PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

The evolution of function in strictosidine synthase-like proteins

Michael A. Hicks,[1] Alan E. Barber II,[1] Lesley-Ann Giddings,[2] Jenna Caldwell,[2] Sarah E. O'Connor,[3,4] and Patricia C. Babbitt[1,5,6]*

[1] Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, California 94158
[2] Department of Chemistry, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139
[3] Department of Biological Chemistry, The John Innes Centre, Colney Lane, Norwich NR4 7UH, UK
[4] School of Chemistry, The University of East Anglia, Norwich NR4 7TJ, UK
[5] Department of Pharmaceutical Chemistry, University of California, San Francisco, California 94158
[6] California Institute for Quantitative Biosciences, University of California, San Francisco, California 94158

ABSTRACT

The exponential growth of sequence data provides abundant information for the discovery of new enzyme reactions. Correctly annotating the functions of highly diverse proteins can be difficult, however, hindering use of this information. Global analysis of large superfamilies of related proteins is a powerful strategy for understanding the evolution of reactions by identifying catalytic commonalities and differences in reaction and substrate specificity, even when only a few members have been biochemically or structurally characterized. A comparison of >2500 sequences sharing the six-bladed β-propeller fold establishes sequence, structural, and functional links among the three subgroups of the functionally diverse N6P superfamily: the arylesterase-like and senescence marker protein-30/gluconolactonase/luciferin-regenerating enzyme-like (SGL) subgroups, representing enzymes that catalyze lactonase and related hydrolytic reactions, and the so-called strictosidine synthase-like (SSL) subgroup. Metal-coordinating residues were identified as broadly conserved in the active sites of all three subgroups except for a few proteins from the SSL subgroup, which have been experimentally determined to catalyze the quite different strictosidine synthase (SS) reaction, a metal-independent condensation reaction. Despite these differences, comparison of conserved catalytic features of the arylesterase-like and SGL enzymes with the SSs identified similar structural and mechanistic attributes between the hydrolytic reactions catalyzed by the former and the condensation reaction catalyzed by SS. The results also suggest that despite their annotations, the great majority of these >500 SSL sequences do not catalyze the SS reaction; rather, they likely catalyze hydrolytic reactions typical of the other two subgroups instead. This prediction was confirmed experimentally for one of these proteins.

Proteins 2011; 00:000–000.

Key words: sequence similarity networks; protein function misannotation; functionally diverse superfamily; gene context; reaction mechanism.

INTRODUCTION

The number of protein sequences available in public databases such as UniProtKB/TrEMBL is now well over 16 million and continues to rise at exponential rates[1] while our ability to functionally characterize these proteins is limited to just a small fraction of these. Computational approaches that can improve our ability to predict and study their molecular functions are therefore critical for leveraging useful information from genome sequencing projects. Yet, the application of these methods to many superfamilies (SFs) can be confounded by complicated patterns of variation across diverse but related protein sequences. Thus, even as an initial step for understanding their functions, annotation transfer from characterized to uncharacterized proteins based on simple similarity metrics can be insufficient to achieve high confidence prediction of their reaction specificities.[2]

One approach to address the widening gap between experimentally characterized and uncharacterized proteins is a global comparison of sequence, structural, and functional features of evolutionarily related proteins to identify common catalytic properties as well as features that distinguish them. Such studies of a number of function-

Abbreviations: ABC, adenosine triphosphate-binding cassette; APMAP, adipocyte plasma membrane-associated protein; DFP, diisopropyl fluorophosphate; DFPase, diisopropyl fluorophosphatase; Drp35, drug-responsive protein-35; LRE, luciferin-regenerating enzyme; N6P, nucleophilic attack 6-bladed β-propeller; NR, Genbank's nonredundant protein database; pNPAc, *p*-nitrophenyl acetate; PON, paraoxonase; RMSD, root mean square deviation; SF, superfamily; SGL, senescence marker protein-30/gluconolactonase/LRE-like; SMP-30, senescence marker protein-30; SS, strictosidine synthase; SSL, strictosidine synthase-like.

Additional Supporting Information may be found in the online version of this article.

*Correspondence to: Patricia C. Babbitt; Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, CA 94158. E-mail: babbitt@cgl.ucsf.edu

Received 11 May 2011; Revised 22 June 2011; Accepted 7 July 2011.
Published online 22 July 2011 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.23135

Table I
Experimentally Characterized Activities of Selected N6P SF Enzymes

Protein accession ID	Subgroup	Enzymatic activity			
		Lactonase	Esterase	Organophosphatase	SS
Rauvolfia serpentina strictosidine synthase (pdb_id: 2fpb, 2fpc, 2fp8, 2pf9, 2vaq, 2v91)	SSL	?[a]	?	?	Yes[b][8]
Catharanthus roseus strictosidine synthase (GI:18222)	SSL	?	?	?	Yes[9]
Ophiorrhiza pumila strictosidine synthase (GI:13928598)	SSL	?	?	?	Yes[10]
Rauvolfia mannii strictosidine synthase (GI: 21097)	SSL	?	?	?	Yes (Predicted)[c][11]
Rauvolfia verticillata strictosidine synthase (GI: 118076220)	SSL	?	?	?	Yes (Predicted)[c][12]
Ophiorrhiza japonica strictosidine synthase (GI: 193792547)	SSL	?	?	?	Yes (Predicted)[c][13]
Homo sapiens adipocyte plasma membrane-associated protein (APMAP) (GI: 24308201)	SSL	?	Yes[32]	?	?
Homo sapiens paraoxonase 1 (PON1) (GI: 130675)	Arylesterase-like	Yes[19]	Yes[19]	Yes[19]	?
Homo sapiens paraoxonase 2 (PON2) (GI: 6174935)	Arylesterase-like	Yes[18]	Yes[18]	No[d][18]	?
Homo sapiens paraoxonase 3 (PON3) (GI: 29788996)	Arylesterase-like	Yes[18]	Yes[18]	Yes[18]	?
PON1 G2E6 mutant (pdb_id: 1v04)	Arylesterase-like	Yes[19]	Yes[19]	Yes[19]	?
Loligo vulgaris diisopropyl fluorophosphatase (DFPase) (pdb_id: 1e1a, 1pjx, 2gvv, 2gvw, 3byc)	SGL	No[27]	?	Yes[67]	?
Fusarium oxysporum lactonohydrolase (GI: 6448475)	SGL	Yes[68]	?	No[7]	?
Rattus norvegicus senescence marker protein-30 (regucalcin) (GI: 68067383)	SGL	Yes[23]	Yes[24]	Yes[24]	?
Staphylococcus aureus drug responsive protein 35 (Drp35) (pdb_id: 2dg0, 2dg1, 2dso)	SGL	Yes[25]	?	?	?
Zymomonas mobilis gluconolactonase (GI: 48655)	SGL	Yes[69]	?	?	?

[a]No available literature reference showed that this activity was detected: "?"
[b]Detected activity: "Yes".
[c]Based on similarities to experimentally characterized SSs (see text): "Yes (Predicted)".
[d]No detectable activity: "No".

ally diverse enzyme SFs have contributed to general hypotheses regarding nature's evolutionary strategies in the diversification of protein function (see Refs. 3–6 for some reviews) and provided useful roadmaps to interpreting their structure–function relationships.

This paper describes a functionally diverse enzyme SF distinguished both by its large size (>2500 sequences) and lack of experimental characterization for all but a tiny handful of its members. Previously described[7] as comprising enzymes that catalyze several different chemical reactions, including strictosidine synthase (SS), paraoxonase (PON), and lactonohydrolase (Table I), all of these enzymes belong to the six-bladed β-propeller fold class and share a common catalytic feature involving nucleophilic attack on an electrophilic substrate. Termed here the nucleophilic attack six-bladed β-propeller (N6P) SF, its members can be clustered into three subgroups based on their sequence similarities. We define a subgroup as sharing similarity in sequence and function within a SF; each subgroup in turn may be comprised of multiple families, each of which catalyzes a distinct overall reaction.

The strictosidine synthase-like (SSL) subgroup, the primary focus of this paper, currently includes three experimentally characterized SSs from *Rauvolfia serpentina*,[8] *Catharanthus roseus*,[9] and *Ophiorrhiza pumila*[10] along with a much larger group of ~500 SSL sequences for which the reaction specificity has not been experimentally identified but that have been annotated in public databases as "putative strictosidine synthases" or "strictosidine synthase family proteins," based on their similarities to the experimentally characterized SSs. In addition to these, proteins from *R. mannii*[11] and *R. verticillata*,[12] which share 100% identity with the SS from *R. serpentina*, and a sixth protein from *O. japonica*,[13] which, based on the metabolic profiles of these species and presence of residues important to their SS activity, are predicted to catalyze the SS reaction as well. SS enzymes are produced by higher plants and catalyze the metal-independent condensation of tryptamine and secologanin via a Pictet–Spengler reaction (Fig. 1a) to generate strictosidine, a precursor molecule of the monoterpenoid indole alkaloid biosynthesis pathway. Starting with strictosidine, many pathways in different plants then lead to the production of about 2000 alkaloid compounds, several of which are useful in the treatment of a variety of human diseases, including cancer, malaria, and schizophrenia.[14] Though in nature SS shows exquisite specificity for its substrates, rationally designed mutations in the enzyme's active site permit modifications of substrate specificity and the formation of altered products.[15] These "unnatural" products have been shown to be incorporated into the pathway, leading to potentially medically useful compounds.[16] Many more SSs will likely be discovered as new plant genomes are solved, opening the possibility of additional "naturally" decorated variants that could enlarge the number of available

