Studies on Enzyme Functional Evolution in the SABATH Multigene Family Using Phylogenetic and Biochemical Approaches

Ruiqi Huang

Follow this and additional works at: https://scholarworks.wmich.edu/masters_theses

Part of the Cell Biology Commons, and the Genetics Commons

Recommended Citation

This Masters Thesis-Open Access is brought to you for free and open access by the Graduate College at ScholarWorks at WMU. It has been accepted for inclusion in Master's Theses by an authorized administrator of ScholarWorks at WMU. For more information, please contact wmu-scholarworks@wmich.edu.
STUDIES ON ENZYME FUNCTIONAL EVOLUTION IN THE SABATH MULTIGENE FAMILY USING PHYLOGENETIC AND BIOCHEMICAL APPROACHES

by

Ruiqi Huang

A Thesis
Submitted to the
Faculty of The Graduate College
in partial fulfillment of the requirements for the
Degree of Master of Science
Department of Biological Sciences
Advisor: Todd Barkman, Ph.D.

Western Michigan University
Kalamazoo, Michigan
April 2012
STUDIES ON ENZYME FUNCTIONAL EVOLUTION IN THE SABATH MULTIGENE FAMILY USING PHYLOGENETIC AND BIOCHEMICAL APPROACHES

Ruiqi Huang, M.S.
Western Michigan University, 2012

Gene duplication is believed to be the major source for providing genetic materials for the innovation and diversification of protein functions; natural selection and/or neutral drift then works on these genetic materials to guide their evolutionary directions. Here, I used the salicylic acid/benzoic acid/theobromine (SABATH) multigene family to study how natural selection acted on duplicated genes to prompt functional diversification. Members in this family methylate plant secondary metabolites by transferring the methyl group from S-adenosyl-L-methionine (SAM) to the carboxyl group or ring nitrogen of the substrates. In the Apocynaceae and Solanaceae lineage of this family, I documented three putative gene duplication events. For each duplication event, I monitored the functional changes of descendents by biochemically characterizing the functions of phylogenetically resurrected ancestors, and used statistical analysis to detect if any sites were under positive selection along these daughter lineages. I also performed forward and reverse mutagenesis studies to validate the impacts of statistically predicted positively selected sites and explained the substrate specificity changes in terms of enzyme kinetics. In addition, the evolutionary pattern of each duplication event was used to fit the contemporary theoretical models for gene duplication and divergence.
ACKNOWLEDGMENTS

It gives me great pleasure in acknowledging the influence of my advisor: Dr. Todd J. Barkman, whose encouragement, supervision and support since September 2008 inspired me to pursue the subject of molecular evolution, and ultimately led to the project contained in this thesis.

Secondly, I would like to thank my committee members: Dr. Brian C. Tripp who taught me protein structures; Dr. Yirong Mo who taught me the basis of computational chemistry; and Dr. Pamela Hoppe who helped me with DNA digestion and ligation. Their instructions were invaluable to me.

Thirdly, I would like to thank the undergraduate student in my lab, Noah Sorrelle, who worked together with me to perform most of the experiments. I also thank Dr. Donald Kane, Dr. John Spitsbergen, Dr. John Geiser, Dr. Silvia Rossbach, Dr. Karim Essani and Dr. Christine Byrd-Jacobs for allowing me to use the equipment in their labs. I would particularly like to thank Dr. James Kiddle from the Chemistry Department for his suggestions regarding my project.

Lastly, I would like to thank all those who supported me in any respect during the completion of the project, especially my parents and my girlfriend, who support me all the time overseas.

Ruiqi Huang
# TABLE OF CONTENTS

**ACKNOWLEDGMENTS** ........................................................................................................ii

**LIST OF TABLES** ..................................................................................................................v

**LIST OF FIGURES** ................................................................................................................vi

**INTRODUCTION** ....................................................................................................................1

- Evolutionary fates of duplicated genes .....................................................................................1
- Neofunctionalization ..................................................................................................................2
- Subfunctionalization ..................................................................................................................3
- Dosage selection ........................................................................................................................4

**METHODS** ..........................................................................................................................10

- Sequences, alignment and phylogenetic analysis ......................................................................10
- Reconstruction of ancestral sequences .....................................................................................12
- Gene subcloning into *Escherichia coli* (*E. coli*) expression system ..................................12
- Generation of ancestral gene variants ......................................................................................13
- Expression of ancestral genes in *E. coli* ................................................................................14
- Purification of ancestral genes expressed in *E. coli* ..............................................................14
- Radiochemical ancestral protein activity assays .....................................................................15
- Determination of kinetic parameters for ancMT-D and its M201H mutant .........................15
- Enzyme product identification ................................................................................................16

**RESULTS** ............................................................................................................................17
LIST OF TABLES

1. Posterior probabilities and positions of original and mutated sites for reconstructed ancestral enzymes and their variants .............................................. 21

2. Michaelis–Menten kinetic parameters for *H. carnosa* SAMT, ancMT-D and their mutants ........................................................................................................... 30
LIST OF FIGURES

1. Comparisons among the four functional divergence models in terms of ancestral conditions, evolutionary forces and possible functional divergence outcomes.......................... 6

2. Methyl transfer reactions catalyzed by representatives of SABATH family members......................................................... 9

3. Estimated SABATH gene tree showing phylogenetic relationships among functionally characterized members............................... 18

4. Phylogenetic relationships among SAMTs, BSMTs and NAMTs in Apocynaceae and Solanaceae................................................. 18

5. Alignment of amino acid sequences from functionally characterized SABATH family members including resurrected ancestral sequences....... 26

6. Relative activities with BA (blue), SA (red), NA (green) of resurrected ancestral enzymes and their mutants at each node............................. 29

7. Michaelis-Menten kinetic graphs of ancMT-D and its M201H mutant with substrate BA and SA.......................................................... 32

8. Estimated relationships among SAMT and Solanaceae BSMT and NAMT............................................................................. 35
INTRODUCTION

Proteins are the most versatile macromolecules of the cell and they are engaged in performing most cellular functions. Major functions of proteins include binding (transporter or receptor), catalysis (almost all enzymes), operation as molecular switches (signaling G proteins), and serving as structural components (actin and tubulin) (Petsko & Ringe 2004). The diversity of these functions observed today is attributed to the evolution of proteins. In both prokaryotes and eukaryotes, the raw genetic materials for protein functional evolution come from gene duplication (Zhang 2003; Conant & Wolfe 2008). Gene duplication refers to any duplication of a piece of DNA that contains a gene; cellular events such as unequal crossing over, retrotransposition, or duplication of an entire chromosome are all cause of gene duplication (Zhang 2003). According to Kimura’s neutral theory of molecular evolution, most duplicated copies should be lost, and it would also take a long time for those surviving copies to increase enough in frequency in a population to be finally fixed (Kimura 1985). So what are the evolutionary outcomes of these surviving gene copies?

Evolutionary fates of duplicated genes

The functions of duplicated genes often determine their later evolutionary fates (Zhang 2003). One common fate of most duplicated copies is to become pseudogenes (which are not expressed or do not have any functions) by genetic disablements, because duplicated genes usually generate functional redundancy and it is often a fitness cost for organisms to have two identical genes (Zhang 2003). For example, a large number of pseudogenes exist in both the C. elegans and human genomes (Harisson et al. 2001; Harisson et al. 2002).
However, some duplicated genes will not become pseudogenes simply because the extra amount of gene products they provide is beneficial. For instance, multiple genes code for rRNAs and histones, both of which are in high demand in cells (Zhang 2003), and they are prevented from functional divergence by gene conversion or strong purifying selection. Furthermore, the serine protease gene family in rat is preserved by gene conversion (Elder & Turner 1995) and the ubiquitin multigene family contains duplicated gene copies maintained by strong purifying selection (Nei et al. 2000).

Other than becoming pseudogenes or being conserved for original functions, duplicated genes may also be stably maintained through functional divergence. Several models have been proposed to explain the maintenance and functional divergence of duplicated gene copies and in most cases, natural selection is believed to play an important role in evolutionary innovation by acting on duplicated gene copies.

Neofunctionalization

The first, or classic, neofunctionalization model for the functional divergence of duplicated genes was proposed by Ohno (1970), and the general mechanism he stated was that after gene duplication, one daughter copy may keep the original ancestral function of the gene, while the other copy, if not lost during neutral drift, is free from purifying selection to accumulate mutations that are adaptive when the environment changes and finally will acquire a new function. This model was also called “Mutation During Nonfunctionality” (MDN) by Hughes (1994), and he further added that rare beneficial mutations may be favored by positive Darwinian selection to prompt a newly duplicated gene copy to acquire a new function shortly after the gene duplication event (Hughes 1994). Examples are the duplicated immunoglobulin V region genes of
mammals that experienced positive diversifying selection after gene duplication (Hughes 1994), and the appearance of the antibacterial activity of the eosinophil cationic protein gene after duplication from the eosinophil-derived neurotoxin gene in hominoids and Old World monkeys (Zhang 1998).

