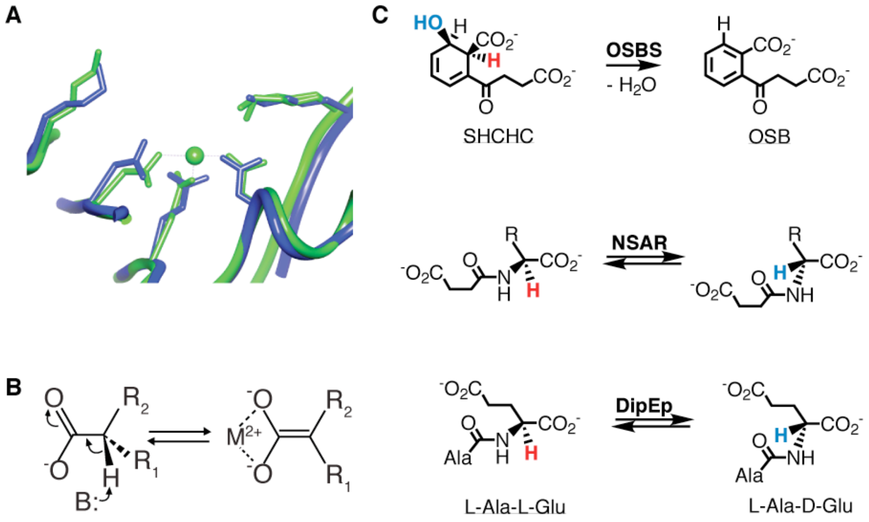

Figure 1 The enolase superfamily. Catalytic residues and a partial chemical reaction are conserved in homologous proteins that have different functions. A) The catalytic residues of o-succinylbenzoate synthase (OSBS; green; PDB entry 1FHV) and dipeptide epimerase (blue; PDB entry 1JPD) from E. coli are conserved (*10, 11*). B) OSBS, dipeptide epimerase and >18 other families in the enolase superfamily utilize the same partial chemical reaction. C) Different chemical reactions catalyzed by OSBS, dipeptide epimerase, and N-succinylamino acid racemase (NSAR).

The Misannotation Problem

Realizing the full potential of genome sequencing technology requires accurate functional annotation. Often, functional divergence cannot be predicted from global sequence similarity, because closely related proteins can have different functions, and distantly related proteins can have the same function. However, sequence similarity is still the primary criterion for functional annotation (*12*). As a result, misannotation levels are unacceptably high, with estimates ranging from 8%-30% (*13-16*). Alarmingly, misannotation rates have apparently increased sharply in recent years (*16*).

Many groups are trying to improve functional annotation methods using combinations of local or global sequence similarity, phylogeny, and genome context (*17-19*). These methods primarily transfer annotations of known functions, so their accuracy hinges on the quality and quantity of available data (*20, 21*). Indeed, inappropriate transfer of annotations to related proteins that have different functions is the main source of misannotation (*16*). A critical problem is that a miniscule fraction of sequenced proteins have been experimentally characterized. In the absence of sufficient data, the boundaries between protein families that have different functions are nebulous. Improving the accuracy of functional annotation will require both more sophisticated function prediction methods and experimental characterization to define family boundaries (Figure 2).

Predicting Functional Differences by Identifying Specificity Determinants

One solution to the misannotation problem is to use methods that predict functional differences. These methods would target proteins with novel functions for experimental characterization. Annotation of related proteins using this additional data would improve initial annotation accuracy and correct misannotation (Figure 2). Methods that identify specificity determinants are ideally suited for this application. They compare different protein families to identif y specific amino acids whose evolutionary divergence is expected to correspond to functional divergence. In addition to identifying misannotation, predicting specificity determinants can also be used to guide experimental characterization of novel protein functions. In particular, this data will be valuable for selecting libraries of compounds to use in high throughput screening or computational ligand docking (*22-24*).

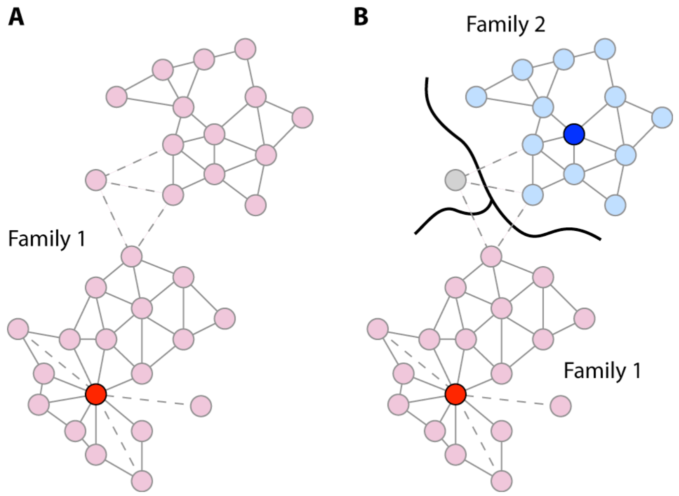

Figure 2 Correcting misannotation by characterizing new protein functions. Circles represent proteins, and the lines represent criteria for transferring functional annotations (sequence similarity, conserved motifs, operon context, etc.). Dashed lines indicate weaker associations. A) The red protein has a known function, and its annotation has been transferred to uncharacterized proteins (pink circles), often through many steps and via weak connections. B) Reannotation after discovering the function of the blue protein defines boundaries between two families that have different functions (black lines). Some proteins that are weakly connected to both families (light grey) might have a third function.

The main focus of research to develop methods that predict specificity determinants or functionally important amino acids revolves around algorithm design (*25-31*). However, the most critical question is whether the identified amino acids truly determine specificity. Homologous proteins that have the *same* function also exhibit sequence variation. Some variation is neutral, but functionally important amino acids also vary due to coevolution with adjacent amino acids or adaptation to new environments. These functionally important amino acids are not specificity determinants, but they would be identified as such by existing methods, leading them to predict that these proteins have different functions.

Existing methods for predicting specificity determinants have two other weaknesses that need to be addressed. First, some of the existing methods assume that specificity determinants are in the same location in all compared proteins (*25, 26*). As discussed below, results from our lab demonstrate that this is a faulty assumption. Second, sequence diversity of the input data is likely to affect the ability of these methods to predict specificity determinants. For example, expanding the LacI/GalR data set to include a larger number and more diverse sequences increased the number of known functionally important residues that were identified (*32*). In contrast, using Evolutionary Trace to predict specificity determinants in our model system, the *o*-succinylbenzoate synthase (OSBS) family, returned no results, suggesting that the sequences were too divergent (the average sequence identity is 28%) (*28, 33*).

Model System: the o-Succinylbenzoate Synthase Family

The experiments in this thesis utilize the *o*-succinylbenzoate synthase (OSBS) family as a model system. The OSBS family belongs to the enolase superfamily. OSBS catalyzes the conversion of 2-succinyl-6-hydroxy-2,4-cyclohexadiene-1-carboxylate

(SHCHC) to *o*-succinylbenzoate (OSB) (Figure 1C). This reaction is required for menaquinone synthesis in a wide variety of bacteria, a few Archaea, and plants.(*33, 34*) OSBS enzymes can share as little as 15% sequence identity, even though they have a single evolutionary origin and a conserved function (*33*). As a result, they are frequently misannotated as other members of the enolase superfamily (*28, 33*). Based on the phylogeny, the OSBS family was originally divided into five subfamilies. One subfamily includes proteins that catalyze a second reaction, *N*-succinylamino acid racemization (NSAR). NSAR is utilized in a pathway for converting D-amino acids to L-amino acids (*35*). All characterized enzymes with NSAR activity also have OSBS activity (A. Sakai and J. Gerlt, personal communication) (*35, 36*). Operon context indicates that some of these proteins are bifunctional *in vivo*, while the biological function of others is either OSBS or NSAR. Due to high levels of sequence similarity and the bifunctionality of some enzymes, the NSARs cannot be easily segregated into a separate family from the OSBS family. The discovery that both NSAR and OSBS are biologically relevant activities supports the hypothesis that new enzyme activities evolve through promiscuous intermediates (*33*).

Like all members of the enolase superfamily, OSBS enzymes are composed of a C-terminal catalytic $(\beta/\alpha)_7\beta$-barrel domain and an N-terminal capping domain with an $\alpha + \beta$ fold that is unique to the enolase superfamily. Two loops from the capping domain, one that is around position 20 (the 20s loop) and one that is around position 50 (the 50s loop) form the top of the active site and help determine specificity in some members of the enolase superfamily (*37, 38*). In the barrel domain, the second lysine found in a KxK motif at the end of the second beta strand is the catalytic base. The first acidic residue in the motifs DxN, ExP, and DEx on beta strands 3, 4, and 5 bind the divalent metal ion, and a lysine or arginine on beta strand 6 helps stabilize the transition state (*39*).

The catalytic motifs on strands 2-5 are the only absolutely conserved residues in the whole OSBS family. The lysine on beta strand 6 is also highly conserved, but it is replaced by arginine in one OSBS subfamily. All of these catalytic amino acids are also conserved in other members of the enolase superfamily, including the muconate lactonizing enzyme (MLE) family, the dipeptide epimerase (DE) family and a number of uncharacterized proteins. Thus, these conserved residues do not determine specificity for the OSBS reaction (*33*).

Specificity Determinants in the OSBS Family

The extreme sequence divergence of the OSBS suggested that the residues that determine specificity for the OSBS reaction are not conserved in the OSBS family. Support for this hypothesis comes from comparing the crystal structures of *E. coli* OSBS (EcOSBS) and the promiscuous OSBS/NSAR from *Amycolatopsis* sp. T-1-60 (AmyOSBS/NSAR). The relative orientation of the barrel and capping domains differs by 18°, which shifts the position of the 20s loop so that it cannot contact the product or the barrel domain the same way in the two structures (*10, 33, 40*). In addition, the conformation of the ligand is different, with the succinyl tail of OSB bent down in EcOSBS and extended in AmyOSBS/NSAR.

Two residues in EcOSBS determine this difference (*41*). Mutations at G288 of EcOSBS instroduce a steric clash that reduces catalytic efficiency >500-fold. However, Most OSBS enzymes also have glycine at this position, but the subfamily that AmyOSBS/NSAR belongs to has an aspartate at this position. Two other families in the enolase superfamily, muconate lactonizing enzyme and dipeptide epimerase, also have acidic residues at this position. Muting this residue to glycine in L-Ala-D/L-Glu epimerase from *Escherichia coli* or MLE II from *Pseudomonas* sp. P51, allows them to catalyze the OSBS reaction (*42*). Thus, G288 is a specificity determinant in some, but not all OSBS family members.