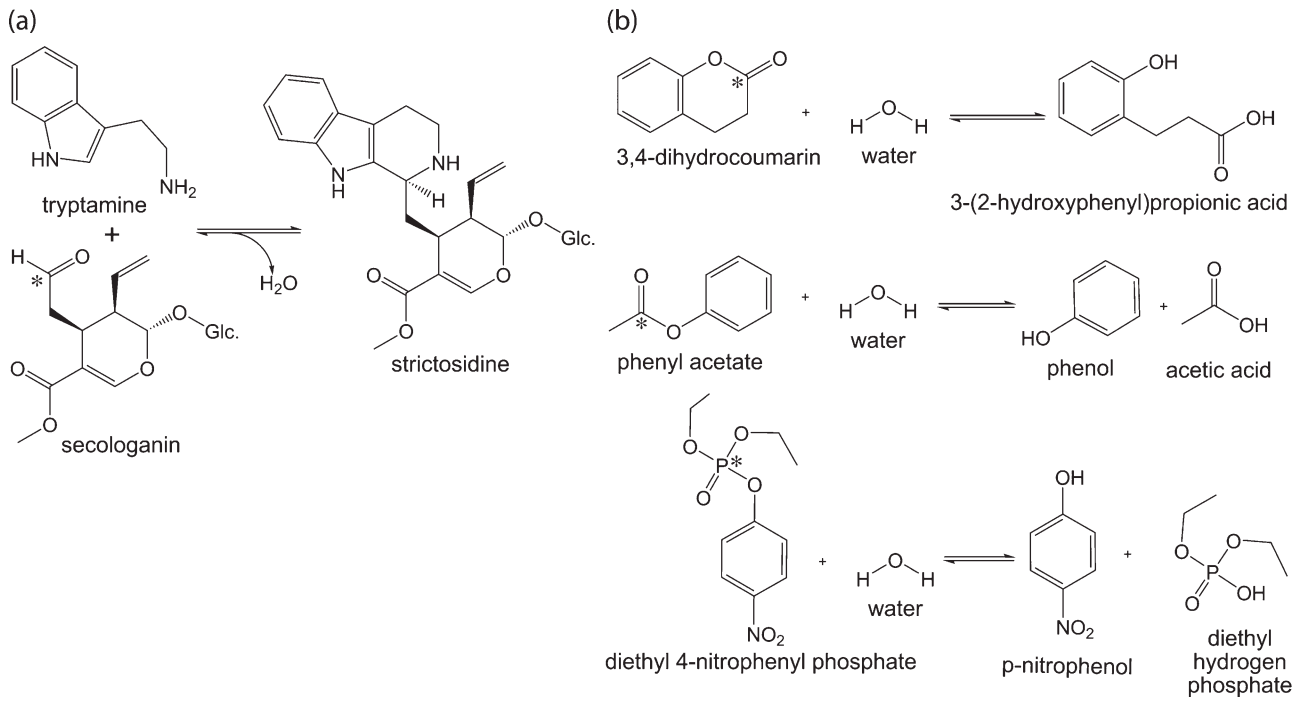

Figure 1

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

precursors for new drugs in the strictosidine class. Sequences that have been experimentally characterized as catalyzing the SS reaction are termed "true" SSs below to distinguish them from the uncharacterized SSL proteins, many of which we suggest catalyze some other reaction instead.

The other two subgroups in the N6P SF are termed herein the arylesterase-like and senescence marker protein-30/gluconolactonase/luciferin-regenerating enzyme (SGL) subgroups, based on the Pfam families of the same name.[17] The arylesterase-like subgroup comprises ~200 proteins. Its characterized members include the human serum paraoxonases (the PON1, PON2, and PON3 families), which have been shown to have lactonase activity.[18] PON1 proteins additionally catalyze ester bond cleavage and organophosphate degradation reactions,[19,20] whereas PON2 and PON3 proteins have very limited arylesterase activity and virtually no organophosphatase activity[21] (Table I, Fig. 1b). *In vitro* studies have also shown that many of these proteins catalyze one or more of these reactions "promiscuously;" that is, at a significant rate enhancement but not at the level expected for native-like activity.[18,20,22] Like the arylesterase-like subgroup, the SGL subgroup, containing about 1800 members, catalyzes a similar set of chemical reactions. Examples include senescence marker protein-30, an enzyme involved in L-ascorbic acid biosynthesis in non-primate mammals[23] that can also breakdown toxic organophosphates in mouse liver,[24] drug-responsive protein-35 (Drp35), involved in the resistance to antibiotics by *Staphylococcus aureus*,[25,26] diisopropylfluorophosphatase (DFPase), which degrades organophosphates[27] but for which the native activity remains unknown, and luciferin-regenerating enzyme, a protein that catalyzes luciferin regeneration in fireflies.[28] Characterized proteins in the arylesterase-like and SGL subgroups are metal dependent and the great majority of sequences in these subgroups share a conserved set of active site residues that are involved in the coordination of a divalent metal ion. These have been implicated by mechanistic studies as being important for catalysis of these lactonase, esterase, and organophosphatase activities.[25,29–31] Our sequence comparisons show that a similar pattern of metal coordinating ligands is conserved in the great majority of SSL sequences as well, including the SSL subgroup member human adipocyte plasma membrane-associated protein (APMAP), which was recently determined to have some arylesterase activity.[32] The vast majority of proteins in both the arylesterase-like and the SGL subgroups have not been biochemically characterized, so that, like the SSL subgroup, their specific functions are not known.

In this paper, we describe a global computational comparison of the >2500 members of the N6P SF to identify their sequence, structural, and mechanistic links, which

are then used to develop hypotheses about the functions of the many sequences of unknown function ("unknowns") represented in the SSL subgroup. This global comparison provides a context for discriminating functional features and lays a foundation for prediction of reaction specificity of the unknowns in the N6P superfamily.

MATERIALS AND METHODS

Data set sources and curation

Full-length protein sequences gathered from Pfam[17] seed set families with near structural homology to any protein sharing the six-bladed β-propeller fold (strictosidine synthase: PF03088; Arylesterase: PF01731; SGL: PF08450; Folate_rec: PF03024; GSDH: PF07795; Ldl_recept_b: PF00058; MRJP: PF03022; NHL: PF01436; PD40: PF07676; Phytase: PF02333) were used as an initial starting point in identifying new homologous sequences. These data were combined into a "seed set" which was used as a query set for a series of BLAST[33] searches against UniRef100.[34] All hits from the initial BLAST search with *E*-values less than 1E-5 were kept and added. The set was filtered using HMMER 3.0 beta and its three defined filtering criteria.[35] Any sequence that did not meet the three filtering criteria for any of the Pfam HMMs (i.e., did not receive a HMMER score against any Pfam model) were dropped. This set was again used as a query set for BLAST using the same procedure from the first iteration. All sequences from the final set were cropped to the sequence corresponding to the apparent six-bladed β-propeller domain. A curated sequence subset was aligned using PROMALS3D.[36] The clusters corresponding closely to the models for strictosidine synthase, SGL and Arylesterase contained conserved metal coordinating residues and were added to a "final SF set."

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

Generating sequence similarity networks

Sequence similarity networks of the SF and the SSL subgroup were generated using an altered version of a previously described methodology[37] and visualized using the Cytoscape program.[38] Networks were generated in which nodes represent sequences and edges represent BLAST-based connections against the UniRef100 database.[34] An edge is drawn between two sequences only if the statistical significance of the similarity score between them is less than (better than) a defined *E*-value cutoff. The organic layout was used to generate the final graph.

Structure-guided sequence alignments

The best aligning chains of serum paraoxonase 1 from the arylesterase-like subgroup (pdb_id: 1v04.pdb[39]), *Loligo vulgaris* ganglion diisopropylfluorophosphatase (pdb_id: 1pjx.pdb[40]) and *Staphylococcus aureus* drug responsive protein-35 (pdb_id: 2dg1.pdb[25]) from the SGL subgroup, and *R. serpentina* strictosidine synthase (pdb_id: 2fpb.pdb[41]) from the SGL subgroup were aligned using the Needleman–Wunsch algorithm[42] as implemented in the Matchmaker program[43] in Chimera.[44] A companion program, Match -> Align, was used to generate a multiple sequence alignment based on the structure alignment. The sequence alignment was then refined by eye using the aligned structures as a guide.

In the case of 2gvv.pdb (DFPase with inhibitor bound), 2fpb.pdb (strictosidine synthase with tryptamine bound) and 2fpc.pdb (strictosidine synthase with secologanin bound), a structure-based sequence alignment was generated using the Matchmaker program[43] in Chimera,[44] as described earlier, without refinement by eye. The structure alignment was further refined by aligning the alpha carbons of the last three metal-coordinating residue positions. Distances for reactive group positions were then measured in Chimera.[44]

Sequence-based alignments of the SSL subgroup were generated for each phylogenetically defined cluster using MUSCLE.[45] For example, proteins in the plant only cluster were aligned to each other prior to generating a full subgroup alignment. Profile alignments, in which each cluster of aligned sequences was aligned with another cluster, were then created. Protein sequences from the arylesterase-like and SGL subgroups with associated structures were then aligned to the SSL subgroup alignment using MUSCLE. This overall alignment was refined by eye using the structure-based multiple sequence alignment as a guide.

Gene context analysis

The amino acid sequence of a SSL gene fused to the transmembrane portion of an ABC transporter (gi|13471676) was used to identify other putative ABC transporter fusions by BLAST searches using the integrated microbial genomics system.[46] The top eight nonredundant hits were selected based on their alignment length and gene neighborhoods evaluated.

Phylogenetic tree

Proteins in the SSL subgroup alignment were filtered to 40% identity using cd-hit,[47] resulting in about 30 clusters. A single protein was selected from each cluster

based on the median length of that cluster. Trees were constructed with MrBayes v3.1.2[48,49] under the WAG amino acid substitution model[50] using a gamma distribution to approximate rate variation among sites.

Structure–function linkage database

Data for the SSL subgroup has been added to the Structure-Function Linkage Database (SFLD) (http://sfld.rbvi.ucsf.edu)[51] as described in the text.

Data for the SGL and arylesterase-like subgroups of this SF will be added to the SFLD.

General methods and analytical techniques

Secologanin was isolated as previously described.[52] All chemicals were purchased from Sigma Aldrich unless otherwise noted.

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

Cloning and protein expression

C. roseus SS in pET28a (+) and empty vector were transformed into *Escherichia coli* BL21 (DE3) cells for protein expression. The *Vitis vinifera* SSL gene (gi|147772032) was synthesized by CODA genomics (now Verdezyne; Carlsbad, CA) and *Nco*I and *Xho*I restriction sites were introduced by PCR for standard directional cloning into pET32b (+). The construct was then transformed into *E. coli* Rosetta DE3 cells for protein expression. PON1 (variant G2E6) in pET32b (+) was transformed into *E. coli* Origami B DE3 cells for protein expression. All liquid and solid media were supplemented with 1 m*M* $CaCl_2$.

Overnight cultures were grown at 37°C in sterile LB-broth containing 1 m*M* $CaCl_2$ and the appropriate antibiotic selection. Cultures of *C. roseus* SS (500 mL), containing 1 m*M* $CaCl_2$ and 50 μg mL^{-1} of kanamycin were inoculated with overnight cultures (1:100 dilution), and grown at 37°C until the optical density at 600 nm reached 0.5–0.75. After cultures were chilled at 4°C for 30 min, protein expression was induced with 1 m*M* isopropyl-β-D-1-thiogalactopyranoside (IPTG). Cells were harvested by centrifugation after 18 h of protein expression at 18°C and stored at –80°C. Cultures of pET28a(+) (500 mL) containing 1 m*M* $CaCl_2$ and 50 μg of kanamycin were inoculated with overnight cultures (1:100 dilution), and grown at 30°C until the optical density at 600 nm of 0.5–0.75 was reached. Protein expression was induced with 1 m*M* IPTG and after 4 h the cells were harvested by centrifugation and stored at –80°C.