Subfunctionalization

The MDN model describes the new functions that arise after gene duplication; however, there are many cases in which the ancestral gene itself has promiscuous activities (one or multiple major activities and several side activities) before gene duplication, and then the daughter copies retain a subset of ancestral functions after gene duplication. Two models are proposed to explain this subfunctionalization scenario: the "duplication, degeneration, complementation" (DDC) model (Force et al. 1999) and the "escape from adaptive conflict" (EAC) model (Hittinger & Carroll 2007). The DDC model emphasizes neutral partitioning of ancestral functions in the descendent copies after gene duplication with no improvement on each function. On the other hand, the EAC model emphasizes initial functional constraint within the multifunctional ancestor which cannot be optimized for performing each function before gene duplication, and positive Darwinian selection in resolving these functional conflicts and improving the specialized functions of each daughter copy after gene duplication. In both DDC and EAC, subfunctionalization can happen in the coding sequences or the gene regulatory elements or both. One example for the DDC model is the duplicated *Engrailed* genes in zebrafish, *Engl* and *Englb*, which have different expression domains, with *Engl* expressed in pectoral appendage buds and *Englb* expressed in specific neurons in the hindbrain and spinal cord. However, the orthologous unduplicated gene *Enl* in tetrapods
was expressed in both domains, suggesting the complementary loss of ancestral gene functions of *Engl* and *Englb* in zebrafish (Force et al. 1999). One example that fits the EAC model involves the duplicated paralogous genes *Gall* (codes for a galactokinase) and *Gal3* (works as a co-inducer) in the galactose-use pathway in *Saccharomyces cerevisiae*. Compared to the promoter of the unduplicated homologue *Gal*, which works as both a co-inducer and a galactokinase in *Kluyveromyces lactis*, the gene duplication and subsequent subfunctionalization resolved the adaptive conflict between the transcriptional regulation of the two gene functions by subdividing the transcriptional regulation of each function into two specialist genes, so that the promoter of *Gall* performs better than the promoter of unduplicated *Gal* on galactose induction (Hittinger & Carroll 2007).

Dosage selection

Recently, the innovation, amplification and divergence (IAD) model that describes the evolution of novel gene functions has been proposed by Bergthorsson et al. (2007). Like DDC and EAC, the IAD model requires a multifunctional ancestor with neither beneficial nor deleterious minor activities (the innovation) in addition to a primary major function; however, after gene duplication (the amplification), it emphasizes the continuous positive selection for the newly favored but previously neutral minor activity by increasing gene dosage through gene duplication, which then provides more selective mutational targets. Finally, this previously minor activity could be improved in any single gene copy through mutations under the same selective force, and further improved through recombination among those mutants (the divergence). During the whole process, the original function would still be maintained in at least one of these
copies by purifying selection while other copies diverge. Once one of those mutants alone could provide enough favored activity, selection on the rest of those copies is relaxed, allowing them to be removed through deleterious mutations or drift (Bergthorsson et al. 2007). The Cairns System in bacteria is an example for the IAD model. In that system, a partially defective lacZ gene (the innovation) was proved to be able to revert to the wild-type lacZ gene (the divergence) by using lactose as the positive selective force to amplify copies of the defective lacZ gene (the amplification). As the number of the partially defective lacZ genes increased, the rate of reverting to the wild-type lacZ gene by mutation also increased (Bergthorsson et al. 2007).

Figure 1 compares the similarities and differences between the four functional divergence models. The MDN model can most easily be distinguished from the other models from the functions of the ancestors (single versus multiple functions). However, distinguishing between the DDC, EAC and IAD models can be difficult because they all show subdivision of ancestral functions in the descendents, although the divergence is prompted by different kinds of evolutionary forces. To solve this problem, we need to apply statistical methods to find out the corresponding selective pressure underlying functional divergence. Evolutionary pressures on protein-coding genes can be quantified by the ratio of the number of non-synonymous substitutions per non-synonymous site (amino acid changing, $d_N$) to the number of synonymous substitutions per synonymous site (silent, $d_S$) along the evolutionary lineages: $d_N / d_S > 1$ indicates that positive selection has worked on a certain lineage to prompt adaptive functional change; $d_N / d_S < 1$ indicates that purifying selection functions to remove deleterious alleles and to stabilize the original function; $d_N / d_S = 1$ indicates that both non-synonymous and synonymous
changes will be neutral, and no selection was involved (Yang & Bielawski 2000). All these functional divergence models are driven by different evolutionary forces, so it is possible to discriminate DDC, EAC and IAD with these statistical calculations (Figure 1).

<table>
<thead>
<tr>
<th>Functional divergence models</th>
<th>Ancestral gene function</th>
<th>Gene duplication</th>
<th>Mode of selection</th>
<th>Descendent gene function</th>
</tr>
</thead>
<tbody>
<tr>
<td>MDN</td>
<td>A</td>
<td></td>
<td></td>
<td>A (Ancestral function)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( d_N/d_S &lt; 1 )</td>
<td>B (New function)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( d_N/d_S &gt; 1 )</td>
<td></td>
</tr>
<tr>
<td>DDC</td>
<td>A &amp; B</td>
<td></td>
<td></td>
<td>A (Partitioned subfunction)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( d_N/d_S = 1 )</td>
<td>B (Partitioned subfunction)</td>
</tr>
<tr>
<td>EAC</td>
<td>A &amp; B</td>
<td></td>
<td></td>
<td>A⁺ (Improved subfunction)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( d_N/d_S &gt; 1 )</td>
<td>B⁻ (Improved subfunction)</td>
</tr>
<tr>
<td>IAD</td>
<td>A &amp; B⁺</td>
<td></td>
<td></td>
<td>A (Ancestral function)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( d_N/d_S &lt; 1 )</td>
<td>B (Improved minor function)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( d_N/d_S &gt; 1 )</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1. Comparisons among the four functional divergence models in terms of ancestral conditions, evolutionary forces and possible functional divergence outcomes. MDN, the mutation during non-functionality model; DDC, the duplication, degeneration, complementation model; EAC, the escape from adaptive conflict model; IAD, the innovation, amplification and divergence model. Functions are represented by “A”, “B” (single function) or “A & B” (two functions). “⁺” means improved functions, “⁻” means minor activity. Evolutionary forces that prompt possible functional divergence are determined by the ratio of the number of non-synonymous substitutions per non-synonymous site \((d_N)\) to the number of synonymous substitutions per synonymous site \((d_S)\). \(d_N/d_S > 1\) indicates positive selection; \(d_N/d_S < 1\) indicates purifying negative selection; \(d_N/d_S = 1\) indicates neutral drift.
As discussed above, to explore the functional changes of the duplicated daughter genes, it is essential to investigate the function of the ancestral gene. Nowadays, we are able to chemically synthesize extinct ancestral genes based on phylogenetically reconstructed ancestral states, and directly investigate their functions; this paleomolecular method allows biologists to explore the environments and conditions in ancient life, and to investigate the structural bases of protein functional evolution. (Malcolm et al. 1990; Stackhouse et al. 1990; Adey et al. 1994; Jermann et al. 1995; Chandrasekharan et al. 1996; Chang et al. 2002; Gaucher et al. 2003). In addition, reconstruction of ancestral states can help to distinguish these aforementioned functional divergence models: we can test if the novel function arose de novo or from one of the ancestor’s promiscuous activities, which would distinguish MDN from DDC, EAC and IAD.

In my research, I used the known SABATH multigene family from plants, which comprises many duplicated gene members, to study how natural selection acted on the functional divergence of duplicated gene copies. For this study, I used combined phylogenetic and biochemical approaches.