The other residue that determines substrate orientation in EcOSBS is R159, which interacts with the succinyl carboxylate of the substrate via an intervening water molecule. Mutating this position to methionine reduced efficiency 200-fold. This residue is conserved in all OSBS family enzymes except the subfamily to which AmyNSAR/OSBS belongs. In AmyOSBS/NSAR, an arginine enters the active site from a different location, which corresponds to a buried leucine in EcOSBS. Thus, R159 helps confer specificity in OSBS enzymes that bind the substrate in the "bent" conformation. However, arginine is also found at this position in the MLE family and the Firmicutes DE subfamily. In the MLE family, it is > 10 Å from the ligand, and its primary role might be to plug the bottom of the barrel. In the Firmicutes DE subfamily, this arginine has no active site accessibility, but it forms a hydrogen bond to an aspartate at the position corresponding to A107, which also contacts the amino terminus of the dipeptide substrate. Because this arginine is conserved in proteins that are not in the OSBS family, R159 would not be detected as a specificity determinant by bioinformatic sequence comparison methods. However, the observation that arginines at different positions help determine substrate binding in EcOSBS and AmyOSBS/NSAR illustrates the faultiness of the assumption that the positions of specificity determinants are conserved.

Goal of Thesis Research

The goal of the work discussed in this thesis is to address the weaknesses in current approaches to specificity determinant prediction using the OSBS family as a model system. Chapter 1 lays the groundwork for determining the optimal sequence diversity for specificity determinant prediction algorithms by redefining subfamily assignments in the OSBS family and evaluating mechanistic diversity of OSBS enzymes. Chapter 2 begins to develop a method for identifying specificity determinants that does not assume that positions of specificity determinants are conserved.

CHAPTER II

SEQUENCE AND MECHANISTIC DIVERSITY OF THE OSBS FAMILY

The initial analysis of the OSBS family determined that, in spite of sequence identities of <15%, all members have a common evolutionary origin (*33*). The OSBS family phylogeny was very similar to species trees constructed using ribosomal RNA or other proteins. To facilitate comparisons within the family, it was divided into five major subfamilies by grouping proteins from the deepest branches where the posterior probability was >0.95, as calculated using a tree constructed with MrBayes (*43*). Many new sequences have become available since the original trees were constructed (*23, 33*). The work described in this chapter updates the OSBS family phylogeny with additional sequences, reevaluates the subfamily divisions, and determines the kinetic parameters of representative OSBS enzymes. Expanding the data set was vital for developing the method in Chapter 2, and the experimental data put the results of the computational analysis into evolutionary and mechanistic perspective.

Methods

Data set

The data set was compiled by Eric Hobbs, Robert Koenig, and Dr. Glasner. Starting with our manually curated alignment from 2006, we expanded our original data set by downloading all sequences annotated as OSBS from the Structure-Function Linkage Database, which uses Hidden Markov Models (HMMs) to divide superfamilies into families of proteins that are expected to have the same function (*33*) (*44*). Because the extreme divergence of the OSBS family increases the likelihood of misannotation, we retained only proteins that share > 40% amino acid sequence identity to OSBSs that had been verified based on phylogeny and operon context (*33*). Previous results demonstrated that all proteins with > 40% sequence identity to known OSBSs fall into a monophyletic clade in the phylogeny of the MLE subgroup. This data set includes both OSBS and NSAR enzymes. NSAR enzymes cannot be segregated out on the basis of sequence similarity.

The data set was divided into clusters in which proteins share > 40% identity with at least one other protein. New sequences in each cluster were aligned to the previously aligned sequences using the profile option in MUSCLE (*45*). The resulting alignment was manually adjusted according to a structural alignment of the OSBS enzymes from *E. coli* (1FHV), *Thermosynechococcus elongatus* (2OZT)*, Desulfotalea psychrophila* (2PGE)*, Thermobifida fusca* (2QVH), *Staphylococcus aureus* (2OKT)*, and Amycolatopsis* sp. T-1-60 (1SJB). Structural alignments and all structural images were produced using the University of California, San Francisco Chimera Package from the Resource for Biocomputing, Visualization and Informatics at UCSF (supported by National Institutes of Health 2P41RR001081) (*46*). The final data set consisted of 408 sequences.

Phylogeny

The phylogeny of the whole OSBS family was determined for a representative set of 198 proteins in which no two proteins share > 70% identity. This set was selected using CD-HIT (*47, 48*). Trees were constructed using MrBayes 3.1.2 under the WAG substitution matrix and a gamma distribution to approximate rate variation among sites (*43, 49*). MrBayes was run on the CIPRES-Portal 2.0 (*50*). The results were analyzed using Tracer to evaluate tree convergence and burn-in (*51*). Trees were also constructed by maximum likelihood using the RaxML BlackBox web server (http://phylobench.vital-it.ch/raxml-bb/) with a WAG substitution matrix and a gamma distribution to approximate rate variation among sites (*52, 53*).

Protein purification

OSBS enzymes were expressed in *E. coli* strain BL21 (DE3) or *E. coli* strain BW25113 (*menC::kan*) (a gift from J.A. Gerlt, University of Illinois at Urbana-Champaign). This strain was converted to a DE3 strain to express T7 RNA polymerase using the λDE3 lysogenization kit from Novagen. Expressing the mutants in the *menC*$^{-}$ strain ensured that the purified proteins would not be contaminated with wild-type OSBS. Cultures were grown overnight at 37 ºC without induction in 300 mL of Luria-Bertani broth supplemented with carbenicillin and kanamycin. Cells were harvested by centrifugation at 1700x*g* for 15 minutes at 4 ºC. They were resuspended in buffer containing 20 mM Tris, pH. 8.0, 500 mM NaCl, and 5 mM imidazole. Resuspended pellets were lysed using a Microfluidizer (Microfluidics Corporation) at 1800 psi. After centrifugation at 18,000x*g* for 30 minutes at 4 ºC, the filtered lysate was applied to a 5 mL HisTrap FF column charged with Ni^{2+} (GE Healthcare). The protein was eluted with a buffer containing 20 mM Tris, pH. 8.0, 500 mM NaCl and 500 mM imidazole using a step to 15% elution buffer followed by a linear gradient to 100% elution buffer. Fractions containing apparently homogenous protein were identified by SDS-PAGE and pooled. Amicon Ultra-15 centrifugal filters (30 kD cutoff) (Millipore) were used to exchange the buffer and concentrate the pooled fractions. Purified proteins were stored in 10 mM Tris, pH. 8.0 and 5 mM $MgCl_2$ supplemented with 25% glycerol for storage at -80 °C. The His-tag was cleaved by Thrombin. The protein was incubated with Thrombin (2 units/mg) on ice for at least 24 hours. Keep some uncleaved proteins as control. Separate the cleaved proteins from any uncleaved protein by running the whole mixture through another Ni-NTA column under the same condition as previous purification. The cleaved proteins should be in just flow through and uncleaved protein will bind to the column and can be eluted with imidazole. Run the collected fraction via SDS-PAGE to look for a gel shift.

OSBS activity assay

Wild-type and mutant OSBS enzymes were assayed with varying concentrations of SHCHC in 50 mM Tris, pH 8.0, 0.1 mM $MnCl_2$ at 25 °C. The assays were performed by quantifying the decrease in absorbance at 310 nm ($\Delta\varepsilon$ = -2400 M^{-1} cm^{-1}), as previously described (*36, 54*). The SHCHC was synthesized by Lance Ferguson. Proteins were assayed before and after cleavage of the His-tag to determine if it affected activity. Initial rates were calculated using VisionPro (Thermo Scientific) and were fit to the Michaelis-Menten equation using Kaleidagraph (Synergy Software).

Circular dichroism

Thermal denaturation circular dichroism spectroscopy was performed on wild-type *E. coli* OSBS and several of its mutants to determine thermodynamic constants using an Aviv spectropolarimeter in the far-UV region. Samples were prepared with a concentration of 0.1 mg/mL in 50 mM inorganic potassium phosphate, 200 mM KCl, and 20% ethylene glycol buffer, pH 8.0 in one cm pathlength cuvettes. A wavelength scan was performed to determine the wavelength at which our proteins had the greatest ellipticity and to elucidate some of the structural properties of the enzyme. The wavelength of the largest peak (where ellipticity is greatest) was used as the wavelength to measure unfolding as each protein is thermally denatured. Thermal denaturation scans were conducted from 5-95 °C at 221 nm. A temperature equilibration time of three minutes was used for each increase in temperature. Temperature was increased at a rate of two degrees per interval and each measurement was averaged for 30 seconds following equilibration. Data was analyzed using Origin 6.1 software. Thermodynamic constants were estimated by fitting the data of the thermal denaturation curve to the equation:

$$\int \quad ((YN+MN*x)+(YD+MD*x)*exp(-\Delta H*(1/(x+273.15)-1/(Tm+273.15))/R))/(1+exp(-\Delta H*(1/(x+273.15)-1/(Tm+273.15))/R))$$

In the equation, YN is the intercept of the y axis for the lower flat part of the curve and MN is the slope of this section. YD is the intercept of the y axis for the upper flat part of the curve and MD is the slope of this section. X is the temperature in degrees C that was reported in each thermal denaturation curve at each point in the curve. R is the gas constant 0.001987 Kcal $K^{-1}mol^{-1}$.

Results and Discussion

Phylogeny of the OSBS family

A phylogenetic tree of the OSBS family was constructed to determine if adding new sequences altered the previously defined subfamilies, in which sequences belonging to the same bacterial phylum were grouped together. Due to the large number of sequences, the OSBS family was filtered to a nonredundant set of 198 proteins in which no two proteins share > 70% identity. The previously defined subfamilies are still well-supported (Figures 3 and 4). The smallest subfamily from our previous analysis, the Bacteroidetes subfamily, grew from 4 to 36 sequences, and expansion of the Chlorobi subfamily from one to 11 sequences defined another major subfamily.

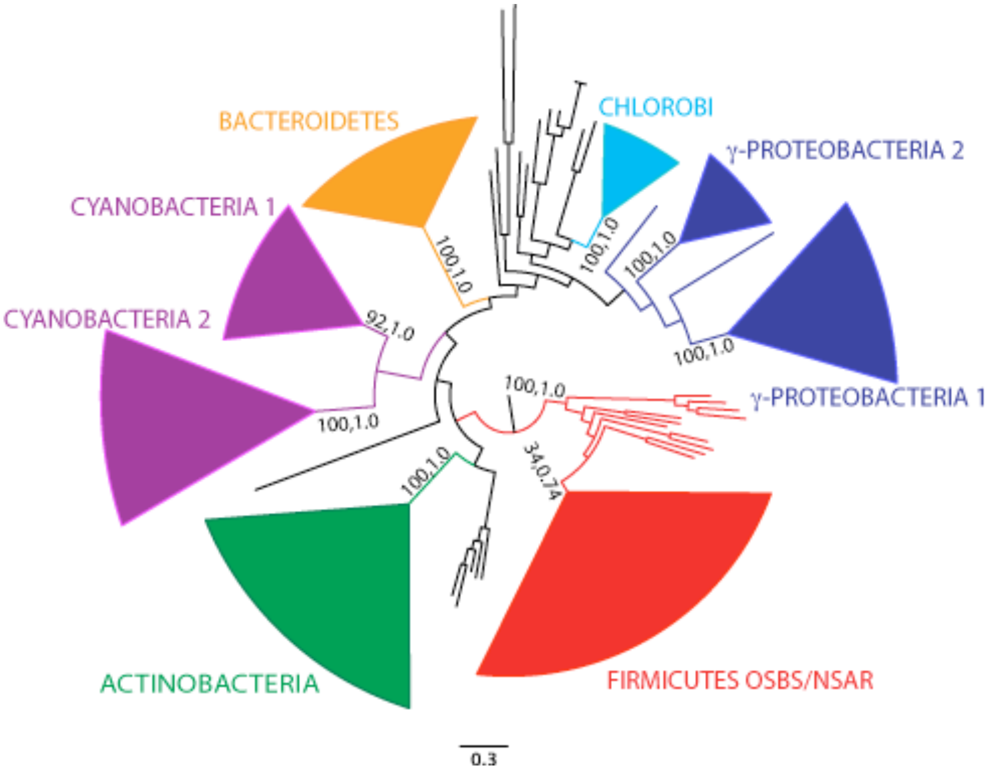

Figure 3 Division of the OSBS family into 8 subfamilies. Width of the wedges is proportional to the number of sequences, and wedge radius corresponds to the longest branch length. Proteins represented by individual branches share too little similarity to be included in the major subfamilies and are therefore left unassigned. Maximum likelihood bootstrap values and Bayesian posterior probabilities are shown for each designated subfamily. The tree is rooted based on the phylogeny of the MLE subgroup (*41*).