Cultures of *V. vinifera* SSL and the empty pET32b (+) vector (500 mL) containing 1 m*M* $CaCl_2$ and 100 μg mL^{-1} of ampicillin and 34 μg mL^{-1} of chloroamphenicol were inoculated with overnight culture (1:100 dilution), and grown at 37°C until the optical density at 600 nm of 0.5–0.75 was reached. Protein expression was induced with 1 m*M* IPTG and after 6 h the cells (2 h for empty vector) were harvested and stored at –80°C. PON1 was expressed as previously reported.[53]

Protein purification

PON1 was purified as described previously.[53] Cells expressing *C. roseus* SS, pET28a (+), pET32b (+), and the *V. vinifera* SSL were lysed by sonication in lysis buffer (pH 8) containing 50 m*M* HEPES, 1 m*M* $CaCl_2$, 300 m*M* NaCl, 10 m*M* imidazole, 10% glycerol, 4 mg lysozyme, and 0.4 μg leupeptin and pepstatin protease inhibitors. The lysate was then incubated in 0.1% tergitol for 2.5 h at 4°C. After centrifugation, the supernatant was incubated with 0.01% v/v pre-equilibrated Ni-NTA resin suspension for 1 h before the flow-through was collected. The resin was washed with one column volume of lysis buffer and two column volumes of wash buffer containing 50 m*M* HEPES, 1 m*M* $CaCl_2$, 300 m*M* NaCl, 20 m*M* imidazole, 10% glycerol, and 0.1% tergitol. The resin was then washed with increasing concentrations of imidazole and the histidine-tagged proteins were eluted in pH 8 buffer containing 50 m*M* HEPES, 1 m*M* $CaCl_2$, 300 m*M* NaCl, 250 m*M* imidazole, and 10 % glycerol. The eluent was concentrated in Amicon Ultra centrifugal filter units by Millipore (Billerica, MA) and buffer exchanged using 50 m*M* HEPES buffer containing 162 m*M* NaCl, 1 m*M* $CaCl_2$, and 10 % glycerol. The final protein concentration was determined using a bichinchoninic acid assay by Pierce (Rockford, IL).

Hydrolase activity with *p*-nitrophenyl acetate

A stock of 300 mM *p*-nitrophenyl acetate (pNPAc) was prepared in HPLC-grade methanol and diluted for enzyme assays. Colorimetric assays to detect the formation of *p*-nitrophenol at 405 nm were prepared in a MICROTEST 96 well plate from Becton Dickinson Labware (Franklin Lakes, NJ) with a final volume of 250 μL containing either 530 nM (*C. roseus* SS, pET28a (+)) or 53 nM (PON1, *V. vinifera* SSL, and pET32b (+)) of

protein in 100 m*M* HEPES buffer containing 1 m*M* $CaCl_2$ and 2.6 m*M* NaCl. After pre-equilibration at room temperature for 10 min, assays were initiated by the addition of pNPAc (0.5, 1, 1.6, 2.1, 2.6, and 3.2 m*M*). Time points were chosen such that the rate of product formation was linear, to ensure accurate measure of initial rates (between 3 and 30 min after initiation of the reaction). The measured pathlength of 0.79 cm and *p*-nitrophenol extinction coefficient of 18,000 M^{-1} cm^{-1} were used to convert the absorbance units of *p*-nitrophenol into concentrations by the Beer–Lambert law. Kinetic parameters were estimated from two kinetic trials by fitting the data to the Lineweaver–Burk plot since limited substrate solubility did not enable an acceptable fit to the Michaelis–Menten equation (maximal substrate concentration was less than the 2–3 $\times$ K_M).

Pictet–Spenglerase activity

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

RESULTS AND DISCUSSION

In the following sections, sequence similarity networks[37] generated from all-by-all pairwise comparisons of >2500 sequences are used to summarize relationships across the subgroups of the N6P SF and provide functional context for more detailed comparisons of their active sites and mechanisms. Functional and biological features mapped to the networks then enable visualization of functional trends across the SF.

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

Similarities and differences among SSL, SGL, and arylesterase-like subgroup proteins are complicated and suggest that most SSL proteins do not catalyze the SS reaction

The sequence similarity network comparing the proteins in the N6P SF shows that SSL proteins share a higher degree of similarity with SGL members than either subgroup does with any member of the arylesterase-like subgroup (Fig. 2), even though comparison of their enzymatic functions suggests a closer relationship between the arylesterase-like and SGL subgroups than for either with the true SSs (Fig. 1). The best connection seen between an arylesterase-like protein and any other subgroup is at an E-value of 1.2 $\times$ 10^{-8}, where a single edge appears between the arylesterase-like and SGL subgroups (data not shown). At an *E*-value threshold of 1 $\times$ 10^{-5}, multiple connections are seen between the arylesterase-like and SSL subgroups. No connections are seen between the SSL and arylesterase-like subgroups at the *E*-value threshold of 1 $\times$ 10^{-5}, which is the least significant E-value at which we feel connections can be considered as minimally confident.[37] Consistent with the functional evidence, comparison of available structures for all three subgroups also shows that the active sites of these arylesterase-like and SGL proteins are similar to each other while the active site of the true SS structures is indeed highly divergent from the other two subgroups (Fig. 3).

Using Drp35 (pdb_id: 2dg1; SGL subgroup) as a reference structure, PON1 (pdb_id: 1v04; arylesterase-like subgroup) aligns at 86 alpha carbon positions with an overall RMSD of 1.18 Å, indicating that the overall structures of these two representative N6P SF members are highly similar. The most striking feature in their

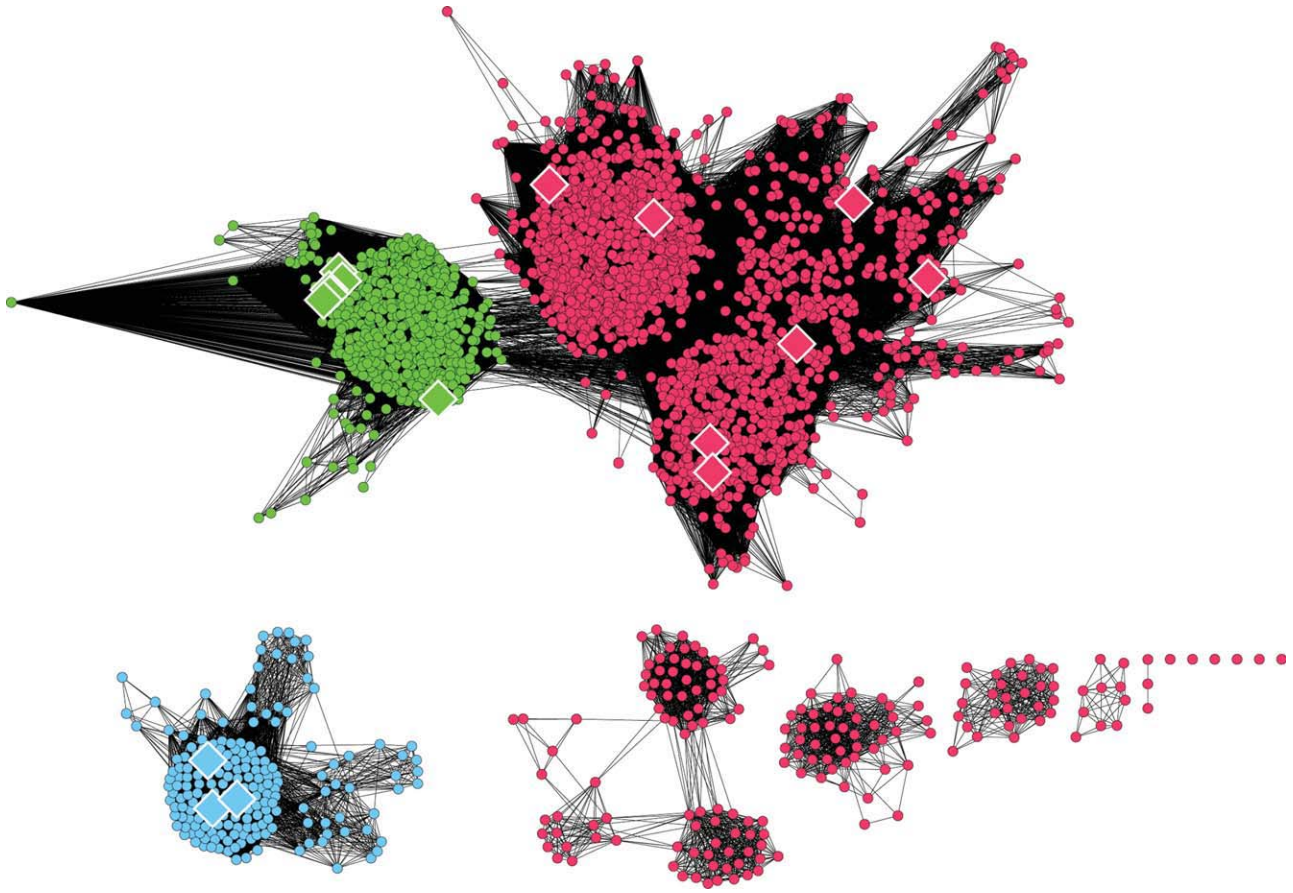

Figure 2

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

superposed active sites is a conserved set of four residues coordinating to a divalent metal ion (Fig. 3a); this metal-dependent active site architecture has been implicated in the phosphotriesterase mechanism of DFPase (SGL subgroup) and perhaps also in the phosphotriesterase mechanism of PON1 (arylesterase-like subgroup).[30] [1] In DFPase, the oxygen from an aspartate from blade 5 is thought to perform a direct nucleophilic attack on a phosphorous atom in a phosphoryl group of DFP, as demonstrated by an $H_2{}^{18}O$ incorporation experiment in which the metal coordinating oxygen of the aspartic acid is replaced by an oxygen atom of a solvent water molecule.[30] The other three residues involved in metal coordination appear to be necessary to maintain the electrostatic environment required for proper orientation of the substrate. The identity of a physiologically relevant divalent metal ion for many of these proteins is unclear; *in vitro* hydrolytic activity has been found using a variety of metals for proteins in the SGL subgroup, including Ca^{2+}, Mg^{2+}, Zn^{2+}, and Mn^{2+} in human SMP-30;[54] Mn^{2+} and Zn^{2+} for the lactonase activity and Mg^{2+}, Mn^{2+}, Co^{2+}, and Cd^{2+} for the DFPase activity in mouse SMP-30;[23] and Ca^{2+} in Drp35.[25] Regardless of their exact identity, these metals appear necessary to polarize the sp^2-hybrid bonds of substrates (such as carbonyl or phosphoryl groups), and to stabilize the resulting negative charge from transition states and other intermediates. In addition to the catalytic Ca^{2+} ion required for lactonase and esterase reactions in PON1, a histidine–histidine dyad has been implicated in activation of a water molecule involved in nucleophilic attack on the carbonyl of the substrate (Supporting Information Fig. S1).[39,55]

[1]Although additional residues may coordinate to the metal in some enzymes (such as a fifth residue seen in the PON1 structure), a mechanistic role has not yet been defined beyond the four residues detailed in our analysis. As such, our analysis is limited to speculating on the role of these four residues in the SF.