The SABATH multigene family is a group of methyltransferases that transfer a methyl group from the methyl donor, S-adenosyl-L-methionine (SAM), to the carboxyl group or ring nitrogen of a substrate, forming S-adenosyl-L-homocysteine (SAH) and methyl esters such as methylsalicylate (MeSA), or N-methylated compounds such as caffeine (Figure 2). This family has recently been identified and named according to three functionally characterized members (SAMT, BAMT, and theobromine synthase) (D’Auria et al. 2003). Representatives in the family include: loganic acid methyltransferase (LAMT) (Madyastha et al. 1973), salicylic acid methyltransferase
(SAMT) (Ross et al. 1999), benzoic acid methyltransferase (BAMT) (Murfitt et al. 2000), caffeine synthase (TCS1) (Kato et al. 2000), theobromine synthase (MXMT) (Ogawa et al. 2001), jasmonic acid methyltransferase (JMT) (Seo et al. 2001), indole-3-acetic acid methyltransferase (IAMT) (Zubieta et al. 2003), benzoic acid/salicylic acid methyltransferase (BSMT) (Chen et al. 2003), farnesoic acid methyltransferase (FAMT) (Yang et al. 2006), gibberellic acid methyltransferase (GAMT) (Varbanova et al. 2007), cinnamate/p-coumarate methyltransferase (CCMT) (Kapteyn et al. 2007), anthranilic acid methyltransferase (AAMT) (Kollner et al. 2010) and nicotinic acid methyltransferase (NAMT) (Hippauf et al. 2010). Figure 2 shows some of the chemical reactions catalyzed by modern members of the SABATH gene family. The substrates share some structural similarities, e.g. all have benzene ring structures and have either the carboxyl group or ring nitrogen atoms as the methyl group acceptors. Specifically, SA differs from BA by one hydroxyl group at the C2 position of the benzene ring and NA differs from BA by one nitrogen atom replacement at the C4 position.

The SABATH family exists in many plant species (in both monocots and dicots), and the number of family members in different plant species varies due to gene duplication and recombination events: there are 24 SABATH members in Arabidopsis thaliana (D’Auria et al. 2003), 41 SABATH members (including pseudogenes) in Oryza sativa (rice) (Zhao et al. 2008), and 33 SABATH members in Populus trichocarpa (poplar) (Zhao 2008). Many of these duplicated members have different and promiscuous activities: BSMT, SAMT and NAMT can catalyze the methylation of benzoic acid (BA), nicotinic acid (NA) and salicylic acid (SA); BAMT / BSMT prefers BA over NA and SA, NAMT prefers NA over BA and SA, and SAMT prefers SA over BA and NA (Ross et al.
1999; Chen et al. 2003; Murfitt et al. 2000; Hippauf et al. 2010; Huang et al. 2012). Thus, it is an ideal system to study the evolutionary pattern of duplicated genes and subsequent functional divergence.

Figure 2. Methyl transfer reactions catalyzed by representatives of SABATH family members. Enzymes in reactions A, B, C, D and E are O-methyltransferases; enzyme in reaction F is an N-methyltransferase.

I investigated the functional divergence of enzymes in the SABATH family by resurrecting a 400-million-year old ancestor and several younger evolutionary intermediates in the Apocynaceae and Solanaceae lineage. To explore how these enzymes evolved to be able to differentiate structurally similar substrates, I chose three highly
similar substrates, BA, SA and NA, all of which can be methylated by most SAMTs, BSMTs and NAMTs. I tested the enzyme functions with the three substrates through heterologous protein expression. In addition, the evolutionary preference changes with the three substrates were used to test the aforementioned protein functional divergence models.

METHODS

Sequences, alignment and phylogenetic analysis

DNA sequences from characterized SABATH gene family members were obtained from GenBank under the following accession numbers: BT022049 for *A. thaliana* BSMT; AY224596 for *A. lyrata* BSMT; AY150400 for *A. thaliana* FAMT; AY008434 for *A. thaliana* JM; NM_118775 for *A. thaliana* GAMT1; NM_125013 for *A. thaliana* GAMT2; AK175586 for *A. thaliana* IAMT; EU375746 for *O. sativa* IAMT; EU033968 for *O. basilicum* CCMT; EU057974 for *C. roseus* LAMT; AF198492 for *A. majus* BAMT; AB048793 for *C. arabica* XMT; AB084125 for *C. arabica* DXMT; AB084126 for *C. arabica* MXMT; AB096699 for *Theobroma* MXMT; AB031280 for *C. sinensis* MXMT1; AB031281 for *C. sinensis* MXMT2; AF133053 for *C. breweri* SAMT; AF515284 for *A. majus* SAMT; AJ863118 for *H. carnos* SAMT; AJ308570 for *S. floribunda* SAMT; EF472972 for *D. wrightii* SAMT; EF472976 for *D. wrightii* BSMT; AY741483 for *C. nocturnum* SAMT; AB049752 for *A. belladonna* SAMT; AY741490 for *B. americana* SAMT; EF472978 for *B. americana* BSMT; AY233466 for *P. Hybrid* BSMT; AY741482 for *S. americana* SAMT; EF472977 for *S. americana* BSMT; AJ628349 for *N. suaveolens* BSMT1; GU014479 for *N. suaveolens* SAMT; GU014480 for *N. suaveolens* BSMT2; GU014481 for *N. suaveolens* BSMT1-2; GU014482 for *N. alata*
SAMT; GU014483 for *N. alata* BSMT2; GU014484 for *N. alata* BSMT1-2; GU014485 for *N. sylvestris* SAMT; GU014486 for *N. sylvestris* BSMT2; GU169286 for *N. gossei* NAMT; GU169287 for *N. gossei* BSMT1-2; GU169288 for *N. gossei* BSMT2; GU169289 for *N. gossei* SAMT.

Multiple sequence alignment was performed by using ClustalX (Thompson et al., 1997) with the following parameters: gap opening 15, gap extension 6.66, delay divergent sequences (%), 30, and DNA transition weight 0.5. The alignment was then adjusted manually to preserve codon integrity.

The phylogenetic tree of the SABATH family was estimated using the maximum likelihood method with 10 random addition sequences and TBR swapping in PAUP* (Swofford 2003) assuming the Hasegawa-Kishino-Yano (HKY) + invariant sites + gamma (HKY + I + G) model of nucleotide substitution as chosen by Modeltest (Posada & Crandall 1998). The HKY model accounts for different rates for transitions and transversions of sequence evolution with “I” meaning extent of static, unchanging sites in a dataset and “G” meaning gamma distributed rate variation among sites. Bootstrapping was performed using 100 replicates.

PAML version 4.2 (Yang 2007) was used to test the hypothesis of positive selection in the SABATH gene family. The branch-site test was implemented because it is expected that positive selection should act only on a subset of sites and branches of a gene tree as functional divergence occurs. In the branch-site model (Yang and Nielsen 2002), branches in the tree are divided a priori into foreground and background categories. In this case, the foreground lineage is the branch on which positive selection will be detected, whereas the background lineages are all other branches in the tree.
likelihood ratio test is then constructed to compare this selection model with the null model in which the foreground branch is not allowed to have any positively selected sites (Zhang et al. 2005). The HKY model of nucleotide substitution was assumed in the analysis.

Reconstruction of ancestral sequences

Ancestral protein sequences were estimated using Codeml in PAML assuming the phylogenetic tree in Figure 3 by using the maximum likelihood method. Because there was uncertainty of some amino acid sites, I made multiple variants of each protein. Sequences for ancMT-B, ancMT-D, ancMT-E and ancMT-F obtained from codon-based estimates using the tree in Figure 3 were used to compare with those obtained using the denser sampling of sequences in Figure 8 to choose the alternative amino acid to mutate. The mutation sites for ancMT-Z and ancMT-A were chosen by comparing sequences estimated from the same data set but assuming different roots: one root is GAMT, the other root is IAMT because this enzyme is shared by both monocots and dicots, as well as both gymnosperms and angiosperms (Zhao et al. 2008).

Gene subcloning into Escherichia coli (E.coli) expression system

Ancestral DNA sequences for node A, node B, node D and node E were synthesized and cloned into the pUC57 vector by GenScript with engineered Ndel and BamHI restriction sites. Target genes were cut with Ndel and BamHI restriction enzymes (New England BioLabs, NEB) in NEBuffer 3 for 5 hours at 37 °C and gel purified using QIAEXII Gel Extraction kit (QIAGEN). The expression vector pET15b (Novagen) with N-terminal 6×His tag was chosen and processed the same way except for the addition of
Calf intestinal alkaline phosphatase and incubation for 30 minutes at 37 °C to remove the 5' end phosphate groups before gel purification. Target genes were then ligated with pET15b at 16 °C overnight by T4 DNA ligase (NEB). The ligation reaction was transformed into *E.coli* XLI-blue supercompetent cells. The ancestral DNA sequence for Node Z was synthesized and cloned into the pGS21a vector by GenScript and then subcloned into the pEXP5-TOPO vector (Invitrogen) to get rid of the GST tag. The ancestral sequence for node F was obtained by reverse mutation of 7 amino acid sites of *N.sauveolens* BSMT1 (Pott et al. 2004), which was cloned in the pET-SUMO vector (Invitrogen). All the recombined plasmids were isolated using the QIAPrep Spin Miniprep Kit (Qiagen), following the manufacturer’s protocol, and verified by DNA sequencing.