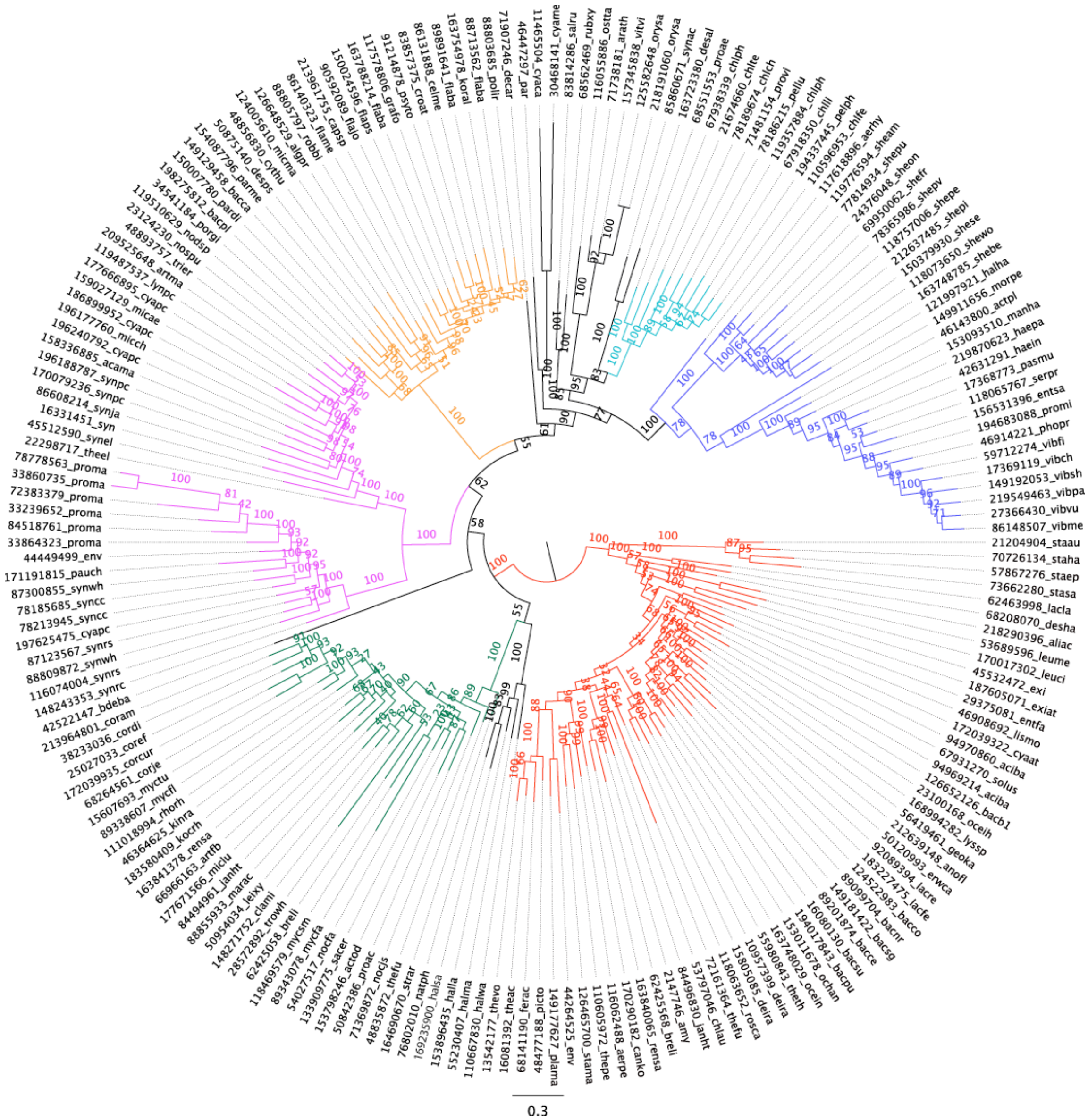

Figure 4 Full phylogenetic tree of the OSBS family constructed using MrBayes. 198 sequences sharing < 70% identity were used to build the tree. The maximum likelihood tree constructed by RaxML was in agreement concerning the subfamily divisions, although there were minor differences in topology within some subfamilies (data not shown). Branches are colored as in Figure 3 (γ-Proteobacteria are blue, Chlorobi are cyan, Bacteroidetes are orange, Cyanobacteria are magenta, Actinobacteria are green, and Firmicutes OSBS/NSAR are red). The tree is rooted based on the phylogeny of the MLE subgroup. The names of the sequences are their gi numbers followed by an abbreviated species name consisting of the first three letters of the genus and the first two letters of the species (if available). Sequences listed as "env" are from environmental sequencing projects.

However, adding more sequences made deep divisions in the previously defined subfamilies very obvious. There are two deeply branching Cyanobacteria groups, and the γ -Proteobacteria subfamily has several small groups that branch deeply. The original Firmicutes OSBS/NSAR subfamily (including the red wedge and other red branches) is also more diverse than most other subfamilies, but divergence of the basal branches does not correlate with NSAR activity. Thus, we redefined subfamilies by restricting membership to sequences that share > 40% sequence identity with at least one other subfamily member. Phylogenetic support for the redefined Firmicutes OSBS/NSAR subfamily is weak using this cutoff. However, the sequence diversity is more uniform between the redefined families, so sequence differences between subfamilies are less likely to be due to differences in evolutionary rate or divergence time.

Mechanistic differences among divergent OSBS enzymes

Prior to this research, members of only two of the eight subfamilies had been enzymatically characterized. It was noted that the k_{cat} and K_M of EcOSBS were much lower than those of AmyNSAR/OSBS, although k_{cat}/K_M was 10-fold higher for EcOSBS (*36*). Because the genome of *Amycolatopsis* ap. T-1-60 has not been sequenced and the NSAR/OSBS catalyzes the NSAR and OSBS reactions with similar efficiency, the biological function of AmyOSBS/NSAR is unknown. The kinetic parameters of AmyOSBS/NSAR are similar to those of the OSBS from *Bacillus subtilis,* which is known to have OSBS activity as its biological function because the gene is encoded in the menaquinone synthesis operon. Thus, the lower efficiency of AmyNSAR/OSBS for the OSBS reaction relative to EcOSBS is probably not due to the degerneration of an activity that is no longer biologically relevant. Instead, the change in mechanism represented by the differences in k_{cat} and K_{M} could have been an important factor in the evolution of NSAR activity in the Firmicutes OSBS/NSAR subfamily.

To begin addressing this possibility, representatives of five OSBS subfamilies were purified and assayed. These proteins came from the γ-Proteobacteria 1, Bacteroidetes, Cyanobacteria 1, Actinobacteria , and Firmicutes OSBS/NSAR subfamilies. In addition, the OSBS from *S. aureus*, which belongs to the Firmicutes phylum but which was too divergent to include in the Firmicutes OSBS/NSAR subfamily, was assayed. The kinetic parameters are shown in Table 1.

Table 1 OSBS wild-type activity assay

	Subfamily	k_{cat} (s^{-1})	K_M (μM)	k_{cat}/K_M $(M^{-1}s^{-1})$
E. coli [a]	γ-Proteobacteria 1	24±0.8	12±1.8	2.0 x 10^6
D. psychrophila	Bacteroidetes	17±1.1	15±3.4	1.2 x 10^6
T. elongatus	Cyanobacteria 1	6±0.2	362±23	1.7 x 10^4
T. fusca	Actinobacteria	188±15	464±72	4.1 x 10^5
Amycolatopsis [b]	Firmicutes	120	480	2.5 x 10^5
S.aureus	Unassigned	n.d.	n.d.	1.8 x 10^6

[a] Assayed by Wan Wen Zhu (*41*). [b] Assayed in reference (*36*). The other assays were performed by Mr.Wang.

Although removing the His-Tag from EcOSBS did not affect its activity, AmyNSAR/OSBS was inactive when purified with a His-tag. Thus, the affect of the His-Tag on the four proteins assayed in this work needed to be determined. The kinetic parameters are shown in Table 2. Removing the His-Tag from *T. fusca* did not change its activity. However, removing the His-tag from *D. psychrophila* appeared to reduce activity. The loss of activity was correlated with the length of time the protein was kept at 4 °C for cleavage and purification and was not dependent on the presence of thrombin. We also noted that the yields of this protein were significantly less than the other OSBS enzymes. This protein is probably less stable than the other OSBS enzymes because *D. psychrophila* was isolated from Arctic sediments that are ~10 °C.

It is intriguing that the magnitude of k_{cat} and K_{M} correlate with the phylogenetic relationships among the subfamilies. The enzymes from γ-Proteobacteria subfamily 1 and Bacteroidetes have relatively low values, the enzyme from Cyanobacteria subfamily 1 is intermediate, and the values for the proteins from the Actinobacteria and Firmicutes subfamilies are relatively high. This probably reflects a change in the rate-limiting step. Kinetic isotoped effects of EcOSBS and AmyNSAR/OSBS indicate that proton abstraction is at least partially rate-limiting for both of them (E.A. Taylor, personal communication) (*55*). It is possible that product release or other catalytic effects relating to substrate orientation are partially rate-limiting for EcOSBS, but not for the Actinobacteria and Firmicutes proteins. Future experiments will determine whether the mechanistic differences among these proteins was critical for the evolution of NSAR activity.

Table 2 OSBS wild-type(w His-tag and w/o His-tag) activity assay

	Subfamily	*His-tag*	*k_{cat} (s^{-1})*	*K_M (μM)*	*k_{cat}/K_M ($M^{-1}s^{-1}$)*
D. psychrophila	Bacteroidetes	with His-Tag	17±1.1	15±3.4	1.2×10^6
		after cleavage	9±0.7	60±12	1.5×10^5
T. fusca	Actinobacteria	with His-Tag	188±15	464±72	4.1×10^5
		after cleavage	228±5.6	375±31	6.0×10^5

Effects of mutating active site residues of E.coli OSBS on stability

Because the OSBS family is so divergent and the only conserved amino acids are also conserved in homologous enzymes that have different functions, we hypothesized that the subset of functionally important amino acids that determine OSBS activity have diverged, so that the locations and identities of the important non-catalytic amino acids are different in each subfamily (*33*). Structural differences between *E. coli* OSBS (γ-Proteobacteria subfamily) and *Amycolatopsis* OSBS/NSAR (Firmicutes OSBS/NSAR subfamily) support this hypothesis. The product is bound in different conformations, and the axis of orientation between the two domains in the structure is rotated by ~20 degrees relative to each other.