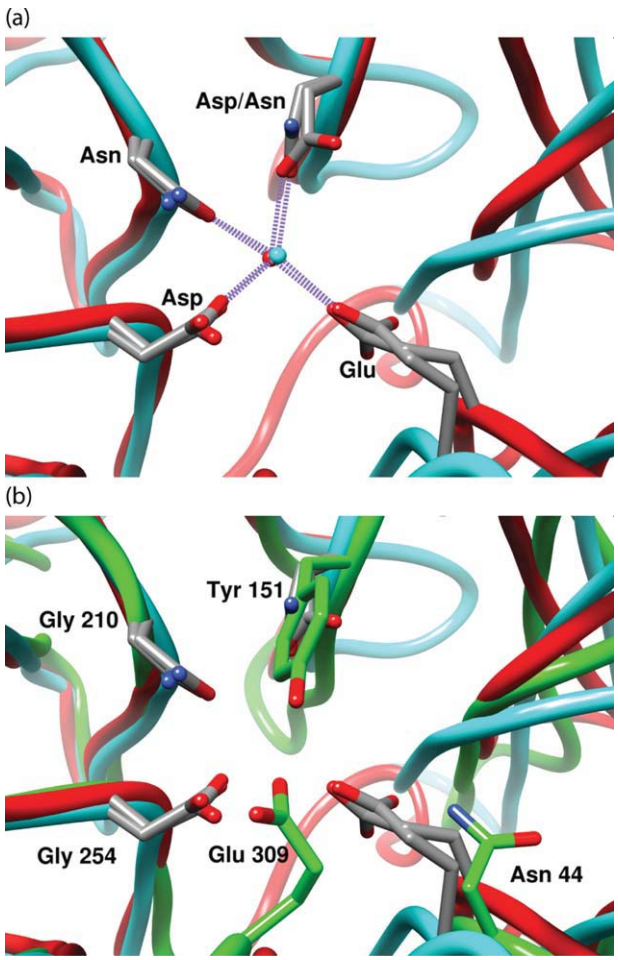

Figure 3

Active site superposition of (a) SGL (Drp35, pdb_id: 2dg1; red) and arylesterase-like (PON1, pdb_id: 1v04; cyan) subgroup proteins. Conserved metal coordinating residues (see text) are colored by element; gray: carbon, blue: nitrogen, red: oxygen. (b) Superposition of Drp35, PON1, and SS (pdb_id: 2fpb; green). SS residues are labeled black that superimpose with alpha carbon positions for metal coordinating residues in Drp35 and PON1. The glutamate (Glu309) required for SS activity in 2fpb is also labeled black. These five residues are colored by element; green: carbon, blue: nitrogen, red: oxygen. The metals shown in (a) have been removed for clarity. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

In contrast, the active site of the true SS from *R. serpentina* lacks all four of these metal coordinating ligands (Fig. 3b). Using Drp35 as a reference structure, SS aligns at 154 alpha carbon positions with an overall RMSD of 1.67 Å, suggesting significant overall structural similarity. However, while the secondary structure features of the SS and Drp35 active sites overlay, SS lacks all four metal-binding ligands, making the active site environment substantially different from those of Drp35 and PON1. This difference is consistent with a metal-independent mechanism in the Pictet–Spengler reaction that is thought to proceed using a catalytic glutamic acid from blade 6 (Glu309), which abstracts a proton from the amine group of the substrate tryptamine, allowing it to attack the aldehyde of secologanin. The indole then attacks the resulting iminium species to form the tetrahydro-β-carboline product, strictosidine.[56]

In an attempt to resolve the apparent contradiction in the relationships between sequence, structure, and reaction specificity among the three subgroups, we examined the SSL subgroup in greater detail as shown in Figures 4 and 5. In the network shown in Figure 4, increasing the stringency at which edges are drawn to an *E*-value less than 1×10^{-50} illustrates the similarity connections among the SSL subgroup proteins in more detail and shows how the network begins to "come apart" into smaller clusters representing higher levels of sequence similarity within each subset. From Figure 4, it is clear that the experimentally characterized true SSs lie at the periphery of the network, sharing similarity connections with only a subset of SSL subgroup proteins. Strikingly, the figure also shows that with the notable exception of the true SSs, the vast majority of the SSL proteins also conserve the four metal-binding ligands generally expected of SF members that catalyze hydrolytic reactions. Of those that do not (129 of the 516 sequences shown in Fig. 4), about half are so diverse from other sequences in the set that their alignments in the active site region are difficult to confirm, making it difficult to evaluate whether or not they are indeed missing one of more metal-binding ligands. Most of the others of these proteins have not been experimentally characterized. As a result, insufficient evidence is available to speculate further regarding their reaction specificities.

Although no structures are yet available for any of these SSL subgroup unknowns, their sequence conservation of the four "canonical" metal-binding ligands suggests that most are indeed metal dependent. Figure 5 provides a structure-guided sequence alignment comparing active site motifs from the structures in Figure 3 with nine SSL sequences (indicated by white arrows in Fig. 4) that are divergent from the true SSs. As indicated in the figure, these unknowns generally appear to conserve all the four metal-binding residues and also likely lack the catalytic glutamate required for catalysis of the SS reaction, again suggesting they are more likely to catalyze a hydrolytic reaction than the SS reaction. The recent identification of a low level of arylesterase activity in human APMAP,[32] an SSL protein that appears to have all four metal coordinating residues (gil24308201 in Fig. 5), provides some initial experimental validation of this prediction. However, more detailed analysis of many more of these proteins will be required to determine their reaction specificities or promiscuous capabilities for known reactions in the arylesterase-like and SGL subgroups. For example, we note that while SSL proteins do not appear to conserve the His–His dyad that PON1 requires for its lactonase/esterase activities (Supporting Information Fig. S1), other lactonase/esterase reactions in the characterized SGL enzymes also lack this feature.

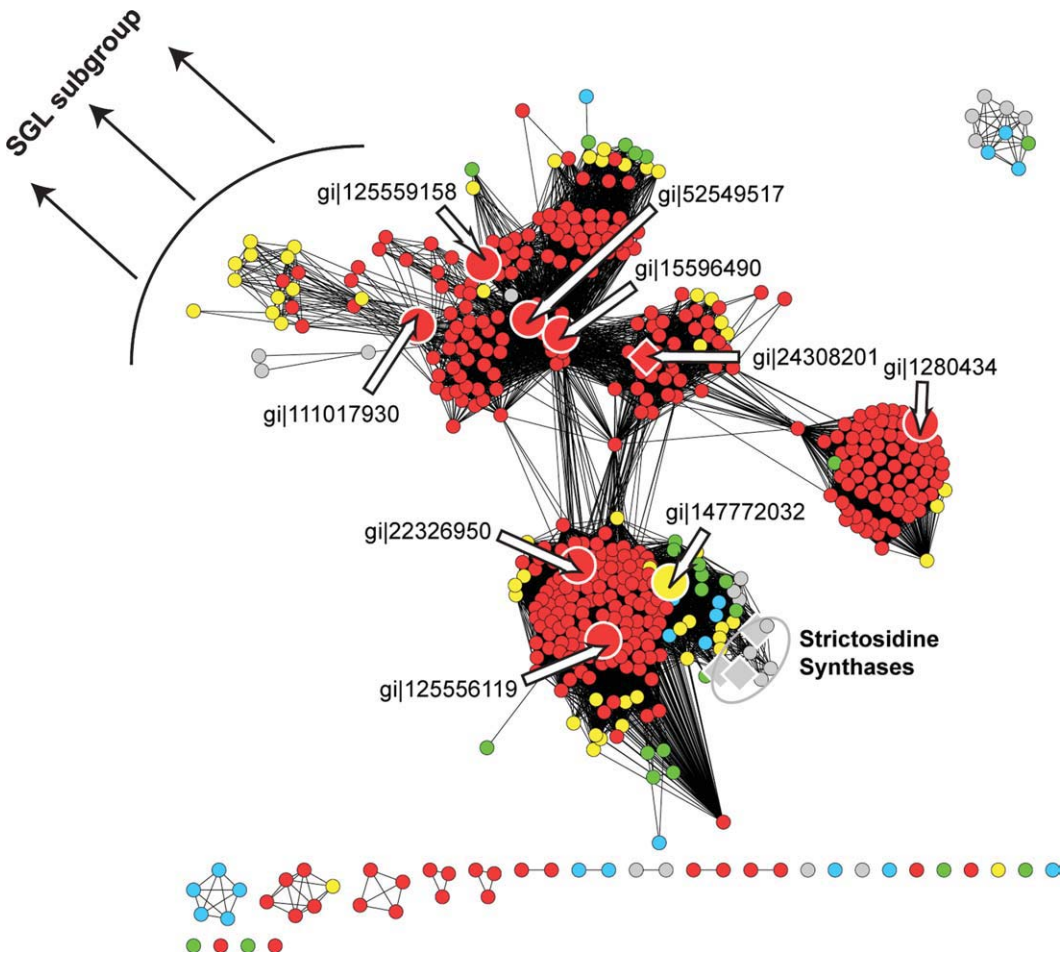

Figure 4

Sequence similarity network showing conservation of metal binding residues for the 516 SSL subgroup sequences generated from all-by-all pairwise comparisons. Nodes, edges, and layout are as in Figure 2 except that edges are drawn only for comparisons scoring better than an *E*-value threshold cutoff of 1E-50 (median alignment length = 297 residues, median percent identity = 41%). The arc with black arrows indicates the region of the network in which most of the connections between SSL proteins and the SGL subgroup are found (Fig. 2). Large diamond-shaped nodes: enzymes that have been biochemically and/or structurally characterized. Large nodes labeled with arrows and gi numbers: proteins shown in the motif alignment in Figure 5. True SSs are labeled and circled, including both biochemically characterized and predicted SSs (see Table I). Nodes are colored by number of metal-coordinating residues; all four present: red, 3 of 4: yellow, 2 of 4: green, 1 of 4: cyan, no metal-coordinating residues present: gray.