**Generation of ancestral gene variants**

Mutation of ancestral gene variants and positively selected sites was carried out using the QuikChange Site-Directed Mutagenesis Kit (Stratagene). Purified plasmids were used as PCR templates following the manufacturer’s protocol. High fidelity *PfuTurbo* DNA polymerase was used. Mutagenic primers were designed to be between 25 and 45 bases in length, and the melting temperature (*T*_m) of the primers was calculated to be greater than or equal to 78 °C by using the following formula: 

\[
T_m = 81.5 + 0.41(\%GC) - 675/N - \%mismatch
\]

(*N* is the primer length in bases) for point mutations, or 

\[
T_m = 81.5 + 0.41(\%GC) - 675/N
\]

(*N* does not include the bases to be inserted or deleted) for insertions or deletions. Cycling parameters were as follows: template denaturation at 95 °C for 30 seconds, cycling then proceeded to denaturation at 95 °C for 30 seconds, annealing at 55 °C for 1 minute, and extension for 1 minute / kb of plasmid length at 68
°C, for 16 cycles. Plasmid transformation into *E. coli XL1*-blue supercompetent cells was performed following *Dpnl* digestion of the templates. All mutated sites were verified by DNA sequencing.

Expression of ancestral genes in *E. coli*

The recombined plasmids were transformed into *E. coli* BL21 Star™ (DE3) competent cells for protein expression. Transformed *E. coli* cells were grown at 37 °C in LB broth containing 100 µg/ml ampicillin or Kanamycin until the OD_{600} reached 0.6 to 0.8. The growths were then induced with 1.0 mM isopropyl 1-thio-galactopyranoside (IPTG) for 4 hours at 25 °C. The cells were harvested by centrifugation at 4 °C, 4000 rpm for 15 minutes. The pellets were resuspended in chilled equilibration buffer (50 mM sodium phosphate, 300 mM NaCl, pH 7), sonicated on ice for 30 seconds with a 10 second pause on ice between each of 3 bursts. The suspension was then centrifuged at 12000 rpm for 20 minutes at 4 °C to get rid of cell debris. The supernatant contained the soluble total proteins and the total proteins were either used directly for enzyme catalytic assays or purified.

Purification of ancestral genes expressed in *E. coli*

For ancMT-D and its M201H mutant, the total soluble proteins were passed over a pre-equilibrated TALON cobalt column, washed with 10 × column volume equilibration buffer, then 5 × column volume wash buffer (equilibration buffer + 10 mM imidizole, pH 7) and eluted with 500 µl elution buffer (equilibration buffer + 150 mM imidizole, pH 7) 4 times according to manufacturer instructions (Clontech, Takara Bio Group, Japan). Purified enzymes were stored in 1 mM DTT and 20% glycerol at -80 °C.
The purity and molecular mass of the purified proteins were evaluated by running on pre-cast NuPAGE 4-12% Bis-Tris gels (Invitrogen). The enzyme concentration was determined by the Bradford assay (Bradford 1976) using bovine serum albumin (BSA) as the standard.

Radiochemical ancestral protein activity assays

Radiochemical activity assays were performed in a 50 μl volume containing 50 mM Tris-HCl, pH 7.5, 1 mM substrate (BA, SA, NA) dissolved in pure ethanol, 1 μl 14C-SAM with a specific activity of 40 mCi/mmol (Perkin-Elmer, Boston, MA) and 20 μl total protein or 0.15 to 0.5 μg purified protein. As a negative control, 1 μl pure ethanol was added as the substrate. The assay was initiated by addition of SAM, maintained at 25 °C for 20 min, and stopped by addition of 200 μl EtOAc. After vigorous shaking and phase separation by centrifugation at 13,200 rpm for 30 seconds, 120 μl upper organic phase was mixed with 4 ml counting cocktail (Bio-Safe II, RPI) and counted using a liquid scintillation counter (Beckman Coulter, Fullerton, CA) to obtain the DPM (disintegrations per minute) value. Relative enzyme activity with each substrate was calculated and the product with the highest DPM value was set to 100.

Determination of kinetic parameters for ancMT-D and its M201H mutant

The reaction rates, measured with increasing concentrations of either purified ancestral proteins or substrates, were found to obey Michaelis-Menten kinetics. These assays were performed using the radiochemical assay described above except 2.5 μl of concentrated HCl was added to stop the reaction immediately. Appropriate enzyme concentrations and incubation times were then determined in time-course assays with low
unsaturated substrate concentrations, such that the reaction velocity was linear during the assay period and less than 10% substrate depletion occurred. For ancMT-D, to determine the $V_{\text{max}}$ and $K_M$ values for BA and SA, SAM was held constant at 200 μM, concentrations of BA were varied from 25 μM to 2 mM, concentrations of SA were varied from 2 μM to 200 μM, 0.5 μg purified enzyme was used for both BA and SA assays, and reactions were conducted at 25 °C for 5 minutes. Assays for the M201H mutant were carried out using the same substrate concentrations as those for ancMT-D, and the reaction time for both BA and SA was 10 minutes at 25 °C with 0.1 μg and 0.15 μg purified enzyme used for BA and SA, respectively. Initial velocities versus BA or SA concentrations were plotted using GraphPad Prism (GraphPad Software, La Jolla, CA) to fit the hyperbolic Michaelis-Menten equation to calculate the $V_{\text{max}}$ and $K_M$. $V_{\text{max}}$ was then converted to apparent $k_{\text{cat}}$ and expressed in units of s$^{-1}$. Assays were done in duplicate and standard errors were calculated using GraphPad Prism.

Enzyme product identification

Transformed BL21 (DE3) cells containing the recombinant plasmids were cultured in the same condition as above. After induction with 1mM IPTG for 4 hours at 25 °C, 1 ml of 10 mM substrate was added to the culture and growth for either 30 minutes at 25 °C for ancestral enzymes at node A, node B, node D, node E and node F or overnight at 25 °C for the ancestral enzyme at node Z. After the cells were harvested by centrifugation at 4 °C, 4000 rpm, the supernatant was extracted with 4 ml of hexane. The hexane phase was concentrated down to 200 μl by air flow and 1 μl of concentrated product was analyzed by gas chromatography-mass spectrometry (GC-MS) using the following oven conditions: 40 °C for 2 minutes, ramping 20 °C / minute to 300 °C with a
2 minute hold. Products were identified by comparing their mass spectra and retention times with available standards from the library of National Institute of Standards and Technology. 0.0188 μmol of pure toluene was added to the 200 μl hexane extracts as an internal standard to quantify the product of ancMT-Z and its mutants; assays were done in triplicate.

RESULTS

Phylogenetic relationships of characterized contemporary SABATH members

To determine the phylogenetic relationships among characterized members in the SABATH gene family, I extracted the gene sequences from GenBank and aligned them using ClustalX with manual adjustment to preserve codon integrity. An unrooted phylogenetic tree, shown in Figure 3, was then obtained using the maximum likelihood method in PAUP* based on the sequence alignment. SAMTs, BSMTs and NAMTs from Apocynaceae and Solanaceae form a monophyletic lineage and are divergent from the other SABATH family members (Figure 3 and Figure 4). Previous studies (Pott et. al, 2004; Hippauf et. al 2010) which determined the substrate profiles of these SABATH members from Apocynaceae and Solanaceae, have shown that almost all SAMTs, BSMTs and NAMTs can methylate BA, SA and NA, with all SAMTs preferring SA over BA and NA, all BSMTs preferring BA over SA and NA, and all NAMTs preferring NA over BA and SA. Here we define preference to be relative enzyme activities. Detailed phylogenetic relationships of SAMTs, BSMTs and NAMTs in Apocynaceae and Solanaceae are shown in Figure 4.
Figure 3. Estimated SABATH gene tree showing phylogenetic relationships among functionally characterized members. The SAMTs, BSMTs, and NAMTs from Apocynaceae and Solanaceae form a monophyletic lineage. Bootstrap percentages (>80) for each lineage are given. Circles point out the nodes where three ancestral amino acid sequences were reconstructed.