Mutating the active site residues of *E. coli* OSBS identified several residues that were important for activity, some of which were conserved in only a subset of the OSBS family (*41*). We are also evaluating the effects of these mutations on protein stability using circular dichroism. Previous work determined that mutating some of the charged catalytic residues stabilized the protein by relieving electrostatic repulsion (*56*). We are determining how mutating non-charged and polar amino acids in the active site affect stability (Table 3).

Table 3 *E.coli* OSBS mutants stabilities

Variants	k_{cat} (s^{-1})[a]	K_{M} (μM)	$k_{\mathrm{cat}}/K_{\mathrm{M}}$ ($\mathrm{M}^{-1}\ \mathrm{s}^{-1}$)	T_{melt}(°C)
WT	24 ± 0.8	12 ± 1.8	2.0×10^6	50.9
L48M/F51Y	27 ± 0.5	73 ± 6	3.7×10^5	48.3
S262G	10 ± 0.5	21 ± 4	4.8×10^5	51
S263G	73 ± 6	158 ± 33	4.6×10^5	48.4
S264A	12 ± 0.5	29 ± 4.7	4.1×10^5	50

[a] Kinetics were performed by Wan Wen Zhu. Circular dichroism was performed by Mr. Wang.

Two mutations were slightly destabilizing. L48M/F51Y change two residues in a loop around position 50 that form a hydrophobic substrate binding pocket to the residues found at those positions in *Amycolatopsis* OSBS/NSAR. Most mutations on these loops in both EcOSBS and AmyOSBS/NSAR decreased the yield of soluble protein (M. Hicks, S. Lucas, L. Ferguson, M. Glasner, data not shown). However, these mutations had a relatively mild effect on catalytic efficiency, reducing it ~10-fold. The other mutation that decreases stability, S263G, actually increases k_{cat} and K_M without changing catalytic efficiency. Several other mutations in *E. coli* OSBS also increase k_{cat} and K_M without changing catalytic efficiency (*41*). If they also decrease stability, that would explain why the lower k_{cat} and K_M of the wild-type enzyme are preferred.

CHAPTER III

DEVELOPING A METHOD TO IDENTIFY DIFFERENCES IN FUNCTIONALLY IMPORTANT AMINO ACIDS

The OSBS family is a good model system for developing new methods to identify specificity determinants because divergence of the subfamilies that have the same activity can be compared to the divergence of sequences that have evolved a new activity (in the Firmicutes OSBS/NSAR subfamily). This will promote the development of models to distinguish between types of amino acids that determine differences in specificity versus those that vary due to neutral mutations or covariation to maintain the structure. For example, differences in polar and charged residues in the active site would be expected to indicate a change in specificity. Current methods for identifying specificity determinants do not take this into account.

Another the weakness of existing methods for identifying specificity determinants is that they assume that the positions of specificity determinants are conserved (*25, 26, 57*). The highest scoring residues will be conserved in both groups of proteins, but the identity of the amino acid would be different. This criterion would have missed one of the critical residue differences between the γ-Proteobacteria subfamily (represented by EcOSBS) and the Firmicutes OSBS/NSAR subfamily (represented by AmyOSBS/NSAR): R159 is conserved in the γ-Protobacteria, but it is variable in the Firmicutes OSBS/NSAR subfamily (*41*).

The work in this chapter discusses the development of an algorithm that avoids this pitfall. In our description of this method, we use the word "function" to include the roles of amino acids in catalysis, binding, folding, and stability. This method is based on the observation that functionally important residues evolve at slower rates than other residues. If the residue is more important for function in one subfamily versus another, its evolutionary rate will be significantly slower in that subfamily. Although calculating evolutionary rates is computationally intensive because it requires a phylogenetic tree, it outperforms many other methods (*29*). This method will initially be validated by comparing two OSBS subfamilies that have the same function, but the algorithm is expected to be generalizable for comparing proteins that have different functions, in which differences in functionally important amino acids reflect changes in specificity as well as covariation and neutral mutations that accumulate to maintain the structure or shared aspects of function.

Methods

Phylogeny

OSBS subfamilies were defined according to the results of Chapter 1. The sequence alignment of the γ-Proteobacteria 1, γ-Proteobacteria 2, Bacteroidetes, Cyanobacteria 1, Cyanobacteria 2, Actinobacteria and Firmicutes OSBS/NSAR subfamilies were extracted from the data set described in Chapter 1. Phylogenies of each

OSBS subfamily were determined for a representative set of proteins in which no two proteins share > 95% identity using MrBayes 3.1.2 (*43*). These sets were selected using CD-HIT (*47, 48*). The parameters for MrBayes 3.1.2 were the as same as for the whole OSBS family in Chapter 1, except that the number of categories for the gamma distribution was set to eight, in order to calculate the evolutionary rates more accurately.

Calculation of evolutionary rate ratios

Raw evolutionary rates for each subfamily were calculated in MrBayes during tree construction. For each pair of subfamilies, the ratio of evolutionary rates for each aligned residue was calculated. Evolutionary rate ratios were treated as continuous distributions. Boxplots were derived to describe the distribution using the software JMP by SAS institute.

Determination of mutations to construct

After selecting sites for mutagenesis based on the boxplots, we created sequence logos of each site for the Bacteroidetes and Actinobacteria subfamilies (*58*). Sites that were predicted to be functionally important in one subfamily were changed to the most common amino acid found in the other subfamily. As a control, the same site in the other subfamily was mutated to the residue that was predicted to be important in the first subfamily. The faster-evolving sites in the other subfamily were predicted to be more tolerant of mutation.

Mutagenesis

Site-directed mutagenesis was performed by the QuickChange Mutagenesis protocol using a 2-stage PCR reaction and the primers listed in Table 4 (*59*). The templates were the *T. fusca* (GI 158430463) and *D. psychrophila* OSBSs (GI 146387140) subcloned into a pET15b vector (Novagen). For each mutagenesis experiment, two reactions were set up, each containing either the forward or reverse primer. Each reaction contained 2.5 μL 10X Pfu buffer, 200 μM of each dNTP, 1 μM forward or reverse primer, 75 ng plasmid template, and 0.5 μL Pfu Turbo polymerase (Strategene) in a total of 25 μL. Following an initial 30" denaturation step at 94 ºC, four cycles of dentaturation at 94 ºC for 30", annealing at 55 ºC for 1', and extension at 68 ºC for 12 minutes were performed. 20 μL of the forward and reverse reactions were combined, and 25 more cycles of PCR were carried out on the combined 40 μL reaction using the cycling conditions above. One μL of DpnI was added to the PCR reaction to digest the template plasmid at 37 ºC for a minimum of 3 hours. The reactions were purified using a QIAquick PCR purification kit (Qiagen), and 2 μL were transformed into electrocompetent DH5α cells. Mutations in plasmids isolated from colonies were confirmed by sequencing in both directions (Eton Bioscience, Inc.). Christopher Gajwesky designed and constructed mutations of *T. fusca* OSBS.

Table 4 A list of mutations created for each residue in 2QVH and 2PGE as well as forward and reverse primer sequences.

Mutation	Forward Primer Sequence (5'-3')	Reverse Primer Sequence (5'-3')
2PGE		
P119A	CCGATGGGCGATTTGCAGCATTGCGTTTCGC	GCGAAACGCAATGCTGCAAATCGCCCATCGG
P119R	CCGATGGGCGATTTCGCGCATTGCGTTTCGC	GCGAAACGCAATGCGCGAAATCGCCCATCGG
G348L	CCACAGGGACTGGGCACGCTGCAGCTCTATACCAAC	GTTGGTATAGAGCTGCAGCGTGCCCAGTCCCTGTGG
G348V	GGGACTGGGCACGGTTCAGCTCTATACC	GGTATAGAGCTGAACCGTGCCCAGTCCC
G217A	GTGTCGATGCCAACGCGGCATTTTCACCC	GGGTGAAAATGCCGCGTTGGCATCGACAC
G217R	GTGTCGATGCCAACCGCGCATTTTCACCCGC	GCGGGTGAAAATGCGCGGTTGGCATCGACAC
A329M	GCAATCTTGGTTTAGCCATGATTGCGCAGTGGACAGCTC	GAGCTGTCCACTGCGCAATCATGGCTAAACCAAGATTGC
L228I	CGAATGCTCCGCAGCGCATCAAGAGACTTTCCCAG	CTGGGAAAGTCTCTTGATGCGCTGCGGAGCATTCG
S29E	CACGGGGGGTGTTGACGGAAAAGCCAACTTGGTTCG	CGAACCAAGTTGGCTTTTCCGTCAACACCCCCCGTG
R284A	GAGTGCGATGCTTGATGCTATTGCTCCGCAGTACATAATC	GATTATGTACTGCGGAGCAATAGCATCAAGCATCGCACTC
I15P	CGTCGCAGTGATTTACTGTTTAAACGTCCGGCGGG	CCCGCCGGACGTTTAAACAGTAAATCACTGCGACG
Q45W	GGACATGGCGGTTGGGGGGAGGTCTCGC	GCGAGACCTCCCCCCAACCGCCATGTCC
2QVH		
R49A	CGGGAATGCGCTGCTTGGTGGGCAGCTTG	CAAGCTGCCCACCAAGCAGCGCATTCCCG
R49P	CGGGAATGCGCTCCGTGGTGGGCAGCTTG	CAAGCTGCCCACCACGGAGCGCATTCCCG
L258G	GCTTGTGGTCTGGCAACTGGCCGTCTGCTGCATGC	GCATGCAGCAGACGGCCAGTTGCCAGACCACAAGC
G133R	CGTATCGATGTTAATCGCGCGTGGGATGTTGAC	GTCAACATCCCACGCGCGATTAACATCGATACG
A238M	CGAGCGTCGGTCTGGCTATGGGTGTAGCTCTGGC	GCCAGAGCTACACCCATAGCCAGACCGACGCTCG
I144A	CAGCCGTACGCATGGCTCGCTTGCTTGACCG	CGGTCAAGCAAGCGAGCCATGCGTACGGCTG
R22E	CCGTGGTATCACTGTGGAAGAAGGTATGTTAGTTCGCGGTG	CACCGCGAACTAACATACCTTCTTCCACAGTGATACCACGG
R22S	CCGTGGTATCACTGTGAGCGAAGGTATGTTAGTTCGC	GCGAACTAACATACCTTCGCTCACAGTGATACCACGG
A196R	GTGCGCGATGCAGAACGCGCTGATGTTGTGG	CCACAACATCAGCGCGTTCTGCATCGCGCAC
P11I	GGCAGAGCGTTTGCCATTATCCTGCGCACGCGTTTC	GAAACGCGTGCGCAGGATAATGGCAAACGCTCTGCC
P11H	GAGCGTTTGCCATTCACCTGCGCACGCGTTTC	GAAACGCGTGCGCAGGTGAATGGCAAACGCTC
W33I	CGCGGTGCAGCTGGTATCGGTGAGTTTAGCCCATTC	GAATGGGCTAAACTCACCGATACCAGCTGCACCGCG

Protein purification

Wild-type EcOSBS was expressed in *E. coli* strain BL21 (DE3). Mutant EcOSBS enzymes were expressed in *E. coli* strain BW25113 (*menC::kan*) (a gift from J.A. Gerlt, University of Illinois at Urbana-Champaign). This strain was converted to a DE3 strain to express T7 RNA polymerase using the λDE3 lysogenization kit from Novagen. Expressing the mutants in the *menC*$^{-}$ strain ensured that the purified proteins would not be contaminated with wild-type OSBS. All other procedures were that same as for Chapter 1.