The SSL sequences most similar to the true SSs share between 37 and 49% identity with them, sufficiently distant to suggest that they may well catalyze reactions other than SS. Importantly, they also appear to lack the catalytic glutamate required for SS activity (data not shown except for the proteins labeled by gi numbers in Fig. 5), again supporting the hypothesis that they do not catalyze the SS reaction. Notably, several of these closest neighbors to the true SSs also lack one or more of the conserved metal-binding ligands typical of the hydrolytic SF members. Because it appears in Figure 5 to clearly lack the metal-binding ligand at position 210 and is also most similar to SS in the motif shown in the figure that is associated with that position, we chose the protein from *V. vinifera*, gi|147772032, for experimental examination of its ability to catalyze either a hydrolytic or the SS reaction. Consistent with our hypotheses, it was found to hydrolyze the model substrate *p*-nitrophenyl acetate (pNPAc), albeit at a lower level than the activity reported for PON1[19] but still significantly higher than background hydrolysis observed with *C. roseus* SS (Supporting Information Fig. S2). Steady-state kinetic analyses of the *V. vinifera* protein assayed with pNPAc revealed that the protein had a V_{max} of 220 ± 140 mol min^{-1} mol^{-1} and an estimated K_M of 8 ± 3 m*M*. Interestingly, the *V. vinifera* protein does not appear to catalyze the Pictet–Spengler condensation between tryptamine and secologanin (Supporting Information Fig. S3).

The low level of hydrolytic activity in this enzyme is not unexpected since neither its true substrate nor the

	44	151	210	254	309
PON1 (1v04.pdb)	GS**E**DLE	SV**N**DIVAVG	DVRVVAEGFDFA**N**GINISP	LV**D**NISVD	QGSTVAAV
Drp35 (2dg1.pdb)	QL**E**GLN	CI**D**DMVFDS	TVTPIIQNISVA**N**GIALST	GP**D**SCCID	LRSTHPQF
DFPase (1pjx.pdb)	GA**E**GPV	GC**N**DCAFDY	QMIQVDTAFQFP**N**GIAVRH	GA**D**GMDFD	EKPSNLHF
SS (2fpb.pdb)	APNSFT	WLYAVTVDQ	ETTLLLKELHVPGGAEVSA	NPGNIKRN	EHF**E**QIQE
gi|147772032	GP**E**AIA	FL**N**AVDVDQ	EVTVLLRGLGGAGGVTISK	TP**D**NIKRN	KTISEVQE
gi|22326950	GP**E**SVA	FT**N**DLDIAD	KAVVLVSNLQFP**N**GVSISR	HP**D**NVRTN	RSVSEVEE
gi|125556119	GP**E**SVA	FT**N**GVDIDQ	QVTVLQSNITYP**N**GVAISA	YP**D**NVRPD	RP-TEVMD
gi|24308201	GP**E**SIA	FV**N**DLTVTQ	EVKVLLDQLRFP**N**GVQLSP	FP**D**NIRPS	TYISEVHE
gi|1280434	GP**E**CLI	IF**N**GVTVSK	VSEVLLDELAFA**N**GLALSP	LP**D**NLTPD	T-ISHVLE
gi|125559158	AP**E**DVY	FA**D**AAIEAS	EASVVLDGLGFA**N**GVALPP	NP**D**NIRLG	NMVTSVTE
gi|111017930	GP**E**DVA	AC**N**NSAVGR	ETDLLAEGLQFA**N**GVGLAS	IP**D**NMTSQ	P-VTGVRE
gi|15596490	GP**E**DTA	FT**D**DLDIAS	KTEVLLKDLYFA**N**GVALSA	LP**D**NLQGD	RMITSAKP
gi|52549517	GP**E**DVA	LT**D**DVDIAA	TTRLVLNNLYFA**N**GVAVSP	FP**D**GISSN	Q-ITSVQE

Figure 5

Active site motif alignments of nine SSL subgroup proteins with structures from the arylesterase-like (blue, pdb_id:1v04), SGL (red, pdb_id's: 2dg1, 1pjx) and SSL subgroups (green, pdb_id: 2fpb). Numbering is from the SS structure, also highlighted in green. Highlighted in bold and yellow: four conserved metal coordinating active site residues that are found in the characterized arylesterase-like and SGL subgroup members and the majority of uncharacterized sequences in the SSL subgroup (indicated by gi numbers, Figure 4), but not in the true SS (row highlighted in green). Highlighted in bold and gray: catalytic glutamate required for SS activity. Note that in PON1 (1v04.pdb), a fifth residue (N255) is thought to coordinate to the calcium; this residue is conserved in many of the SSL proteins shown and may play a similar role. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

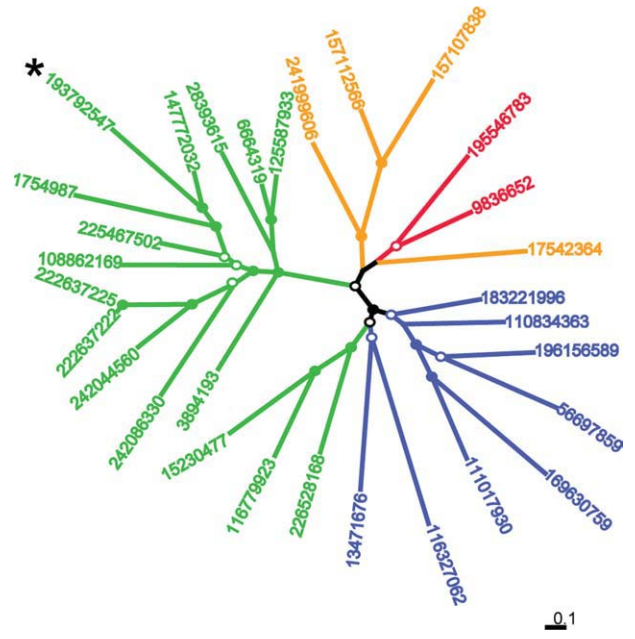

Figure 6

Bayesian phylogenetic tree of a representative subset of SSL subgroup. Thirty proteins for which no two proteins share greater than 40% identity were used. Coloring of leaves and branches is by type of life: plants, green; bacteria, blue; vertebrate animals, red; invertebrate animals, orange. Characterized SS (gi|193792547) is indicated by an asterisk. Branch confidence values: >0.95, filled circle; 0.70–0.94, open circle; 0.65–0.69, no circle.

consequences of missing one of the metal-binding ligands are known. This missing metal ligand, a highly conserved Asn that aligns in all of the presumed hydrolytic enzymes shown in Figure 5 (position 210), is replaced by a Gly in both the *V. vinifera* protein and the SS. Although this asparagine has been shown to play a role in maintaining the electrostatic environment necessary for the degradation of toxic organophosphates by DFPase in the SGL subgroup,[30] it might be less important for other hydrolytic reactions such as those catalyzed by some lactonases or esterases.[25] Alternatively, a subset of SSL proteins lacking one or more canonical catalytic residues may play a regulatory role in some as yet unknown cellular process (e.g. in other systems, see Ref. 57).

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

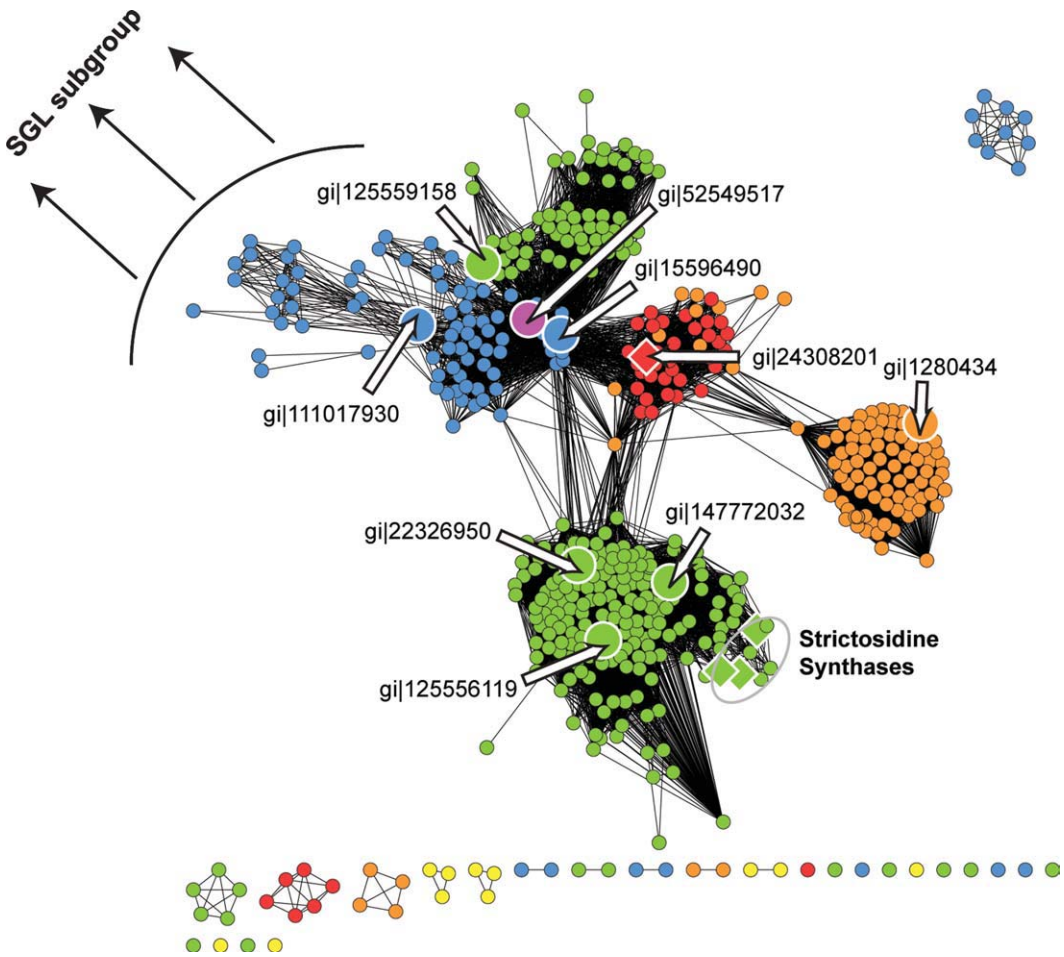

Figure 7

Sequence similarity network showing types of life for the 516 SSL subgroup sequences generated from all-by-all pairwise comparisons. Nodes, edges, and layout are as in Figure 2 except that edges are drawn only for comparisons scoring better than an *E*-value threshold cutoff of 1E-50 (median alignment length = 297 residues, median percent identity = 41%). The arc with black arrows indicates the region of the network in which most of the connections between SSL proteins and the SGL subgroup are found (Figure 2). Large diamond-shaped nodes: enzymes that have been biochemically and/or structurally characterized. Large nodes labeled with arrows and gi numbers: proteins shown in the motif alignment in Figure 5. True SSs are labeled and circled, including both biochemically characterized and predicted SSs (see Table I). Coloring is by "type of life"; bacteria: blue, plants: green, archaea: magenta, vertebrate animals: red, invertebrate animals: orange, protozoa: yellow.