Figure 4. Phylogenetic relationships among SAMTs, BSMTs and NAMTs in Apocynaceae and Solanaceae. Bootstrap percentages (>80) are shown for each corresponding node. Lineages are colored according to their favored substrates: BA (blue), SA (red), and NA (green). Ancestral amino acid sequences were estimated at node
A, B, D, E and F; node A, D and E indicated by arrows 1, 2 and 3 represent nodes where putative gene duplication events occurred. All SAMTs prefer SA over BA and NA, all BSMTs prefer BA over SA and NA, and all NAMTs prefer NA over BA and SA.

Gene duplication events and resurrection of ancestral proteins

In order to understand the functional divergence of SAMT, BSMT and NAMT in Apocynaceae and Solanaceae, we need to know the function of their ancestors. Previous studies (Barkman et al. 2007; Hippauf et al. 2010; Huang et al. 2012) have postulated three gene duplication events in the Apocynaceae and Solanaceae lineage: the first gene duplication event at node A separated SAMTs, BSMTs and NAMTs in Apocynaceae and Solanaceae from the other functionally diverse SABATH members including D/MXMT, FAMT, LAMT, GAMT and IAMT (Figure 3); the second gene duplication event at node D gave rise to SAMTs and BSMTs in the Solanaceae lineage; the third gene duplication event at node E generated BSMT2, as well as BSMT1 and NAMT in *Nicotiana suaveolens* and NAMT in *Nicotiana gossei*. Therefore, ancestral amino acid sequences at nodes A, B, D, E and F were estimated using the amino acid sequences of characterized SABATH members based on the phylogenetic tree in Figure 3, and they were named ancMT-A, ancMT-B, ancMT-D, ancMT-E and ancMT-F. In addition, the amino acid sequence of a more ancient ancestor deep in the tree at node Z was also estimated and named ancMT-Z (Figure 3).

Although the level of sequence divergence in the SABATH family was high, these ancestral enzymes were still estimated with high confidence. The mean overall posterior probability for ancMT-Z is 0.77, for ancMT-A is 0.92, for ancMT-B is 0.93, for ancMT-D is 0.94, for ancMT-E is 0.94 and for ancMT-F is 0.99. For site-specific posterior probabilities, sites with lower probabilities are mostly found on the surface or
the loops (which are distant from the active site) of the resurrected enzymes, while nearly all sites in the active site of these proteins have high posterior probabilities (> 0.8). Meanwhile, to assess the robustness of the ancestral enzyme function to the ancestral sequence uncertainty, two variants of ancMT-Z (ancMT-Z1 and ancMT-Z2), five variants of ancMT-A (ancMT-A1, ancMT-A2, ancMT-A3, ancMT-A4 and ancMT-A5), five variants of ancMT-B (ancMT-B1, ancMT-B2, ancMT-B3, ancMT-B4 and ancMT-B5), five variants of ancMT-D (ancMT-D1, ancMT-D2, ancMT-D3, ancMT-D4 and ancMT-D5), 4 variants of ancMT-E (ancMT-E1, ancMT-E2, ancMT-E3 and ancMT-E4), and 4 variants of ancMT-F (ancMT-F1, ancMT-F2, ancMT-F3 and ancMT-F4) were generated through site-directed mutagenesis. Mutated sites were chosen based on comparisons between ancestral sequences obtained using different ancestral state estimation strategies as described above. In addition, mutated sites were chosen from among different structural components or domains of the protein, including the active site, surface, α-helix, β-sheet and loops. For each ancestral enzyme and its variants, the posterior probabilities (PP) of the original and mutated sites as well as their positions in the three dimensional structure of the protein (Zubieta et al. 2003) are shown in Table 1. An amino acid sequence alignment of the characterized modern-day SABATH members and the five resurrected ancestors are shown in Figure 5. Mutated sites in these variants are indicated by arrows, and active site residues, as identified through the crystal structure of C.breweri SAMT, are marked by inverted triangles.
Table 1. Posterior probabilities and positions of original and mutated sites for reconstructed ancestral enzymes and their variants

<table>
<thead>
<tr>
<th>Site number</th>
<th>Wild type</th>
<th>PP for Original site</th>
<th>Variant</th>
<th>PP for Mutated site</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>398</td>
<td>ancMT-Z</td>
<td>C (0.497)</td>
<td>ancMT-Z1</td>
<td>S (0.712)</td>
<td>α-helix</td>
</tr>
<tr>
<td>198</td>
<td>ancMT-Z</td>
<td>Y (0.981)</td>
<td>ancMT-Z2</td>
<td>F (0.755)</td>
<td>β-sheet, active site</td>
</tr>
<tr>
<td>45</td>
<td>ancMT-A</td>
<td>K (0.526)</td>
<td>ancMT-A1</td>
<td>N (0.987)</td>
<td>loop, surface</td>
</tr>
<tr>
<td>96</td>
<td>ancMT-A</td>
<td>L (0.998)</td>
<td>ancMT-A2</td>
<td>F (0.898)</td>
<td>loop, surface</td>
</tr>
<tr>
<td>212</td>
<td>ancMT-A</td>
<td>S (0.79)</td>
<td>ancMT-A3</td>
<td>D (0.951)</td>
<td>loop, surface</td>
</tr>
<tr>
<td>279</td>
<td>ancMT-A</td>
<td>S (0.592)</td>
<td>ancMT-A4</td>
<td>G (0.872)</td>
<td>loop, surface</td>
</tr>
<tr>
<td>437</td>
<td>ancMT-A</td>
<td>E (0.996)</td>
<td>ancMT-A5</td>
<td>L (0.902)</td>
<td>loop, surface</td>
</tr>
<tr>
<td>45</td>
<td>ancMT-B</td>
<td>N (0.676)</td>
<td>ancMT-B1</td>
<td>S (0.835)</td>
<td>loop, surface</td>
</tr>
<tr>
<td>117</td>
<td>ancMT-B</td>
<td>E (0.145)</td>
<td>ancMT-B2</td>
<td>deletion</td>
<td>loop, surface</td>
</tr>
<tr>
<td>148</td>
<td>ancMT-B</td>
<td>Q (0.498)</td>
<td>ancMT-B3</td>
<td>E (0.496)</td>
<td>loop, surface</td>
</tr>
<tr>
<td>345</td>
<td>ancMT-B</td>
<td>D (0.622)</td>
<td>ancMT-B4</td>
<td>N (0.460)</td>
<td>loop, surface</td>
</tr>
<tr>
<td>430</td>
<td>ancMT-B</td>
<td>E (0.665)</td>
<td>ancMT-B5</td>
<td>D (0.534)</td>
<td>α-helix, surface</td>
</tr>
<tr>
<td>38</td>
<td>ancMT-D</td>
<td>T (0.482)</td>
<td>ancMT-D1</td>
<td>I (0.981)</td>
<td>loop, surface</td>
</tr>
<tr>
<td>117</td>
<td>ancMT-D</td>
<td>E (0.139)</td>
<td>ancMT-D2</td>
<td>deletion</td>
<td>loop, surface</td>
</tr>
<tr>
<td>171</td>
<td>ancMT-D</td>
<td>N (1.0)</td>
<td>ancMT-D3</td>
<td>deletion</td>
<td>loop, surface</td>
</tr>
<tr>
<td>272</td>
<td>ancMT-D</td>
<td>F (0.682)</td>
<td>ancMT-D4</td>
<td>L (0.822)</td>
<td>β-sheet</td>
</tr>
<tr>
<td>396</td>
<td>ancMT-D</td>
<td>S (0.853)</td>
<td>ancMT-D5</td>
<td>A (0.952)</td>
<td>α-helix</td>
</tr>
<tr>
<td>107</td>
<td>ancMT-E</td>
<td>Y (0.454)</td>
<td>ancMT-E1</td>
<td>D (0.613)</td>
<td>α-helix, surface</td>
</tr>
<tr>
<td>260</td>
<td>ancMT-E</td>
<td>K (0.724)</td>
<td>ancMT-E2</td>
<td>E (0.915)</td>
<td>α-helix, surface</td>
</tr>
<tr>
<td>282</td>
<td>ancMT-E</td>
<td>F (0.495)</td>
<td>ancMT-E3</td>
<td>C (0.942)</td>
<td>loop, surface</td>
</tr>
<tr>
<td>432</td>
<td>ancMT-E</td>
<td>A (0.643)</td>
<td>ancMT-E4</td>
<td>S (0.984)</td>
<td>α-helix, surface</td>
</tr>
<tr>
<td>226</td>
<td>ancMT-F</td>
<td>K (1.0)</td>
<td>ancMT-F1</td>
<td>I (0.973)</td>
<td>loop, surface</td>
</tr>
<tr>
<td>295</td>
<td>ancMT-F</td>
<td>L (1.0)</td>
<td>ancMT-F2</td>
<td>I (0.912)</td>
<td>loop, active site</td>
</tr>
<tr>
<td>304</td>
<td>ancMT-F</td>
<td>N (1.0)</td>
<td>ancMT-F3</td>
<td>K (0.894)</td>
<td>α-helix, surface</td>
</tr>
<tr>
<td>403</td>
<td>ancMT-F</td>
<td>I (1.0)</td>
<td>ancMT-F4</td>
<td>T (0.749)</td>
<td>α-helix</td>
</tr>
<tr>
<td>Species</td>
<td>Accession</td>
<td>Length</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------</td>
<td>-----------</td>
<td>--------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clarkia SAMT</td>
<td></td>
<td>66</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Snapdragon SAMT</td>
<td></td>
<td>67</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stephanocis SAMT</td>
<td></td>
<td>65</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hoya SAMT</td>
<td></td>
<td>65</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Petunia BSMT</td>
<td></td>
<td>65</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Datura SAMT</td>
<td></td>
<td>66</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cestrum SAMT</td>
<td></td>
<td>66</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N. suav SAMT</td>
<td></td>
<td>65</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N. alata SAMT</td>
<td></td>
<td>65</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N. sylv BSMT</td>
<td></td>
<td>65</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arabidopsis BSMT</td>
<td></td>
<td>66</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arabidopsis IAMT</td>
<td></td>
<td>79</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Populus IAMT</td>
<td></td>
<td>77</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ocimun IAMT</td>
<td></td>
<td>64</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arabidopsis GAMT</td>
<td></td>
<td>80</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arabidopsis GAMT</td>
<td></td>
<td>80</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coffee IAMT</td>
<td></td>
<td>67</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coffee DXMT</td>
<td></td>
<td>67</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Theobroma MXMT</td>
<td></td>
<td>67</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Camellia MXMT</td>
<td></td>
<td>72</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ancMT-Z</td>
<td></td>
<td>67</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ancMT-A</td>
<td></td>
<td>66</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ancMT-B</td>
<td></td>
<td>66</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ancMT-D</td>
<td></td>
<td>65</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ancMT-E</td>
<td></td>
<td>65</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ancMT-F</td>
<td></td>
<td>65</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ancMT-D1 = T381  ancMT-A1 = K45N  ancMT-B1 = N45S
Figure 5. Alignment of amino acid sequences from functionally characterized SABATH family members including resurrected ancestral sequences. Mutated sites of these ancestral variants and sites for adaptive functional changes are indicated by arrows. Active site residues, deduced from the crystal structure of *C. breweri* SAMT, are marked by inverted triangles.
Functional divergence after the first gene duplication event at node A