OSBS activity assay

OSBS activity was assayed as described in Chapter 2.

Results and Discussion

We constructed phylogenetic trees of each subfamily and calculated the evolutionary rate at each aligned residue using two methods (MrBayes and Consurf) (*43, 53, 60*). Statistical tests show that the distributions of evolutionary rates calculated by these two methods are similar (Table 5). P value (Sig.) is below the critical point that the differences between two methods are not significant. For each pair of subfamilies, the ratio of the evolutionary rates for each residue was calculated. We set the significance threshold by using boxplots of the rate ratios to identify outliers whose ratio is 1.5 x the interquartile distance and whose evolutionary rates are among the slowest 5% (Fig. 5). This is a relatively stringent threshold and may require revision as we experimentally test the predictions. Pairwise comparisons of the seven main OSBS subfamilies identified ~30 residues in each one that evolve more slowly in one subfamily versus another. In most subfamilies, a majority of these are not in the active site. This is not unexpected, because the proteins have the same activity.

Table 5 Distributions of evolutionary rates calculated by MrBayes and Consurf are similar

		Paired Differences						
					95% Confidence Interval of the Difference			
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper	df	Sig. (2-tailed)
Pair 1	MrBayes - CONSURF	-.07130	.2496	.01420	-.0992	-.0434	308	.000

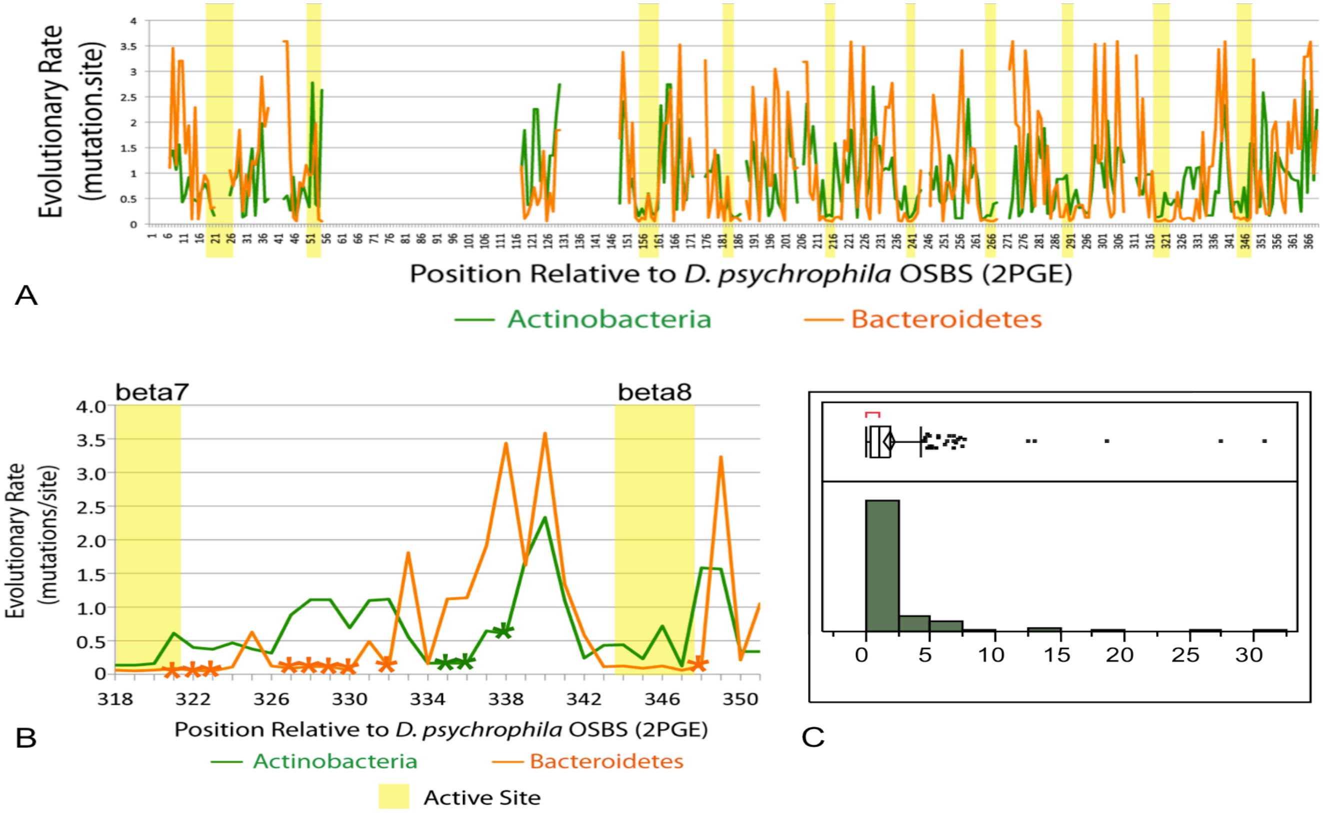

Figure 5 Identifying differences in functionally important amino acids by comparing evolutionary rates. A) Plot of the evolutionary rates calculated for the Bacteroidetes and Actinobacteria subfamilies. Highlighted regions are in the active site. B) Plot of the evolutionary rates calculated for the Bacteroidetes and Actinobacteria subfamilies. The segment between beta-strand 7 and beta-strand 8 of the C-terminal domain is shown. Highlighted regions are in the active site. Asterisks indicate residues that are outliers. C) Boxplot of the ratio of evolutionary rates between the Bacteroidetes and Actinobacteria subfamilies. The outliers evolve at least 5-fold more slowly in Bacteroidetes.

We also compared evolutionary rates among all the subfamilies individually. Calculating evolutionary rates produces a continuous distribution. Evolutionary rates are measured in substitutions per site, so the rate should correlate with the tolerance to amino acid substitutions at that site. Thus, the distribution of evolutionary rates ranks amino acids according to their expected functional importance. The statistical analysis of evolutionary rates individually are shown in Figure 6. There are differences in the subfamilies' evolutionary rates distributions. The sequence diversity and average sequence identity might cause the difference. Those might also affect the performance of bioinformatic functional prediction methods.

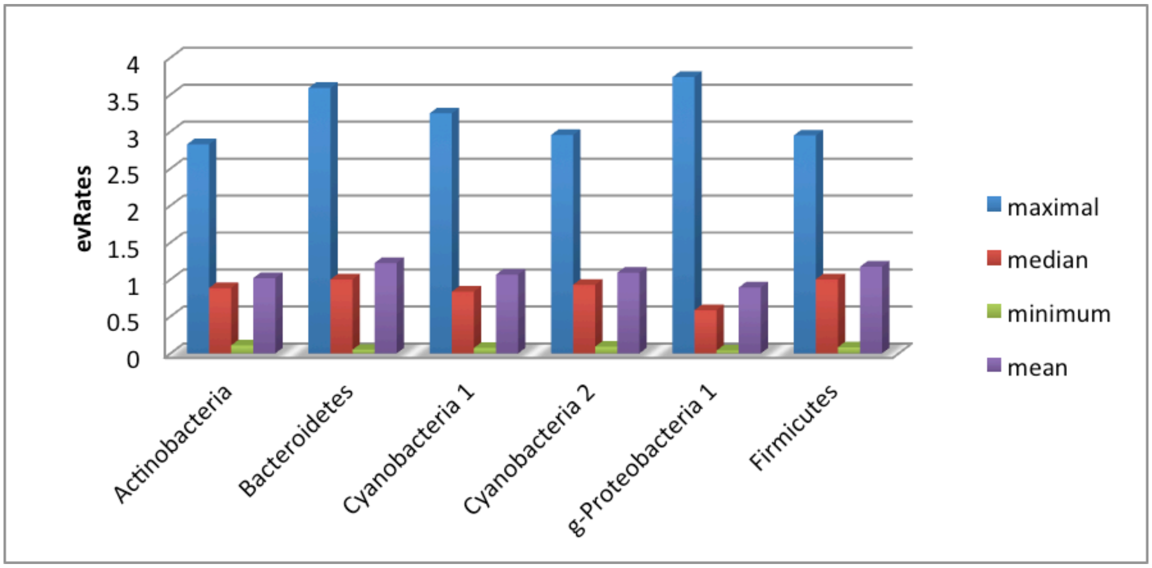

Figure 6 Descriptive statistics analysis of subfamilies raw evolutionary rates. The maximum, median, mean and minimum values of individual distributions are as shown.

Roles of functionally important amino acid in the Actinobacteria and Bacteroidetes Subfamilies

In order to determine the functional roles of the identified amino acids and to verify predictions of the evolutionary rate ratio method, the predictions from the comparison of the Bacteroidetes and Actinobacteria subfamilies were experimentally tested. These subfamilies were chosen because they have similar numbers of sequences (36 versus 42, respectively) and similar sequence diversity (54% versus 51% average sequence identity, respectively). Outliers determined from boxplots of the evolutionary ratio between the Actinobacteria and Bacteroidetes subfamilies are listed in Tables 6 and 7.