SS may have evolved from an ancestor with metal-coordinating active site residues

To gain additional clues regarding how the SS reaction could have evolved in the context of SSL subgroup ancestry, we constructed a phylogenetic tree for a representative subset of the subgroup (Fig. 6). The clustering patterns of the leaves of the tree appear similar to the clustering patterns seen in the sequence similarity networks (Fig. 4), with clusters generally correlating with type of life (Fig. 7). Two distinct clades of plant proteins are present: one corresponding to the plant only cluster of the SSL subgroup network, and the other clade joined with bacterial proteins corresponding to the "mixed" cluster of the SSL subgroup in Figure 7. Additionally, separate clades for vertebrate and for invertebrate animal proteins can be distinguished; however, the interior node joining these two clades is not well resolved.

As noted earlier, gi|147772032 from *V. vinifera* may reflect features of an evolutionary transition between hydrolytic SSLs and SS in the ancestry of the subgroup, displaying characteristics of the metal-coordinated active site common to most of the SF, as well as characteristics more similar to those in the SS active site motif (Fig. 5). An evaluation of the position of this sequence in the tree relative to characterized SS may provide clues about how the SS function could have arisen. It is therefore interesting that a characterized SS protein (gi|193792547) is the nearest neighbor of gi|147772032 in the tree. The next most interior node, gi|1754987 (not shown in Fig. 5), also appears to share three of the four metal coordinating residues common to most of the SF members, with the Asn/Gly position

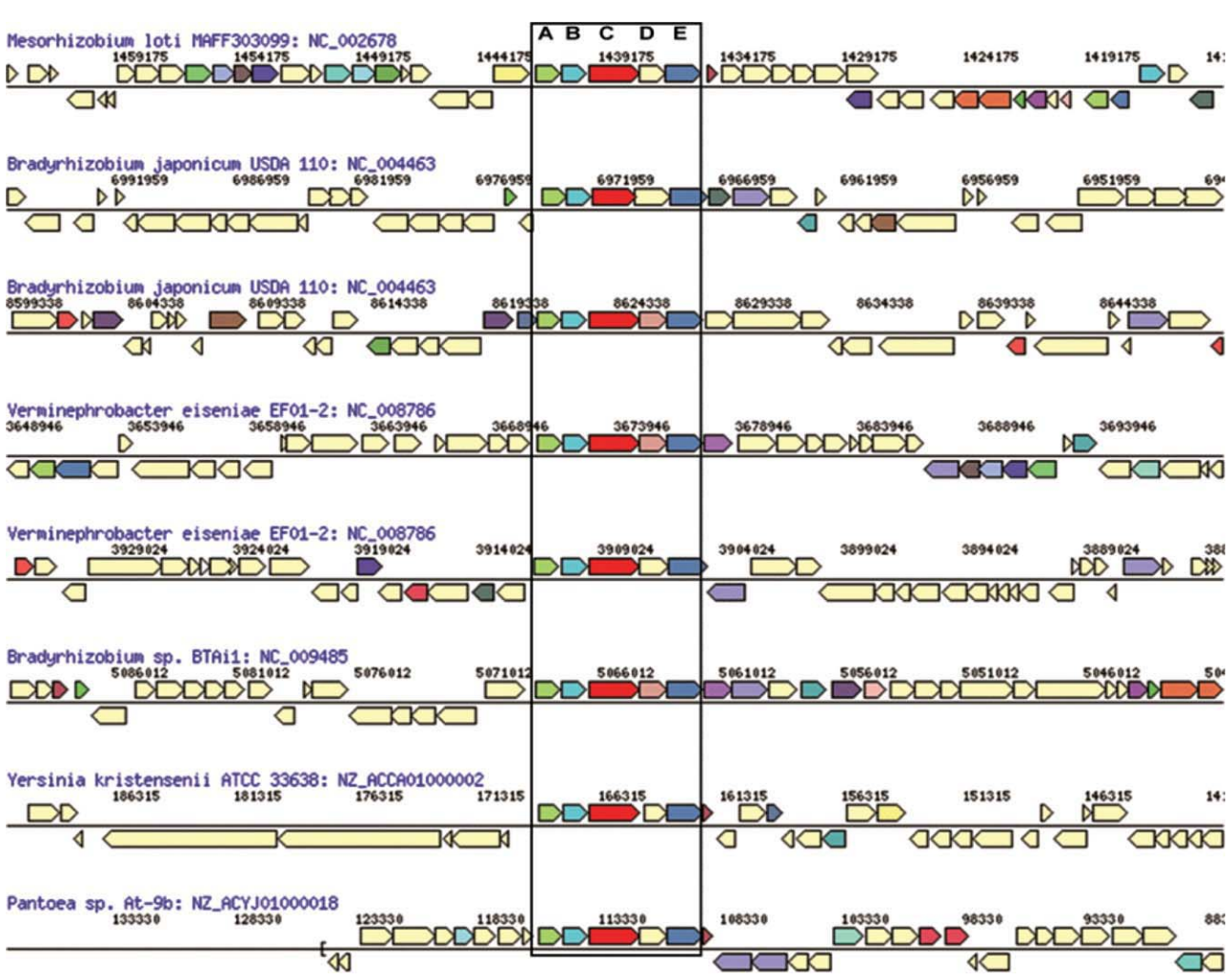

Figure 8

Gene neighborhood for SSL proteins in gram-negative bacteria as seen on the Integrated Microbial Genomes (IMG) system.[46] Colored genes (other than light-yellow) have orthologous components in another organism. Boxed: permease protein of sugar ABC transporters (light green, labeled A); binding protein components of sugar ABC transporter (cyan, labeled B); a second permease protein of sugar ABC transporter, but in this case fused to a SSL protein (red, labeled C); a second, independent SSL protein (light-pink or in some cases, light-yellow, when an orthologous component is not detected, labeled D); and an ATP binding protein (light blue, labeled E). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

(position 254 in Fig. 5) substituted by an Ala. Though the interior node joining the next closest sequence (gi|225467502) with these three others is not well-resolved, the interior node distinguishing the plant only cluster from the rest of the SSL subgroup is resolved to a posterior probability of 1. All the proteins contained therein, with the sole exception of the true SS, conserve some or all four of the metal coordinating active site residues common to the rest of the SF, lending support to the idea that the SSs, found only in recently evolved higher plants, may have diverged from a metal-dependent ancestral-like protein.

It is also possible that there are more complex evolutionary origins to the functions represented in the SSL subgroup, and that SS and SSL proteins may be paralogous rather than orthologous. Sequencing of the genomes of additional higher plants to fill in the sequence links among these very diverse proteins, along with further experimental characterization of SSL proteins, will be required to address this issue.

Biological information provides clues about functional properties of some SSL subgroup proteins

Besides information from homology, many other types of information have been used to infer functional properties of sequences discovered in genome projects (see Ref. 58 for a recent review). Gene context suggests that several SSL proteins from gram-negative bacteria may function in ABC transport, most likely involved in the

uptake of carbohydrates (Fig. 8). These proteins are sometimes fused to ABC transmembrane domains (such as gi|13471676 from *Mesorhizobium loti MAFF303099*), but can also appear as independent domains proximal to genes encoding ABC transport machinery, such as an ATP-binding domain or substrate-binding protein. Based on the sequence similarity of these ABC transporter components (not shown), we can hypothesize that these proteins are members of the carbohydrate uptake transporter-2 (CUT2) subfamily. Proteins in this subfamily are known to import monosaccharides such as ribose[59] and xylose[60] as well as ribonucleotides.[61] The proximity of at least two of these proteins (gi|238761435 from *Yersinia kristensenii ATCC 33638* and gi|258637446 from *Pantoea* sp. *At-9b*) to genes encoding proteins similar to carboxymuconolactone decarboxylases suggest that these transporters likely import lactones and related compounds. Consistent with that interpretation, the presence of metal-coordinating residues in all of the SSL domains in this set of proteins suggests a potential enzymatic role as well, perhaps functioning as the first step in a secondary sugar metabolism pathway.

Other SSL proteins appear to be involved in immunity. For example, hemomucin, an innate immune receptor in flies,[62,63] is represented in the cluster of proteins from invertebrates (Fig. 7; orange). All of the proteins in this cluster have an SSL domain while only some have the additional mucin-type repeats described elsewhere.[64] Interestingly, nearly all these proteins appear to have the metal coordinating residues associated with hydrolytic activity (Fig. 4). Whether the presence of these residues confers hydrolytic activity and whether its unknown activity is required for signal transduction by the hemomucin receptor remain to be tested.

Comparison of liganded SGL and SSL subgroup structures reveals intriguing similarities in active site mechanisms despite the differences between their divergent reactions and active sites

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

In contrast to the metal-independent Pictet–Spengler reaction[56] catalyzed by SS, conservation of metal ligands across the rest of the N6P SF appears to be a fundamental requirement for the activation of substrates for hydrolysis and the stabilization of intermediates and transition states. For its lactonase and esterase activities,[39] PON1 requires these conserved metal coordinating residues as well as a fifth metal-binding residue and a His–His dyad (Supporting Information Fig. S1). DFPase lacks both the His–His dyad and the fifth metal-binding residue and uses its four metal-coordinating residues directly in the phosphotriesterase mechanism.[30,31,65,66] Superposition of the liganded DFPase and SS structures reveals similarity in the positions of two of these metal ligands in DFPase with that of bound tryptamine from SS (Fig. 9). Remarkably, the position from which the direct nucleophilic attack that forms a phosphoenzyme intermediate in DFPase occurs is in a similar spatial orientation to that involved in the nucleophilic attack by tryptamine on the aldehyde of secologanin to form the carbinolamine intermediate in SS (Fig. 9, Supporting Information Fig. S4).