Enzyme functional divergence following the first gene duplication event at node A (Figure 6) was explored using biochemical assays and statistical analyses. As mentioned above, this duplication event separated SAMTs, BSMTs and NAMTs in Apocynaceae and Solanaceae from the other characterized SABATH members. The resurrected ancMT-A gene was cloned and expressed, and the catalytic activity of total protein extract was tested with three substrates (BA, SA and NA) through in vitro radiochemical biochemical assays. The relative activity levels with these three substrates were then compared using the DPM values. Results indicate that the ancestor at node A prefers BA over SA and NA by more than 4-fold, and five ancMT-A variants produce similar activity patterns (Figure 6A). The catalytic activities of resurrected ancMT-Z and its two variants, which are even older ancestors, were also tested with the three substrates, and they also prefer BA over SA by more than 2-fold, but have no activity with NA (Figure 6A). These results suggest that ancestors at node A favored BA before gene duplication.

To detect functional changes from node A to node B, catalytic activities of resurrected ancMT-B and its five variants at node B were tested in the same way. Compared to ancestral enzymes at node A, ancestral enzymes at node B preferred SA over BA and NA by more than 2-fold (Figure 6B), suggesting that SA evolved to be the favored substrate in this daughter lineage after gene duplication.

In order to test whether the substrate preference change from BA to SA was subjected to positive selection or neutral drift, branch-site statistical analysis (Yang 2000) was used to detect positive selection between node A and node B. Of more than 50 site
differences between node A and node B, one change in the active site, from histidine (His or H) to methionine (Met or M) at position 201 (PP = 0.998), was predicted to be under positive selection ($d^v/d_s = 35$, $P < 0.05$) (Figure 6).

To test the functional consequences of this positively selected site, histidine was mutated to methionine at site 201 in both the ancMT-A and ancMT-Z2 backgrounds to recapitulate this putatively important change. Both mutants changed their preference from BA to SA, specifically, from preferring BA over SA by more than 4-fold to preferring SA over BA by more than 1.6-fold (Figure 6B). In addition, the evolutionary change was investigated in terms of enzyme kinetics. In a previous study (Huang et al. 2012), methionine at site 201 in \textit{H. carnosa} SAMT was mutated back to ancestral histidine, and the $k_{\text{cat}}$ and $K_M$ values of both BA and SA were calculated for the wild-type enzyme and the M201H mutant (Table 2). Taking the M201H mutant as the ancestral state at node A, replacement of histidine with methionine at position 201 increased the $k_{\text{cat}}$ for SA by more than 3-fold, while the $K_M$ for SA was almost unchanged; in this case, the $k_{\text{cat}}/K_M$ for SA increased more than 3-fold. Although this replacement decreased the $k_{\text{cat}}$ for BA by half, it also decreased the $K_M$ for BA by more than 3-fold. Thus the $k_{\text{cat}}/K_M$ for BA increased, but it was still less than the $k_{\text{cat}}/K_M$ for SA. In conclusion, these data suggest that the preference for SA by the ancestor at node B due to increased activity with SA rather than a simple decrease in its activity with BA.
Figure 6. Relative activities with BA (blue), SA (red), NA (green) of resurrected ancestral enzymes and their mutants at each node. The tree shown is simplified from that shown in Figure 4 with the same node labels. Lineages are colored according to their favored substrates. Mean values and standard deviations are calculated from at least two replicates. A. ancMT-A, ancMT-Z and their variants have highest activity with BA; B.
ancMT-Z H201M mutant, ancMT-A H201M mutant, ancMT-B and its variants have highest activity with SA; D. ancMT-D and its variants have highest activity with SA; E. ancMT-D M201H mutant, ancMT-E and its variants, have highest activity with BA; F. ancMT-F and its variants, ancMT-F Y402F mutant and N. gossei NAMT have highest activity with NA. There is no panel C because ancestral sequence for node C was not resurrected. The first preference change occurred between node A and node B where His was replaced by Met at position 201 by positive selection \((P < 0.05)\); the second preference change happened between node D and node E where His reappeared at position 201, an evolutionary reversal of the change from node A to node B, and this change was also concomitant with positive selection \((P < 0.05)\). The third evolutionary change was the improvement of NA activity from node F to N. gossei NAMT where Phe was replaced by Tyr at position 402; however, whether positive selection was concomitant with this change was not clear \((P > 0.05)\).

Table 2. Michaelis–Menten kinetic parameters for \(H.\text{carnosa}\) SAMT, ancMT-D and their mutants

<table>
<thead>
<tr>
<th></th>
<th>(H.\text{carnosa}) SAMT</th>
<th>(H.\text{carnosa}) SAMT, M201H</th>
<th>ancMT-D</th>
<th>ancMT-D, M201H</th>
</tr>
</thead>
<tbody>
<tr>
<td>(K_M) (BA), (\mu)M</td>
<td>156.5</td>
<td>548.7</td>
<td>295.3</td>
<td>78.1</td>
</tr>
<tr>
<td>(K_M) (SA), (\mu)M</td>
<td>61.5</td>
<td>79.84</td>
<td>15.77</td>
<td>6.9</td>
</tr>
<tr>
<td>(k_{cat}) (BA), (s^{-1})</td>
<td>(5.41 \times 10^{-3})</td>
<td>(9.77 \times 10^{-3})</td>
<td>(0.978 \times 10^{-3})</td>
<td>(4.94 \times 10^{-3})</td>
</tr>
<tr>
<td>(k_{cat}) (SA), (s^{-1})</td>
<td>(17.1 \times 10^{-3})</td>
<td>(4.75 \times 10^{-3})</td>
<td>(2.31 \times 10^{-3})</td>
<td>(1.14 \times 10^{-3})</td>
</tr>
<tr>
<td>(k_{cat}/K_M) (BA), (s^{-1} \times \mu)M(^{-1})</td>
<td>34.6</td>
<td>17.8</td>
<td>3.31</td>
<td>63.3</td>
</tr>
<tr>
<td>(k_{cat}/K_M) (SA), (s^{-1} \times \mu)M(^{-1})</td>
<td>278.1</td>
<td>59.48</td>
<td>146.5</td>
<td>165</td>
</tr>
</tbody>
</table>

Note. Kinetics data for \(H.\text{carnosa}\) SAMT and its M201H mutant are cited from Huang et al. (2012)

Functional divergence after the second gene duplication event at node D

Gene duplication at node D generated modern-day enzymes that favored BA over SA in the Solanaceae lineage (Hippauf et. al 2010). The origin of BSMT activity suggested that an evolutionary reversal event happened between node D and E (Figure 6). Resurrected ancestral enzymes ancMT-D and its variants were expressed and functionally characterized with BA, SA and NA as above. Results indicate that the ancestor at node D
preferred SA over BA and NA by more than 2.5-fold (Figure 6D). Assuming the ancestral condition at node D, BA evolved to be the favored substrate for the ancestor at node E after gene duplication, as indicated by the characterized functions of resurrected ancMT-E and its four variants which preferred BA over SA and NA by more than 10-fold (Figure 6E).