Table 6 Residues that are predicted to be more important for function in the Bacteroidetes subfamily than the Actinobacteria subfamily.

residues in *T. fusca* OSBS	Evolutionary Rate	2QVH residue #	residues in *D. psychrophila* OSBS	Evolutionary Rate	2PGE residue #	evRate in Actinobacteria / evRate in Bacteroidetes OSBS
G	2.63	43	S	0.05	55	50.80
R	1.84	49	P	0.10	119	17.63
G	1.58	133	G	0.12	217	12.78
L	1.58	258	G	0.12	348	12.75
G	0.69	239	I	0.06	330	12.03
T	1.35	95	C	0.13	180	10.79
I	0.93	144	L	0.09	228	10.36
A	1.95	263	N	0.19	353	10.15
A	1.12	241	Q	0.11	332	9.89
L	0.88	236	L	0.09	327	9.81
A	1.11	237	A	0.12	328	9.62
A	1.11	238	A	0.12	329	9.62
E	0.84	107	F	0.09	190	9.18
G	1.05	92	G	0.12	177	8.47
V	0.97	225	W	0.13	316	7.47
G	0.92	34	G	0.12	46	7.45
E	0.41	56	E	0.06	126	7.25
L	1.54	147	L	0.21	231	7.20
T	0.37	232	S	0.05	323	7.20
L	0.93	216	W	0.13	304	7.10
E	0.40	231	E	0.06	322	6.97
L	0.37	165	M	0.05	251	6.86
S	0.39	181	E	0.06	267	6.85
V	0.61	230	L	0.09	321	6.77
A	0.95	134	A	0.14	218	6.70
E	0.45	153	H	0.07	237	6.53
V	0.88	200	I	0.14	288	6.44
V	0.88	201	I	0.14	289	6.44

Table 7 Residues that are predicted to be more important for function in the Actinobacteria subfamily than the Bacteroidetes subfamily.

residues in *D. psychrophila* OSBS	Evolutionary Rate	2PGE residue #	residues in *T. fusca* OSBS	Evolutionary Rate	2QVH residue #	evRate in Bacteroidetes / evRate in Actinobacteria
N	3.41	257	R	0.11	171	30.74
G	3.04	272	R	0.11	184	27.35
A	2.08	256	R	0.11	170	18.69
	1.44		R	0.11	183	12.98
Q	1.90	275	D	0.15	187	12.52
R	1.53	284	A	0.20	196	7.59
R	3.20	11	A	0.43	7	7.49
G	3.59	43	A	0.49	31	7.32
F	1.16	34	L	0.16	26	7.10
A	3.59	273	A	0.52	185	6.88
L	1.14	336	A	0.17	245	6.86
C	0.75	255	R	0.11	169	6.77
A	1.12	335	A	0.17	244	6.76
G	3.59	44	G	0.56	32	6.46
G	2.61	199	A	0.42	116	6.28
G	1.48	364	G	0.24	275	6.08
Q	3.43	338	P	0.61	247	5.66
P	0.97	31	E	0.17	23	5.60
L	3.29	366	L	0.61	277	5.38
Q	1.42	167	A	0.28	85	5.08
G	1.97	53	E	0.39	41	5.05
I	2.29	15	P	0.46	11	4.97
G	3.54	302	L	0.72	214	4.94
Q	1.27	45	W	0.26	33	4.79
L	0.75	196	R	0.16	113	4.77
D	2.28	38	A	0.49	30	4.65
L	1.83	355	V	0.40	265	4.63
L	1.91	37	G	0.43	29	4.49

We designed mutations at several of the predicted positions. Looking at Sequence Logos of the sites predicted to be more important for function in one subfamily, we designed mutations at the highly conserved positions in one subfamily by swapping them for the most common amino acid found at the corresponding weakly conserved position in the alignment of the compared subfamily (Figure 7).

Mutations made at respectively larger evolutionary rate residues were considered negative controls, as according to our hypothesis, they should have little effect on the protein. At some positions, we designed additional mutations to alanine or mutations that caused side chain changes that could affect interactions (polar to non-polar, negative to positive and vice versa). A complete list of the mutations can be found in Table 4**.**

Effects of these mutations on protein solubility and kinetics data are listed in Table 8 and Table 9. Using UCSF Chimera, we analyzed the structures of these proteins to understand effects of the mutations we made (Figure 8).

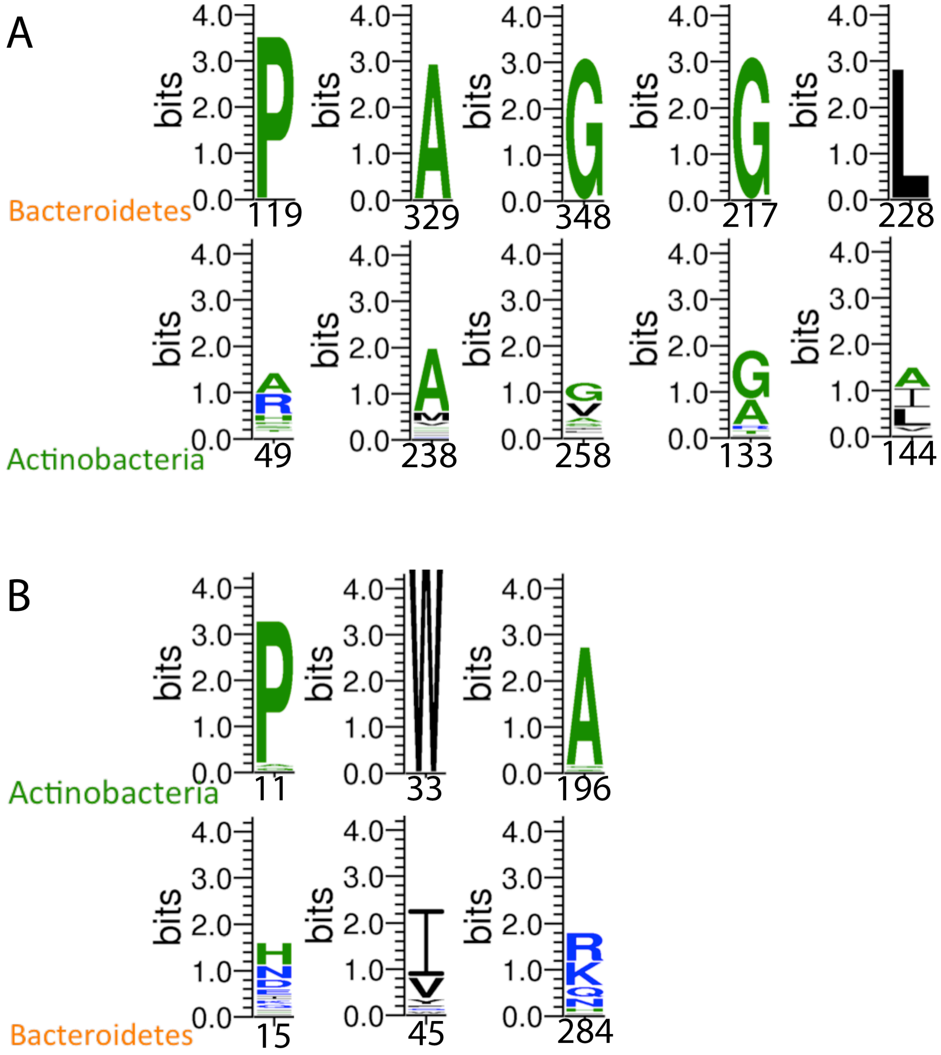

Figure 7 Weblogos used for each aligned residue studied in T. fusca OSBS (2QVH) and its aligned residue in *D. psychrophila* OSBS (2PGE). The numbers at the bottom of each picture correspond to the position in PDB structure. A.) Predicted important resides in *D. psychrophila* OSBS are showed at the top row. The corresponding residues in *T. fusca* OSBS are shown at the bottom. B.) Predicted important resides in *T. fusca* OSBS are showed at the top row. The corresponding residues in *D. psychrophila* OSBS are shown at the bottom.

Table 8 Experimental validation of evolutionary rate ratio method: effect of mutating positions that are expected to be more important for function in *D. psychrophila* (Bacteroidetes)

	k_{cat}	K_M	k_{cat}/K_M	*Actino. evRate/*		k_{cat}	K_M (μM)	k_{cat}/K_M
	(s^{-1})	(μM)	$(M^{-1}s^{-1})$	*Bacter. evRate*		(s^{-1})		$(M^{-1}s^{-1})$
D. psych. WT	17±1.1	14±3.4	1.2 x 10^{6}	--	*T. fus.* WT	188±15	464±72	4.1 x 10^{5}
P119A		insoluble		18	R49A	32±2.5	96±16.0	3.3 x 10^{5}
P119R		insoluble			R49P	n.d.	n.d.	2.0 x 10^{3}
A329M		insoluble		10	A238M	219±25	367±96	6.0 x 10^{5}
G348L	5±0.2	10 ±2.2	5.0 x 10^{5}	13	L258G	109±10	352±64	3.1 x 10^{5}
G348V	6±0.5	31 ±8.2	1.9 x 10^{5}					
G217R	0.5±0	579±111	8.6 x 10^{2}	13	G133R		insoluble	
G217A	0.08±0	176±49	4.6 x 10^{2}					
L228A		insoluble		10	I144A		insoluble	
L321A	9.8±1.3	96.3±28	1.0 x 10^{5}	7				

Table 9 Experimental validation of evolutionary rate ratio method: effect of mutating positions that are expected to be more important for function in *T. fusca* (Actinobacteria)

	k_{cat}	K_M *(μM)*	k_{cat}/K_M	*Bacter. evRate/*		k_{cat}	K_M *(μM)*	k_{cat}/K_M
	(s^{-1})		*($M^{-1}s^{-1}$)*	*Actino. evRate*		*(s^{-1})*		*($M^{-1}s^{-1}$)*
T. fus. WT	188±15	464±72	4.1×10^5	--	*D. psych.* WT	17±1.1	15±3.4	1.2×10^6
A196R		insoluble		8	*R284A*	1.6±01	60±7.4	2.6×10^4
P11I	49±4.5	43±14	1.2×10^6	5	I15P			
P11H	29±4.3	93±35	3.2×10^5					
W33I	n.d.	n.d.	7.0×10^4	5	Q45W			
R22S	13±3.7	195±91	6.7×10^4					
R22E		insoluble						

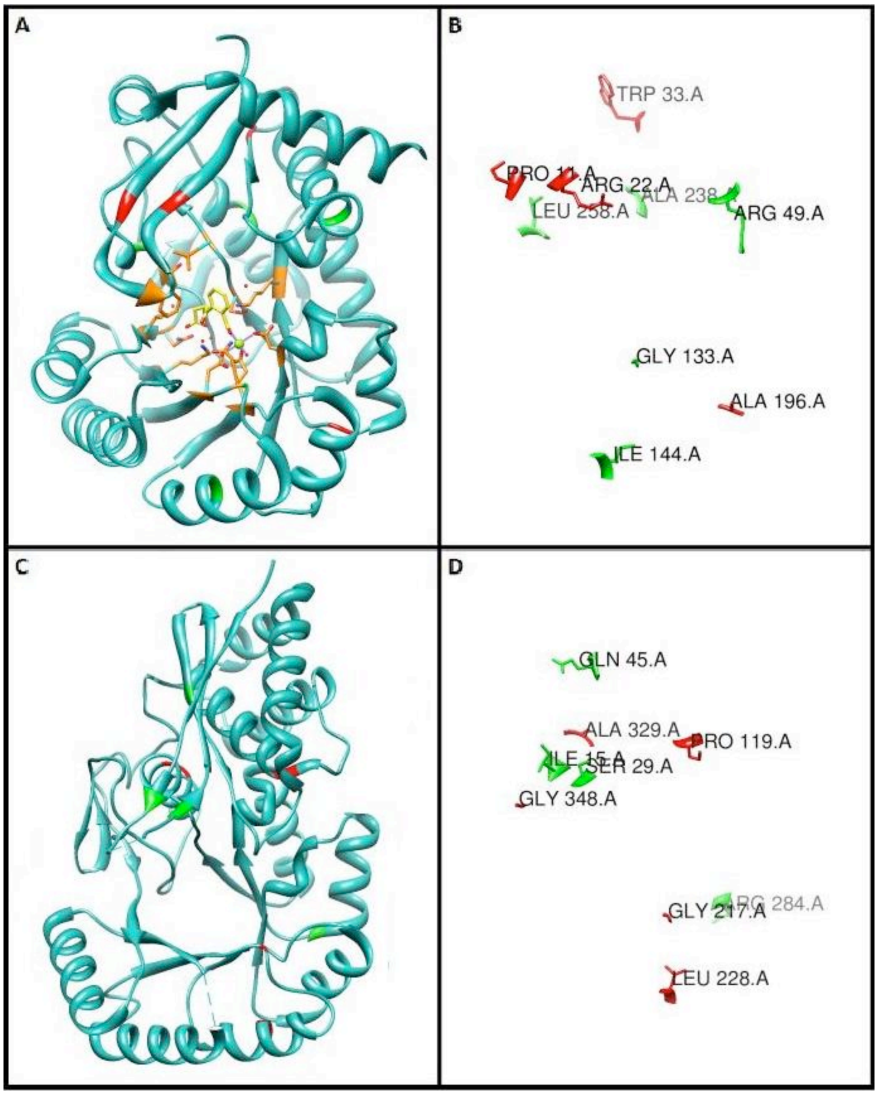

Figure 8 Structure analysis. A.) A ribbon structural depiction of *T. fusca* OSBS (2QVH) with OSB bound in the active site. Residues highlighted in red are predicted to be more important for function in Actinobacteria than Bacteroidetes based on our predictions. Residues highlighted in green are the “unimportant” residues that will be mutated. Orange residues are active site residues and are bound to OSB (yellow). B.) A depiction of *T. fusca* OSBS only showing residues that were mutated. C.) The structure of *D. psychrophila* OSBS (2PGE). Residues highlighted in red are predicted to be more important for function in Bacteroidetes than Actinobacteria based on our predictions. Residues highlighted in green are the “unimportant” residues that will be mutated. D.) The residues that will be mutated in 2PGE after hiding the rest of the structure shown in C.