Based on this structural similarity and the known mechanisms of these enzymes, we speculate that the substitution of two conserved metal residues to glycines in SS (positions 210 and 254 in Fig. 5), which includes the loss of the aspartate required in the phosphotriesterase mechanism in DFPase, may have created space for the binding of tryptamine in the SS active site. The eventual loss of the other, now unneeded, metal coordinating residues could then have occurred over the course of evolution. Consistent with this notion, the amino group in tryptamine seems to play a role in the mechanism of SS that is analogous to the role that the nucleophilic oxygen atom of the conserved aspartic acid plays in the phosphotriesterase mechanism of DFPase. That is, both reaction mechanisms involve a nucleophilic attack on a sp^2-hybridized electrophilic atom from the same part of the active site, one using metal-assisted catalysis and the other using substrate-assisted catalysis.

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

SUMMARY

In this article, we examine the largely uncharacterized SSL subgroup of the N6P SF in the context of the known

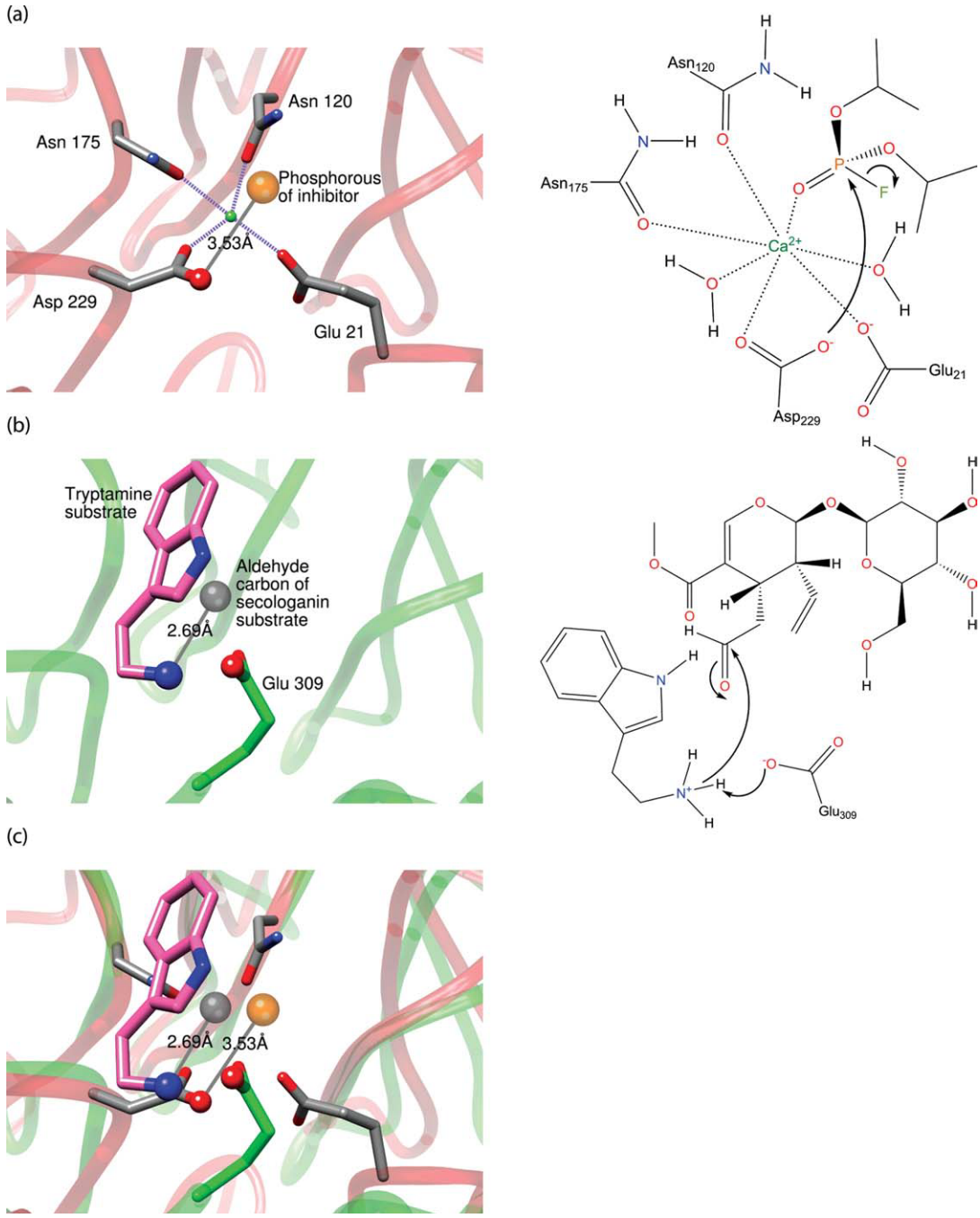

Figure 9

A related catalytic strategy unites the SS enzymes with the rest of the N6P SF. Left panels: active site representations; right panels: diagrams depicting steps in the catalytic mechanisms as described in Refs. 30 and 56. (a) The left panel shows the active site of DFPase (pdb_id: 2gvv; red backbone, SGL subgroup) with dicyclopentyl phosphoramidate inhibitor bound (orange; only the phosphorous of the inhibitor is depicted). The nucleophilic oxygen of the catalytic Asp229, and the electrophilic phosphorous atom of the inhibitor are depicted as ball and stick. Two water molecules and the phosphoryl oxygen coordinated to the metal have been removed for clarity. The right panel depicts the nucleophilic attack of the oxygen of Asp229 on the phosphorous of the substrate DFP. (b) The left panel shows the superposition of SS (green backbone) with tryptamine bound (pdb_id: 2fpb; pink) and with secologanin bound (pdb id: 2fpc; gray, only the aldehyde carbon is depicted). The acidic oxygen atom of the catalytic Glu309, the reactive nitrogen atom of tryptamine, and the electrophilic carbon of secologanin are shown as ball and stick. The right panel shows the deprotonation of the amine group of tryptamine by Glu309 and the subsequent nucleophilic attack of the amine on the aldehyde of secologanin. (c) Superposition of the proteins shown in (a) and (b). Atoms and ligands are depicted as in (a) and (b), with the metal from (a) removed for clarity. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

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

Access to data from this work

The data presented here for the SSL subgroup, including SSL and SS sequence and structure data and full-length alignments of representative sequences identifying key conserved amino acids, have been added to the Structure-Function Linkage Database (SFLD) (http://sfld.rbvi.ucsf.edu).[51] Interactive versions of the networks, allowing users to examine the larger similarity context for specific proteins in the set, are freely available for download.

ACKNOWLEDGMENTS

This work was supported by NIH R01 (GM074820) to SEO and NIH R01 GM60595 and U54 GM093342 to PCB. AEB was supported by the ARCS Foundation and the PhRMA Foundation Predoctoral Informatics fellowship. LAG was supported by a National Science Foundation Predoctoral fellowship and the Ford Foundation Predoctoral fellowship. Molecular graphics images were produced using the UCSF Chimera package from the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (supported by NIH P41 RR001081). We also thank Peter Bernhardt for CrSTR-pET28a (+), Dan Tawfik for PON1-pET32b (+) constructs.

REFERENCES

1. The Uniprot Consortium. Ongoing and future developments at the Universal Protein Resource. Nucleic Acids Res 2011;39(Suppl 1):D214–D219.
2. Schnoes AM, Brown SD, Dodevski I, Babbitt PC. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Comput Biol 2009;5:e1000605.
3. Gerlt JA, Babbitt PC. Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. Annu Rev Biochem 2001;70:209–246.
4. Khersonsky O, Tawfik DS. Enzyme promiscuity: a mechanistic and evolutionary perspective. Annu Rev Biochem 2010;79:471–505.
5. Todd AE, Orengo CA, Thornton JM. Evolution of function in protein superfamilies, from a structural perspective. J Mol Biol 2001;307:1113–1143.
6. Glasner ME, Gerlt JA, Babbitt PC. Evolution of enzyme superfamilies. Curr Opin Chem Biol 2006;10:492–497.
7. Kobayashi M, Shinohara M, Sakoh C, Kataoka M, Shimizu S. Lactone-ring-cleaving enzyme: genetic analysis, novel RNA editing, and evolutionary implications. Proc Natl Acad Sci USA 1998;95:12787–12792.
8. Kutchan TM, Hampp N, Lottspeich F, Beyreuther K, Zenk MH. The cDNA clone for strictosidine synthase from *Rauvolfia serpentina* DNA sequence determination and expression in *Escherichia coli*. FEBS Lett 1988;237:40–44.
9. McKnight TD, Roessner CA, Devagupta R, Scott AI, Nessler CL. Nucleotide sequence of a cDNA encoding the vacuolar protein strictosidine synthase from *Catharanthus roseus*. Nucleic Acids Res 1990;18:4939.
10. Yamazaki Y, Sudo H, Yamazaki M, Aimi N, Saito K. Camptothecin biosynthetic genes in hairy roots of *Ophiorrhiza pumila*: cloning, characterization and differential expression in tissues and by stress compounds. Plant Cell Physiol 2003;44:395–403.
11. Bracher D, Kutchan TM. Strictosidine synthase from *Rauvolfia serpentina*: analysis of a gene involved in indole alkaloid biosynthesis. Arch Biochem Biophys 1992;294:717–723.
12. Wang H, Chen R, Chen M, Sun M, Liao Z-H. Cloning and analysis of strictosidine synthase in *Rauvolfia verticillata*. Xibei Zhiwu Xuebao 2006;26:900–905.
13. Lu Y, Wang H, Wang W, Qian Z, Li L, Wang J, Zhou G, Kai G. Molecular characterization and expression analysis of a new cDNA encoding strictosidine synthase from *Ophiorrhiza japonica*. Mol Biol Rep 2009;36:1845–1852.
14. Stockigt J, Barleben L, Panjikar S, Loris EA. 3D-structure and function of strictosidine synthase—the key enzyme of monoterpenoid indole alkaloid biosynthesis. Plant Physiol Biochem 2008;46:340–355.
15. Chen S, Galan MC, Coltharp C, O'Connor SE. Redesign of a central enzyme in alkaloid biosynthesis. Chem Biol 2006;13:1137–1141.
16. McCoy E, O'Connor SE. Directed biosynthesis of alkaloid analogs in the medicinal plant *Catharanthus roseus*. J Am Chem Soc 2006;128:14276–14277.
17. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer ELL, Eddy SR, Bateman A. The Pfam protein families database. Nucleic Acids Res 2010;38(Suppl_1):D211–222.
18. Draganov DI, Teiber JF, Speelman A, Osawa Y, Sunahara R, La Du BN. Human paraoxonases (PON1, PON2, and PON3) are lactonases with overlapping and distinct substrate specificities. J Lipid Res 2005;46:1239–1247.
19. Khersonsky O, Tawfik DS. Structure-reactivity studies of serum paraoxonase PON1 suggest that its native activity is lactonase. Biochemistry 2005;44:6371–6382.
20. Billecke S, Draganov D, Counsell R, Stetson P, Watson C, Hsu C, Du BNL. Human serum paraoxonase (pon1) isozymes Q and R hydrolyze lactones and cyclic carbonate esters. Drug Metab Dispos 2000;28:1335–1342.
21. Draganov DI. Lactonases with oragnophosphatase activity: structural and evolutionary perspectives. Chem-Biol Interact 2010;187:370–372.
22. Teiber JF, Draganov DI, Du BNL. Lactonase and lactonizing activities of human serum paraoxonase (PON1) and rabbit serum PON3. Biochem Pharmacol 2003;66:887–896.