Branch-site statistical analysis (Yang 2000) was applied to detect whether positive selection was associated with the functional change from node D to node E. Of more than 50 evolutionary changes, the same active site position that changed between nodes A and B was found under positive selection ($d_{av}/d_s = 38, P < 0.05$) (position 201), which reverted back to histidine ($PP = 0.99$) to regain high BA methylating activity.

To verify the relevance of this positively selected site to the functional change, methionine was mutated to histidine at site 201 in the ancMT-D background. This M201H mutant was functionally characterized, and results showed a 3-fold higher activity with BA comparing to SA and NA (Figure 6E).

To clarify the substrate preference change between node D and node E in terms of enzyme kinetics, the $k_{cat}$ and $K_M$ values of BA and SA for ancMT-D and its M201H mutant were measured using purified enzymes (Figure 7). Compared to ancMT-D, the M201H mutation decreased the $K_M$ for BA by nearly 4-fold and increased the $k_{cat}$ for BA by 5-fold, whereas the $K_M$ and the $k_{cat}$ for SA both decreased by 2-fold (Table 2). It is clear that this positively selected change increased ancMT-D’s preference for BA by increasing the $k_{cat} / K_M$ by 20-fold, whereas the $k_{cat} / K_M$ for SA did not change substantially. It should be noted that for the M201H mutant, the $k_{cat} / K_M$ for SA is still higher than that for BA, and its catalytic abilities with BA, SA and NA are also different.
from those of ancMT-E and its variants (Figure 6F), suggesting that there must be other mutations contributing to the evolution of the M201H mutant to the enzyme condition at node E; however, the branch-site statistical analysis cannot detect such sites under positive selection.

Figure 7. Michaelis-Menten kinetic graphs of ancMT-D and its M201H mutant with BA and SA. Assays were done in duplicate. Initial velocities versus substrate concentrations were plotted to fit the hyperbolic Michalis-Menten equation in GraphPad Prism. The $V_{\text{max}}$, $K_M$ and their standard deviations were calculated based on the obtained hyperbolic curve. $V_{\text{max}}$ was then converted to apparent $k_{\text{cat}}$ in units of s$^{-1}$.

Functional divergence after the third gene duplication event at node E

The third gene duplication event is postulated to have happened within Nicotiana at node E (Hippauf et al. 2010), and after this gene duplication, NA evolved to be the
preferred substrate for one descendent lineage (Figure 4 and Figure 6). To determine when the NA activity became favored, the catalytic activities of ancMT-F and its four variants were assayed with BA, SA and NA. All of these ancestral enzyme variants at node F showed highest activity with NA and lowest activity with BA, which was the ancestrally favored substrate at node E (Figure 6F). Although the branch-site test was applied to detect sites that were responsible for the adaptive change from node E to node F, no positively selected sites were found.

Functional improvement from node F to *N. gossei* NAMT

Assuming the ancestral condition at node F, NA methylating activity seems to be further improved in *N. gossei* NAMT, which has negligible activity with BA and SA. Due to the low level of sequence divergence, the branch-site analysis was unable to detect any positively selected sites for the functional improvement from Node F to *N. gossei* NAMT. However, in another branch-site analysis using a denser sampling of SAMT and BSMT lineage members in Solanaceae (Figure 8), positive selection (*d_s/d_\text{S} = 65*) was detected, but is not statistically significant (*P* > 0.05). Two out of eight site changes were predicted to be adaptive. One of the two sites, position 402, from tyrosine (Tyr or Y) to phenylalanine (Phe or F) was mutated in the ancMT-F background, and functionally characterized with BA, SA and NA (Figure 6F). Results indicate that this mutation increased the relative activity with NA when compared to ancMT-F and its variants. However, further kinetic studies are needed to clarify whether this Y402F mutant increased its absolute activity with NA, decreased its absolute activity with BA and SA, or both situations happened.
Evolutionary reversal from node F to *N. suaveolens* BSMT1

Another descendent of node F, *Nicotiana suaveolens* BSMT1, prefers BA over SA and NA (Hippauf et al. 2010), an evolutionary reversal of substrate preference change from node E to node F. Sequence divergence from node F to *N. suaveolens* BSMT1 is quite low and they differ at only six amino acid sites: Phe was replaced by Leu at position 146; Tyr was replaced by Phe at position 198; His was replaced by Gln at position 230; Leu was replaced by Ile position 295; Val was replaced by Ala at position 319; and Cys was replaced by Ser at position 441 (Figure 5). In this case, the branch-site analysis lacks the power to detect any significant signature of positive selection from node F to *N. suav*BSMT1 (Nozawa et al. 2009). Based on the biochemical properties of amino acids, of these six residue replacements, the replacements from Tyr (polar) to Phe (nonpolar) at position 198 and from His (polar, basic) to Gln (polar, uncharged) at position 230 seem to be important. What’s more, the two positions are close to the active site of enzymes in the SABATH family (Figure 5). However, it is still possible that other combinations of these replacements achieved the functional reversal. Further mutagenesis studies are needed to verify the role of these replacements for the functional change.
Figure 8. Estimated relationships among SAMT and Solanaceae BSMT and NAMT. This gene tree includes a dense sampling of SAMT and BSMT members in Solanaceae, and is used in the branch-site analysis to detect positive selection from node F to *N. gossei* NAMT. Likelihood bootstrap values > 80 are shown for corresponding nodes. The BSMT1-2 members are not functionally characterized.

**DISCUSSION**

In my research, I used the SABATH multigene family to investigate how duplicated genes evolved different functions and how natural selection prompted such functional diversification. Specifically, by resurrecting extinct ancestral genes and testing their functions, I disclosed the rough evolutionary patterns of part of this gene family, the SAMT, BSMT and NAMT lineages in Apocynaceae and Solanaceae. I applied branch-site statistical analysis to detect positive Darwinian selection along each descendent
branch after gene duplication events. Sites under positive selection were mutated in the ancestral background to re-evolve the observed functional changes. Using the characterized ancestral gene functions, I tried to fit the evolutionary patterns into the aforementioned functional divergence models in terms of BA, SA and NA preferences. Almost all SABATH members and their ancestors are multifunctional enzymes specialized for one of these three substrates (Figure 6) (Huang et al. 2012). Of the four functional divergence models, only the DDC, EAC and IAD models require a multifunctional ancestor before gene duplication (Figure 1).

Divergence models for the first duplication and divergence event at node A

The ancestor at node A prefers BA over SA and NA before gene duplication; after gene duplication, the descendant at node B changed its relative preference from BA to SA. Positive selection prompt this change by the replacement of His to Met (Figure 6 A and C). As shown in Figure 1, the DDC model requires that daughter copies partition the ancestral functions through neutral drift without any functional improvements in both copies. In the EAC or IAD model, daughter copies not only partition the ancestral functions, but also improve both inherited functions (the EAC model) or one of the ancestral functions (the IAD model). Therefore, given that a significant signature of positive selection was detected in the lineage between node A and node B, the functional divergence after node A fits the EAC or IAD model rather than the DDC model. However, due to the uncertainty of the other descendant lineage of node A in the phylogenetic tree (Figure 3 and Figure 4), it is unclear whether positive selection, purifying negative selection or neutral drift acted on that lineage; thus, it is hard to distinguish EAC from IAD for the functional divergence at node A. Nevertheless,
assuming that efficient methylation of BA or SA is advantageous, given the observation that the ancestral enzyme at node A cannot be fully efficient on both BA and SA simultaneously, there appears to be an adaptive conflict between specializations on these two substrates. After gene duplication, the ancestral enzyme at node B evolved higher activity with SA and resolved this adaptive conflict. In this case, the EAC model probably explains the evolutionary change from node A to node B better than the IAD model.