In *D. psychrophila* OSBS, residues P119, L228, G217, L321 and A329 are predicted to be more important for function in the Bacteroidetes subfamily than the Actinobacteria subfamily by our method, while the aligned residues R49, A238, G133R and I144A in *T. fusca* OSBS are predicted to be negative controls. P119, L228, G217, L321 and A329 In *D. psychrophila* OSBS all have small evolutionary rates so that they evolve slower than R49, A238, G133R and I144A in *T. fusca* OSBS, which have much higher evolutionary rates. P119 and A329 fit our predictions. When we mutated P119 to alanine and arginine in *D. psychrophila*, the mutants result in insolubility. The R49A mutation in *T. fusca* OSBS does not change the OSBS efficiency, and worked as negative control as expected. A329 in *D. psychrophila* also fits our predictions. The variant A329M is insoluble while the corresponding A238M in *T. fusca* has similar OSBS efficiency as wild-type *T.fusca*.

The variants G217R and G217A in *D. psychrophila* decrease OSBS efficiency by 10,000-fold compared to wild-type, which agrees with our prediction, but G133 in *T. fusca* does not fit our prediction as a negative control. When we mutated G133 to arginine, it results in insolubility. One explanation might be that although this position tolerates mutations, an arginine adjacent to the conserved active site residue, N132 disrupts the structure. In the sequence alignment of the Actinobacteria subfamily, alanine, cysteine and threonine occur at positions aligned with G133. More conservative mutations at this site, such as G133A, might show that this site can tolerate some mutations that G217 in *D. psychrophila* cannot.

The variants G348L and G348V in *D. psychrophila* decrease OSBS efficiency by 2.5-fold and 6.3-fold, respectively. Considering the error bar, we concluded that G348L does not provide a strong evidence to support our prediction. However, the assigned residues L258 in *T. fusca* fits our prediction as a negative control.

L228 in *D. psychrophila* fits our prediction, as the variant L228A is insoluble. However the aligned residue I144 in *T. fusca* does not fit our prediction as negative control. The variant I144A is also insoluble. In the sequence alignment of the Actinobacteria subfamily, we find alanine, valine, isoleucine, and leucine at positions aligned to I144. Further conservative muations at this position in *D. psychrophila* and *T. fusca* OSBS enzymes will be required to determine if L288 in *D. psychrophila* is more tolerant of mutations. The variant L321A in *D.psychrophila* decreases OSBS efficiency by 10-fold, but we have not investigated the aligned residues in *T. fusca* OSBS.

In *T .fusca* OSBS, A196, P11 and W33 are predicted to be more important for function in the Actinobacteria subfamily than the Bacteroidetes subfamily by our evolutionary rate-ratio method. The residue A196 in *T. fusca* OSBS fits our prediction because mutating A196 to arginine leads to insolubility. This is not surprising, because A196 is buried. However the aligned residue R284 in *D. psychrophila* does not fit our prediction as a negative control very well, because variant R284A decreases OSBS efficiency by 50-fold. It is not clear why the activity of R284A should decrease, because R284 is a surface residue that does not appear to interact with anything other than water.

The residue P11 in *T. fusca* does not appear to fit our predictions. The variant P11I increases OSBS efficiency by 2-fold while P11H decreases it by ~1.3-fold, which are not significantly different from wild-type. However, both k_{cat} and K_{M} decrease 5-10-fold, indicating a change in the rate-limiting step that appears to convert the enzyme to the slow k_{cat}-class, like *D. psychrophila* and *E. coli* OSBS enzymes. Without knowing whether k_{cat} and K_{M} are under selective pressure or if only the efficiency matters, it is difficult to evaluate the success of our predictions.

It is also not clear whether W33 fits our predictions. Efficiency of the W33I mutant was ~5.8-fold lower. This is not a large drop in efficiency, but this effect was mostly due to an increase in $K_{\mathrm{M.}}$ Depending on substrate concentrations in *vivo*, this decrease could be significant. We have not investigated the aligned residues in *D. psychrophila*.

Performance evaluation of our evolutionary rate-ratio method

We evaluated the performance two ways. First, we applied an accounting method on the data in Table 8 to see how well the effects of the mutations matched our expectations. We excluded the experiments that still lack experiments for negative controls. In *D. psychrophila*, we predicted and verified five functionally important and corresponding negative control pairs of residues, excluding the duplicate mutations at position P119 of *D. psychrophila*. If the experimental result fits our prediction for residues predicted to be more important in one subfamily, we assign two credits. If the experimental result fits our prediction for the corresponding negative control, we assign one credit. If both mutations fit both our predictions, we assign 3 credits. According to this analysis, our method performance on *D. psychrophila* OSBS has a 73.3% successful rate (11/15). Second, we considered the pair of mutations at each predicted site as one entity to determine how well our method identified positions that appear more tolerant of mutations in one subfamily than another. By this metric, the success rate was 40% out of five pairs of residues.

The low success rate could be due to several factors. First, we tested a very small number of positions. Second, we selected residues to mutate that were at various ranks in Tables 6 and 7, instead of selecting only the top-ranked residues. Third, we only tested one or two amino acids at each position. Fourth, we measured success by changes to enzyme efficiency, which might not be the correct parameter. We noted above that some mutations affect k_{cat} and K_M without significantly changing $k_{cat}/K_{M,}$ and we do not know if this is important *in vivo*. Also, we have not determined how these mutations affect stability, which could be important since most of them are not in the active site.

Full evaluation of this method will require testing additional sites and correlating the results with the rank of the predicted residue pair to identify an appropriate scoring cutoff. We also need to determine if different scoring schemes, such as normalizing the evolutionary rate, improve performance. Testing a library of amino acids at the identified positions would determine the tolerance of each site to mutations in a more systematic way than selecting one or two individual mutations. Finally, experiments to test the effects of the mutations on stability and determining the extent to which k_{cat} and K_M, as opposed to k_{cat}/K_M are under natural selection will be necessary in order to determine which parameters are relevant.

CHAPTER IV

SUMMARY

The OSBS family consists of several hundred enzymes that catalyze a step in menaquinone (Vit. K2) synthesis. Based on phylogeny, the OSBS family can be divided into eight major subfamilies. We assayed wild-type OSBS enzyme activities. The results show that the enzymes from γ-Proteobacteria subfamily 1 and Bacteroidetes have relatively low values, the enzyme from Cyanobacteria subfamily 1 is intermediate, and the values for the proteins from the Actinobacteria and Firmicutes subfamilies are relatively high. We apply computational and experimental methods to identify functionally important amino acids in each subfamily. Our data suggest that each subfamily has a different set of functionally important residues. These differences may have accumulated because different mutations were required in each subfamily to compensate for deleterious mutations or to adapt to changing environments. We assessed the roles of these amino acids in enzyme structure and function. Our method achieved 70% successful rate to identify positions that play important roles in one family but not another. The residues P119 and A329 play important role in *D. psychrophila* but not in *T.fusca* OSBS. We also observed two class switch mutations in *T.fusca*, P11 and P22. The mutations at these two position have a similar kinetic parameters as wild-type *D. psychrophila* OSBS. We will test additional sites and correlate the results with the rank of the predicted residue pair to identify an appropriate scoring cutoff in future.

REFERENCES

(1) Glasner, M. E., Gerlt, J. A., and Babbitt, P. C. (2007) Mechanisms of Protein Evolution and Their Application to Protein Engineering, In *Advances in Enzymology and Related Areas of Molecular Biology, Volume 75: Protein Evolution* (Toone, E. A., Ed.), pp 193-239, Wiley & Sons.

(2) Gerlt, J. A., and Babbitt, P. C. (2001) Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies, *Annu Rev Biochem 70*, 209-246.

(3) Todd, A. E., Orengo, C. A., and Thornton, J. M. (2001) Evolution of function in protein superfamilies, from a structural perspective, *J Mol Biol 307*, 1113-1143.

(4) Gerlt, J. A., Babbitt, P. C., and Rayment, I. (2005) Divergent evolution in the enolase superfamily: the interplay of mechanism and specificity, *Arch Biochem Biophys 433*, 59-70.

(5) Rakus, J. F., Kalyanaraman, C., Fedorov, A. A., Fedorov, E. V., Mills-Groninger, F. P., Toro, R., Bonanno, J., Bain, K., Sauder, J. M., Burley, S. K., Almo, S. C., Jacobson, M. P., and Gerlt, J. A. (2009) Computation-facilitated assignment of the function in the enolase superfamily: a regiochemically distinct galactarate dehydratase from *Oceanobacillus iheyensis*, *Biochemistry 48*, 11546-11558.

(6) Glasner, M. E., Gerlt, J. A., and Babbitt, P. C. (2006) Evolution of enzyme superfamilies, *Curr Opin Chem Biol 10*, 492-497.

(7) Mildvan, A. S., Xia, Z., Azurmendi, H. F., Saraswat, V., Legler, P. M., Massiah, M. A., Gabelli, S. B., Bianchet, M. A., Kang, L. W., and Amzel, L. M. (2005) Structures and mechanisms of Nudix hydrolases, *Arch Biochem Biophys 433*, 129-143.

(8) Seibert, C. M., and Raushel, F. M. (2005) Structural and catalytic diversity within the amidohydrolase superfamily, *Biochemistry 44*, 6383-6391.

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

(10) Thompson, T. B., Garrett, J. B., Taylor, E. A., Meganathan, R., Gerlt, J. A., and Rayment, I. (2000) Evolution of enzymatic activity in the enolase superfamily: structure of *o*-succinylbenzoate synthase from *Escherichia coli* in complex with Mg^{2+} and *o*-succinylbenzoate, *Biochemistry 39*, 10662-10676.

(11) Gulick, A. M., Schmidt, D. M., Gerlt, J. A., and Rayment, I. (2001) Evolution of enzymatic activities in the enolase superfamily: crystal structures of the L-Ala-D/L-Glu epimerases from Escherichia coli and Bacillus subtilis, *Biochemistry 40*, 15716-15724.