23. Kondo Y, Inai Y, Sato Y, Handa S, Kubo S, Shimokado K, Goto S, Nishikimi M, Maruyama N, Ishigami A. Senescence marker protein 30 functions as gluconolactonase in L-ascorbic acid biosynthesis, and its knockout mice are prone to scurvy. Proc Natl Acad Sci USA 2006;103:5723–5728.
24. Kondo Y, Ishigami A, Kubo S, Handa S, Gomi K, Hirokawa K, Kajiyama N, Chiba T, Shimokado K, Maruyama N. Senescence marker protein-30 is a unique enzyme that hydrolyzes diisopropyl phosphorofluoridate in the liver. FEBS Lett 2004;570:57–62.
25. Tanaka Y, Morikawa K, Ohki Y, Yao M, Tsumoto K, Watanabe N, Ohta T, Tanaka I. Structural and mutational analyses of Drp35 from *Staphylococcus aureus*: a possible mechanism for its lactonase activity. J Biol Chem 2007;282:5770–5780.
26. Murakami H, Matsumaru H, Kanamori M, Hayashi H, Ohta T. Cell wall-affecting antibiotics induce expression of a novel gene, drp35, in *Staphylococcus aureus*. Biochem Biophys Res Commun 1999;264:348–351.
27. Scharff EI, Koepke J, Fritzsch G, Lucke C, Ruterjans H. Crystal structure of diisopropylfluorophosphatase from *Loligo vulgaris*. Structure 2001;9:493–502.
28. Gomi K, Kajiyama N. Oxyluciferin, a luminescence product of firefly luciferase, is enzymatically regenerated into luciferin. J Biol Chem 2001;276:36508–36513.
29. Katsemi V, Lucke C, Koepke J, Lohr F, Maurer S, Fritzsch G, Ruterjans H. Mutational and structural studies of the diisopropylfluorophosphatase from *Loligo vulgaris* shed new light on the catalytic mechanism of the enzyme. Biochemistry 2005;44:9022–9033.
30. Blum M-M, Löhr F, Richardt A, Rüterjans H, Chen JCH. Binding of a designed substrate analogue to diisopropyl fluorophosphatase: implications for the phosphotriesterase mechanism. J Am Chem Soc 2006;128:12750–12757.
31. Blum M-M, Chen JCH. Structural characterization of the catalytic calcium-binding site in diisopropyl fluorophosphatase (DFPase)—comparison with related β-propeller enzymes. Chem-Biol Interact 2010;187:373–379.
32. Ilhan A, Gartner W, Nabokikh A, Daneva T, Majdic O, Cohen G, Böhmig GA, Base W, Hörl WH, Wagner L. Localization and characterization of the novel protein encoded by C20orf3. Biochem J 2008;414:485–495.
33. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
34. The UniProt Consortium. The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res;38(Suppl_1):D142–D148.
35. Eddy SR. A new generation of homology search tools based on probabilistic inference. Genome Inform 2009;23:205–211.
36. Pei J, Kim B-H, Grishin NV. PROMALS3D: a tool for multiple protein sequence and structure alignments. Nucleic Acids Res 2008;36:2295–2300.
37. Atkinson HJ, Morris JH, Ferrin TE, Babbitt PC. Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. PLoS One 2009;4:e4345.
38. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003;13:2498–2504.
39. Harel M, Aharoni A, Gaidukov L, Brumshtein B, Khersonsky O, Meged R, Dvir H, Ravelli RB, McCarthy A, Toker L, Silman I, Sussman JL, Tawfik DS. Structure and evolution of the serum paraoxonase family of detoxifying and anti-atherosclerotic enzymes. Nat Struct Mol Biol 2004;11:412–419.
40. Koepke J, Scharff EI, Lucke C, Ruterjans H, Fritzsch G. Statistical analysis of crystallographic data obtained from squid ganglion DFPase at 0.85 A resolution. Acta Crystallogr D Biol Crystallogr 2003;59(Part 10):1744–1754.
41. Ma X, Panjikar S, Koepke J, Loris E, Stockigt J. The structure of *Rauvolfia serpentina* strictosidine synthase is a novel six-bladed beta-propeller fold in plant proteins. Plant Cell 2006;18:907–920.
42. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 1970;48:443–453.
43. Meng E, Pettersen E, Couch G, Huang C, Ferrin T. Tools for integrated sequence-structure analysis with UCSF Chimera. BMC Bioinformatics 2006;7:339.
44. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. UCSF chimera—a visualization system for exploratory research and analysis. J Comput Chem 2004;25:1605–1612.
45. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004;32:1792–1797.
46. Markowitz VM, Chen IMA, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Anderson I, Lykidis A, Mavromatis K, Ivanova NN, Kyrpides NC. The integrated microbial genomes system: an expanding comparative analysis resource. Nucleic Acids Res 2010;38(Suppl 1):D382–D390.
47. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006;22:1658–1659.
48. Altekar G, Dwarkadas S, Huelsenbeck JP, Ronquist F. Parallel metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. Bioinformatics 2004;20:407–415.
49. Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 2003;19:1572–1574.
50. Whelan S, Goldman N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 2001;18:691–699.
51. Pegg SCH, Brown SD, Ojha S, Seffernick J, Meng EC, Morris JH, Chang PJ, Huang CC, Ferrin TE, Babbitt PC. Leveraging enzyme structure-function relationships for functional inference and experimental design: the structure-function linkage database. Biochemistry 2006;45:2545–2555.
52. McCoy E, Galan MC, O'Connor SE. Substrate specificity of strictosidine synthase. Bioorg Med Chem Lett 2006;16:2475–2478.
53. Aharoni A, Gaidukov L, Yagur S, Toker L, Silman I, Tawfik DS. Directed evolution of mammalian paraoxonases PON1 and PON3 for bacterial expression and catalytic specialization. Proc Natl Acad Sci USA 2004;101:482–487.
54. Chakraborti S, Bahnson BJ. Crystal structure of human senescence marker protein 30: insights linking structural, enzymatic, and physiological functions. Biochemistry 2010;49:3436–3444.
55. Khersonsky O, Tawfik DS. The histidine 115-histidine 134 dyad mediates the lactonase activity of mammalian serum paraoxonases. J Biol Chem 2006;281:7649–7656.
56. Maresh JJ, Giddings LA, Friedrich A, Loris EA, Panjikar S, Trout BL, Stockigt J, Peters B, O'Connor SE. Strictosidine synthase: mechanism of a Pictet-Spengler catalyzing enzyme. J Am Chem Soc 2008;130:710–723.
57. Pils B, Schultz J. Inactive enzyme-homologues find new function in regulatory processes. J Mol Biol 2004;340:399–404.
58. Rentzsch R, Orengo CA. Protein function prediction—the power of multiplicity. Trends Biotechnol 2009;27:210–219.
59. Bell AW, Buckel SD, Groarke JM, Hope JN, Kingsley DH, Hermodson MA. The nucleotide sequences of the rbsD, rbsA, and rbsC genes of *Escherichia coli* K12. J Biol Chem 1986;261:7652–7658.
60. Erbeznik M, Hudson SE, Herrman AB, Strobel HJ. Molecular analysis of the xylFGH Operon, coding for xylose ABC transport, in *Thermoanaerobacter ethanolicus*. Curr Microbiol 2004;48:295–299.
61. Webb AJ, Hosie AHF. A member of the second carbohydrate uptake subfamily of ATP-binding cassette transporters is responsible for ribonucleoside uptake in *Streptococcus mutans*. J Bacteriol 2006;188:8005–8012.

62. Theopold U, Samakovlis C, Erdjument-Bromage H, Dillon N, Axelsson B, Schmidt O, Tempst P, Hultmark D. *Helix pomatia* lectin, an inducer of drosophila immune response, binds to hemomucin, a novel surface mucin. J Biol Chem 1996;271:12708–12715.
63. Schmidt O, Söderhäll K, Theopold U, Faye I. Role of adhesion in arthropod immune recognition. Annu Rev Entomol 2010;55: 485–504.
64. Fabbri M, Delp G, Schmidt O, Theopold U. Animal and plant members of a gene family with similarity to alkaloid-synthesizing enzymes. Biochem Biophys Res Commun 2000; 271:191–196.
65. Melzer M, Chen JCH, Heidenreich A, Gäb J, Koller M, Kehe K, Blum M-M. Reversed enantioselectivity of diisopropyl fluorophosphatase against organophosphorus nerve agents by rational design. J Am Chem Soc 2009;131:17226–17232.
66. Blum MM, Timperley CM, Williams GR, Thiermann H, Worek F. Inhibitory potency against human acetylcholinesterase and enzymatic hydrolysis of fluorogenic nerve agent mimics by human paraoxonase 1 and squid diisopropyl fluorophosphatase. Biochemistry 2008;47:5216–5224.
67. Hoskin FCG, Roush AH. Hydrolysis of nerve gas by squid-type diisopropyl phosphorofluoridate hydrolyzing enzyme on agarose resin. Science 1982;215:1255–1257.
68. Sakayu S, Michihiko K, Kentaro S, Masao H, Keiji S, Hideaki Y. Purification and characterization of a novel lactonohydrolase, catalyzing the hydrolysis of aldonate lactones and aromatic lactones, from *Fusarium oxysporum*. Eur J Biochem 1992;209:383–390.
69. Kanagasundaram V, Scopes R. Isolation and characterization of the gene encoding gluconolactonase from *Zymomonas mobilis*. Biochim Biophys Acta 1992;1171:198–200.