Divergence models for the second duplication and divergence event at node D

Resurrected ancestor at node D prefers SA over BA and NA; after gene duplication, one daughter lineage, resurrected enzyme at node E, was driven by positive selection to become highly efficient at methylating BA; the other daughter lineage, including the modern day SAMTs in Solanaceae, inherited the ancestral function, preferring SA over BA and NA (Figure 4, Figure 6 D and E). Because significant positive selection prompting the replacement of Met to His was detected with improved ability of the ancestor at node E to methylate BA (Figure 6), the DDC model is not suitable to explain this type of change. To figure out whether this functional divergence fits the EAC or IAD model, it is important to know what type of evolutionary force acted on the lineage from node D to modern day SAMTs. There are two pieces of evidence that support the EAC model over the IAD model. On one hand, comparing ancMT-D to modern day SAMTs in terms of relative activity levels with SA, some of the descendent SAMTs seem to have evolved to be more preference in methylating SA (Huang et al. 2012), suggesting that positive selection might associate with the functional improvement, even though the branch-site test lacks the power to detect any significant positive
selection on this branch. On the other hand, if we assume efficient production of MeBA and MeSA is advantageous, the ancestor at node D was only able to methylate SA efficiently with 2.5-fold less activity with BA, in this case, adaptive conflict of specializing simultaneously with both BA and SA also existed within the ancestor at node D. A significant signature of positive selection has been detected to improve the ability to methylate BA along the lineage from node D to node E through the Met to His mutation at position 201, and the improvement of the ability to methylate BA at node E appears to have resolved the adaptive conflict (Figure 1).

Divergence models for the third duplication and divergence event at node E

The evolutionary pattern of relative activity with NA after gene duplication at node E fits the IAD model over the DDC and EAC models. Throughout the SAMT, BSMT and NAMT lineages, relative activity levels with NA seem to be neutral for ancestors from node A to node E, as these ancestral enzymes varied their substrate preferences between BA and SA, while activity with NA was only minimal (Figure 6). Subsequently, NA evolved to be the primary substrate of descendents at node F, and the relative activity level with NA got further improved in \textit{N. gossei} NAMT. According to the IAD model, the promiscuous activity with NA at node A could be considered as the innovation for the start of the IAD process. Gene duplications at node D and node E, as well as the polyploidization events in \textit{N. gossei} and \textit{N. suaveolens}, suggest the amplification of gene copies occurred. The divergence happened when the ancestor at node F evolved the highest activity with NA from the ancestral condition at node E which preferred BA, and \textit{N. gossei} NAMT further evolved to be specialized only with NA. The IAD model also requires that at least one descendent copy retained the ancestral function,
which was maintained by purifying selection. BSMT2, the other descendent lineage of node E, appears to have inherited the ancestral function of high relative activity with BA (Figure 3) (Huang et al. 2012). However, the branch-site analysis can’t provide any statistical support for positive selection on either descendent lineage of node E, although apparent substrate preference change was identified from node E to node F and N.gossei NAMT (Figure 6). In this case, the DDC model, which claims neutral partitioning of ancestral functions, could be an alternative explanation for this functional divergence.

Structural basis of adaptive changes

Effects of His or Met at site 201 on the substrate preference change between BA and SA were explored based on the three dimensional structure of ancMT-Z built by homology modeling. Although the chemical property of Met is nonpolar and hydrophobic, and the chemical property of His is polar and hydrophilic, the amino acid residue at site 201 seems only to interact with the benzene ring of the two substrates and not with the hydroxyl group of SA. Thus, the homology modeling did not reveal the structural basis of Met to His (or His to Met) mutation.

Impact of adaptive changes on the catalytic efficiency

By performing forward mutagenesis studies which introduced these historical mutations directly in the ancestral background, I successfully achieved the substrate preference changes between node A and node B, node D and node E as well as node F and N.gossei NAMT. I further investigated how such mutations affected substrate preference change in terms of catalytic efficiency, $K_M / k_{cat}$, using data measured from activities of ancMT-D and ancMT-D M201H with BA and SA. The positively selected
change M201H increased the $k_{\text{cat}} / K_M$ for BA by 20-fold, and this was achieved by increasing $k_{\text{cat}}$ and decreasing $K_M$ for BA (Table 2). However, due to the low $K_M$ for SA, the $k_{\text{cat}} / K_M$ for BA was still less than that for SA, even though the $k_{\text{cat}}$ for BA evolved to be four times larger than the $k_{\text{cat}}$ for SA in the M201H mutant. In this case, the enzyme would interact with SA first if the two substrates were present in equal molar concentration within cell. The kinetic analysis suggests that there must be other mutations that might further increase the $K_M$ or decrease the $k_{\text{cat}}$ for SA, increase the $k_{\text{cat}}$ or decrease the $K_M$ for BA, or a combination of these factors to evolve the relative enzyme activity at node E (Figure 6).

Evolutionary reversal events in the SABATH family

Studies on the evolution of vertebrate glucocorticoid and mineralocorticoid receptors (Ortlund et. al 1997) and evolution of red fluorescence proteins (Field & Matz, 2010) demonstrated that there are usually strong epistatic interactions between permissive and adaptive changes during the evolution of protein functions, and these interactions may constrain evolutionary reversals; however, for the evolutionary reversal event identified from node D to node E in the Solanaceae, one ancestral active site change from Met back to His was sufficient for the functional reversal (Figure 6E), and there were no accumulated mutations to constrain this reversal during about 50 to 75 million years prior to the origin of Solanaceae (Hippauf et. al 2010). This phenomenon may suggest that epistasis between permissive and adaptive changes might be weak for the evolution of enzymes involved in secondary metabolism.

In conclusion, substrate promiscuity in the ancestors of the SABATH gene family had an important role in the functional diversification of modern day members. This type
of ancestral functional variation for protein evolution has been applied in several protein engineering studies to evolve the novel function from promiscuous ancestral functions (Aharoni et al. 2005; Arnold 2009). Ancestral functional promiscuity was found in both plants and animals. For example, the ligand-binding promiscuity of the ancestral vertebrate steroid receptors provides the structural basis to evolve novel receptor-ligand interactions in modern day receptors (Bridgham et al. 2009). Reconstruction and chemical synthesis of ancestral proteins offer an intriguing opportunity to study the protein functional evolution as implicated in my study here. With the continuing development of statistical methods and models of molecular evolution for inferring ancestral sequences, we would gain more accurate understanding of the properties of ancient proteins and the evolution of their physiological roles in cells.
REFERENCES


Zhao, N., J. L. Ferrer, J. Ross, J. Guan, Y. Yang, E. Pichersky, J. P. Noel and F. Chen (2008). Structural, biochemical, and phylogenetic analyses suggest that indole-3-acetic acid methyltransferase is an evolutionarily ancient member of the SABATH family. *Plant Physiology* 146(2): 455-467.

Appendix A

Project approval certification
Project Title: Ghosts of Evolution Past: Resurrecting Ancestral Enzymes to Understand the Evolution of Modern-Day Enzyme Activities

Principal Investigator: Todd Barkman

RDPC Project Number: 12-TBa

Date Received by the rDNA Biosafety Committee: October 31, 2011

☐ Reviewed by the rDNA Biosafety Committee

☒ Approved

☐ Approval not required

Vice Chair of rDNA Biosafety Committee Signature: 12/01/2011

Date: 12/01/2011

Revised 11/2011 WMU RDPC
Appendix B

Approval of radioactive material usage
To Whom It May Concern,

This letter is provided to verify that Ruiqi Huang has completed the Western Michigan University's Radiation Safety Training course. Ruiqi has shown the required knowledge to maintain compliance with Western Michigan University's Radiation Safety Program, state regulations, and Federal regulations. He has demonstrated the ability to work safely around radioactive materials and/or radiation producing machines as it pertains to the ALARA (as low as reasonably achievable) radiation exposure practices and follow general laboratory safety practices.

Ruiqi had his initial training in July, 2011 with John G. Center. He scored 100 percent on his written exam. His next annual refresher training will be held in September, 2012.

With respect,

John G. Center, Jr.
Research Safety Officer
3922 Wood Hall
Western Michigan University
1903 West Michigan Ave
Kalamazoo, Mi 49008-5410
(269) 387-5933