(12) Frishman, D. (2007) Protein annotation at genomic scale: the current status, *Chem Rev 107*, 3448-3466.

(13) Brenner, S. E. (1999) Errors in genome annotation, *Trends Genet 15*, 132-133.

(14) Devos, D., and Valencia, A. (2001) Intrinsic errors in genome annotation, *Trends Genet 17*, 429-431.

(15) Jones, C. E., Brown, A. L., and Baumann, U. (2007) Estimating the annotation error rate of curated GO database sequence annotations, *BMC Bioinformatics 8*, 170.

(16) Schnoes, A. M., Brown, S. D., Dodevski, I., and Babbitt, P. C. (2009) Annotation error in public databases: misannotation of molecular function in enzyme superfamilies, *PLoS Comput Biol 5*, e1000605.

(17) Friedberg, I. (2006) Automated protein function prediction--the genomic challenge, *Brief Bioinform 7*, 225-242.

(18) Lee, D., Redfern, O., and Orengo, C. (2007) Predicting protein function from sequence and structure, *Nat Rev Mol Cell Biol 8*, 995-1005.

(19) Rentzsch, R., and Orengo, C. A. (2009) Protein function prediction--the power of multiplicity, *Trends Biotechnol 27*, 210-219.

(20) Karp, P. D. (1998) What we do not know about sequence analysis and sequence databases, *Bioinformatics 14*, 753-754.

(21) Godzik, A., Jambon, M., and Friedberg, I. (2007) Computational protein function prediction: are we making progress?, *Cell Mol Life Sci 64*, 2505-2511.

(22) Song, L., Kalyanaraman, C., Fedorov, A. A., Fedorov, E. V., Glasner, M. E., Brown, S., Imker, H. J., Babbitt, P. C., Almo, S. C., Jacobson, M. P., and Gerlt, J. A. (2007) Prediction and assignment of function for a divergent *N*-succinyl amino acid racemase, *Nat Chem Biol 3*, 486-491.

(23) Kalyanaraman, C., Imker, H. J., Fedorov, A. A., Fedorov, E. V., Glasner, M. E., Babbitt, P. C., Almo, S. C., Gerlt, J. A., and Jacobson, M. P. (2008) Discovery of

a dipeptide epimerase enzymatic function guided by homology modeling and virtual screening, *Structure. 16*, 1668-1677.

(24) Hermann, J. C., Marti-Arbona, R., Fedorov, A. A., Fedorov, E., Almo, S. C., Shoichet, B. K., and Raushel, F. M. (2007) Structure-based activity prediction for an enzyme of unknown function, *Nature. 448*, 775-779. Epub 2007 Jul 2001.

(25) Kalinina, O. V., Novichkov, P. S., Mironov, A. A., Gelfand, M. S., and Rakhmaninova, A. B. (2004) SDPpred: a tool for prediction of amino acid residues that determine differences in functional specificity of homologous proteins, *Nucleic Acids Res 32*, W424-428.

(26) Pei, J., Cai, W., Kinch, L. N., and Grishin, N. V. (2006) Prediction of functional specificity determinants from protein sequences using log-likelihood ratios, *Bioinformatics 22*, 164-171.

(27) Brandt, B. W., Feenstra, K. A., and Heringa, J. (2010) Multi-Harmony: detecting functional specificity from sequence alignment, *Nucleic Acids Res 38*, W35-40.

(28) Lichtarge, O., Bourne, H. R., and Cohen, F. E. (1996) An evolutionary trace method defines binding surfaces common to protein families, *J Mol Biol 257*, 342-358.

(29) Capra, J. A., and Singh, M. (2007) Predicting functionally important residues from sequence conservation, *Bioinformatics 23*, 1875-1882.

(30) Gloor, G. B., Martin, L. C., Wahl, L. M., and Dunn, S. D. (2005) Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions, *Biochemistry 44*, 7156-7165.

(31) Lockless, S. W., and Ranganathan, R. (1999) Evolutionarily conserved pathways of energetic connectivity in protein families, *Science 286*, 295-299.

(32) Tungtur, S., Parente, D. J., and Swint-Kruse, L. (2011) Functionally important positions can comprise the majority of a protein's architecture, *Proteins 79*, 1589-1608.

(33) Glasner, M. E., Fayazmanesh, N., Chiang, R. A., Sakai, A., Jacobson, M. P., Gerlt, J. A., and Babbitt, P. C. (2006) Evolution of structure and function in the *o*-succinylbenzoate synthase/*N*-acylamino acid racemase family of the enolase superfamily, *J Mol Biol 360*, 228-250.

(34) Meganathan, R. (2001) Biosynthesis of menaquinone (vitamin K_2) and ubiquinone (coenzyme Q): a perspective on enzymatic mechanisms, *Vitam Horm 61*, 173-218.

(35) Sakai, A., Xiang, D. F., Xu, C., Song, L., Yew, W. S., Raushel, F. M., and Gerlt, J. A. (2006) Evolution of enzymatic activities in the enolase superfamily: *N*-succinylamino acid racemase and a new pathway for the irreversible conversion of D- to L-Amino Acids, *Biochemistry 45*, 4455-4462.

(36) Palmer, D. R., Garrett, J. B., Sharma, V., Meganathan, R., Babbitt, P. C., and Gerlt, J. A. (1999) Unexpected divergence of enzyme function and sequence: "*N*-acylamino acid racemase" is *o*-succinylbenzoate synthase, *Biochemistry 38*, 4252-4258.

(37) Klenchin, V. A., Schmidt, D. M., Gerlt, J. A., and Rayment, I. (2004) Evolution of enzymatic activities in the enolase superfamily: structure of a substrate-liganded complex of the L-Ala-D/L-Glu epimerase from *Bacillus subtilis*, *Biochemistry 43*, 10370-10378.

(38) Lukk, T., Sakai, A., Kalyanaraman, C., Brown, S. D., Imker, H. J., Song, L., Fedorov, A. A., Fedorov, E. V., Toro, R., Hillerich, B., Seidel, R., Patskovsky, Y., Vetting, M. W., Nair, S. K., Babbitt, P. C., Almo, S. C., Gerlt, J. A., and Jacobson, M. P. (2012) Homology models guide discovery of diverse enzyme specificities among dipeptide epimerases in the enolase superfamily, *Proc Natl Acad Sci U S A 109*, 4122-4127.

(39) Klenchin, V. A., Taylor Ringia, E. A., Gerlt, J. A., and Rayment, I. (2003) Evolution of enzymatic activity in the enolase superfamily: structural and mutagenic studies of the mechanism of the reaction catalyzed by *o*-succinylbenzoate synthase from *Escherichia coli*, *Biochemistry 42*, 14427-14433.

(40) Thoden, J. B., Taylor Ringia, E. A., Garrett, J. B., Gerlt, J. A., Holden, H. M., and Rayment, I. (2004) Evolution of enzymatic activity in the enolase superfamily: structural studies of the promiscuous *o*-succinylbenzoate synthase from *Amycolatopsis*, *Biochemistry 43*, 5716-5727.

(41) Zhu, W. W., Wang, C., Jipp, J., Ferguson, L., Lucas, S. N., Hicks, M. A., and Glasner, M. E. (2012) Residues required for activity in Escherichia coli o-succinylbenzoate synthase are not conserved in all OSBS enzymes, *Biochemistry*.

(42) Schmidt, D. M. Z., Mundorff, E. C., Dojka, M., Bermudez, E., Ness, J. E., Govindarajan, S., Babbitt, P. C., Minshull, J., and Gerlt, J. A. (2003) Evolutionary potential of $(b/a)_8$-barrels: functional promiscuity produced by single substitutions in the enolase superfamily, *Biochemistry 42*, 8387-8393.

(43) Ronquist, F., and Huelsenbeck, J. P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models, *Bioinformatics 19*, 1572-1574.

(44) Pegg, S. C., Brown, S. D., Ojha, S., Seffernick, J., Meng, E. C., Morris, J. H., Chang, P. J., Huang, C. C., Ferrin, T. E., and Babbitt, P. C. (2006) Leveraging enzyme structure-function relationships for functional inference and experimental design: the Structure-Function Linkage Database, *Biochemistry 45*, 2545-2555.

(45) Edgar, R. C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucl. Acids Res. 32*, 1792-1797.

(46) Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004) UCSF Chimera–a visualization system for exploratory research and analysis, *J Comput Chem 25*, 1605-1612.

(47) Li, W., and Godzik, A. (2006) CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences, *Bioinformatics 22*, 1658-1659.

(48) Huang, Y., Niu, B., Gao, Y., Fu, L., and Li, W. (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences, *Bioinformatics 26*, 680-682.

(49) Whelan, S., and Goldman, N. (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach, *Mol Biol Evol 18*, 691-699.

(50) Miller, M. A., Pfeiffer, W., and Schwartz, T. (2010) Creating the CIPRES Science Gateway for inference of large phylogenetic trees, In *Proceedings of the Gateway Computing Environments Workshop (GCE)*, pp 1 - 8, New Orleans, LA.

(51) Rambaut, A., and Drummond, A. J. (2007) Tracer v1.4, In *http://beast.bio.ed.ac.uk/Tracer*.

(52) Stamatakis, A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models, *Bioinformatics 22*, 2688-2690.

(53) Stamatakis, A., Hoover, P., and Rougemont, J. (2008) A rapid bootstrap algorithm for the RAxML Web servers, *Syst Biol 57*, 758-771.

(54) Taylor Ringia, E. A., Garrett, J. B., Thoden, J. B., Holden, H. M., Rayment, I., and Gerlt, J. A. (2004) Evolution of enzymatic activity in the enolase

superfamily: functional studies of the promiscuous *o*-succinylbenzoate synthase from *Amycolatopsis*, *Biochemistry 43*, 224-229.

(55) Taylor, E. A., Palmer, D. R., and Gerlt, J. A. (2001) The lesser "burden borne" by *o*-succinylbenzoate synthase: an "easy" reaction involving a carboxylate carbon acid, *J Am Chem Soc 123*, 5824-5825.

(56) Nagatani, R. A., Gonzalez, A., Shoichet, B. K., Brinen, L. S., and Babbitt, P. C. (2007) Stability for function trade-offs in the enolase superfamily "catalytic module", *Biochemistry 46*, 6688-6695.

(57) Meinhardt, S., and Swint-Kruse, L. (2008) Experimental identification of specificity determinants in the domain linker of a LacI/GalR protein: bioinformatics-based predictions generate true positives and false negatives, *Proteins 73*, 941-957.

(58) Crooks, G. E., Hon, G., Chandonia, J. M., and Brenner, S. E. (2004) WebLogo: a sequence logo generator, *Genome Res 14*, 1188-1190.

(59) Wang, W., and Malcolm, B. A. (1999) Two-stage PCR protocol allowing introduction of multiple mutations, deletions and insertions using QuikChange site-directed mutagenesis, *Biotechniques 26*, 680-682.

(60) Landau, M., Mayrose, I., Rosenberg, Y., Glaser, F., Martz, E., Pupko, T., and Ben-Tal, N. (2005) ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures, *Nucleic Acids Res 33*, W299-302.