Protein salvage and repurposing in evolution: phospholipase D toxins are stabilized by a remodeled scrap of a membrane association domain

Matthew H.J. Cordes[1*], Alexandra K. Sundman[1], Holden C. Fox[1] and Greta J. Binford[2]

[1]Department of Chemistry and Biochemistry, University of Arizona, Tucson, Arizona, 85701

[2]Department of Biology, Lewis & Clark College, Portland, Oregon, 97219

*To whom correspondence should be addressed: cordes@email.arizona.edu

Running title: Salvage and repurposing of protein domain scraps in evolution

Manuscript: 24 pages including references, 5 figures, 5 supplementary tables (S1 to S5; Cordes_supplementary_tables.xlsx), 1 supplementary figure

ABSTRACT

The GDPD-like SMaseD/PLD domain family, which includes phospholipase D (PLD) toxins in recluse spiders and actinobacteria, evolved anciently in bacteria from the glycerophosphodiester phosphodiesterases (GDPD). The PLD enzymes retained the core $(\beta/\alpha)_8$ barrel fold of GDPD, while gaining a signature C-terminal expansion motif and losing a small insertion domain. Using sequence alignments and phylogenetic analysis, we infer that the C-terminal motif derives from a segment of an ancient bacterial PLAT domain. Formally, part of a protein containing a PLAT domain repeat underwent fusion to the C terminus of a GDPD barrel, leading to attachment of a segment of a PLAT domain, followed by a second complete PLAT domain. The complete domain was retained only in some basal homologs, but the PLAT segment was conserved and repurposed as the expansion motif. The PLAT segment corresponds to strands β7-β8 of a β-sandwich, while the expansion motif as represented in spider PLD toxins has been remodeled as an α-helix, a β-strand, and an ordered loop. The GDPD-PLAT fusion led to two acquisitions in founding the GDPD-like SMaseD/PLD family: 1) a PLAT domain that presumably supported early lipase activity by mediating membrane association, and 2) an expansion motif that putatively stabilized the catalytic domain, possibly compensating for, or permitting, loss of the insertion domain. Of wider significance, messy domain shuffling events can leave behind scraps of domains that can be salvaged, remodeled and repurposed.

IMPACT AND IMPORTANCE

The basic architecture of proteins consists of functional and structural modules called domains that can be mixed and matched by gene recombination and fusion during evolution. We show that evolution can also subdivide these modules in creative ways, using remodeled scraps of domains to build additions onto other domains.

KEYWORDS

Domain expansion, phospholipase D, spider venom, structural evolution, chimeric domain

ABBREVIATIONS AND SYMBOLS

Introduction

Domains are basic modules of protein evolution and are commonly shuffled or recombined between proteins as whole functional and structural units to yield diverse multidomain architectures with an array of capabilities [1-3]. The structure of a classic, independently folded domain is not cleanly divisible into smaller units, but depends on a cooperative, global network of interactions. To shuffle incomplete domains between different proteins is generally disruptive, especially if the recombination is not homologous. The evolution of protein architecture appears to principally involve modular exchange of complete domains, which is easier and somewhat more predictable in outcome.

Recombination involving parts of domains can happen, however, and has considerable potential to generate molecular novelty. Domains have some internal modularity, being composed of secondary structure elements and motifs that recur in different domains. Design and artificial selection studies have shown that recombination between motifs or segments from completely unrelated domains can yield chimeras with novel folded structures [4-6]. Interestingly, the global control of domain structure still asserts itself, such that the structure of a segment can be remodeled and repurposed in unexpected ways when incorporated into a new domain [7]. Credible theories also posit that early in the history of life, domains originated by mixing and joining of antecedent domain segments [8-11]. Finally, the potential exists both in design and evolution to expand existing domains with little disruption by adding fragments of other domains. Despite the creative potential of domain chimerism, however, it remains an open question what role, if any, it plays in ongoing natural evolution. Here we provide evidence that chimerism aided the evolution of a family of phospholipases through domain expansion.

The GDPD-like SMaseD/PLD family is a group of phospholipase D (PLD) enzyme domains descended from the ubiquitous glycerophosphodiester phosphodiesterases (GDPD) [12-14]. In a previous phylogenetic study, we showed that this family arose anciently in bacteria and radiated into a sparse, diverse collection of organisms, at least in part through ancient lateral gene transfer [15]. Some extant members of the family are important necrotoxins and hemotoxins to mammals and/or neurotoxins to invertebrates. These include recluse spider and scorpion toxins [16-18], as well as toxins from pathogenic *Corynebacteria* and *Arcanobacteria* [19].

GDPD-like SMaseD/PLD toxins differ in function from their GDPD ancestors. First, the major known SMaseD/PLD substrates include common membrane lipids such as sphingomyelin [20], while GDPDs principally act on nonlipid substrates such as glycerophosphocholine [21]. Multiple members of both families have activity against lysophospholipids, however [21-25]. Second, the PLD toxins catalyze conversion of the substrate to a cyclic phosphate by intramolecular attack of a hydroxyl group at phosphorus [24,26]; GDPD enzymes have been proposed to produce cyclic phosphates as intermediates but then catalyze their hydrolysis in a second step [27]. Third, the PLD toxins are extracellular enzymes [16,20,28,29] that bind membrane surfaces [30-34] and can damage or remodel them [31,35,36]. Classic bacterial GDPDs are cytoplasmic or periplasmic enzymes; as a whole GDPDs vary widely in cellular localization and membrane association [21].

Both GDPD and GDPD-like SMaseD/PLD domains fold as $(\beta/\alpha)_8$ barrels and have a similar active site [13,14], but also have several signature differences in structure (see Figure 1) [15]. First, the PLD domains conserve some version of a signature C-terminal expansion motif, which we and others have proposed to stabilize or gird the domain structure [13,37,38]. We originally dubbed it a

"plug" motif because it caps the end of the barrel opposite the active site [13]; it has also been termed a "tail" [38]. Second, GDPD enzymes have a small insertion domain (GDPD-I) [39] that is severely truncated in the PLDs and reduced to a short catalytic loop in spider toxins. The function of GDPD-I is unknown, but the corresponding catalytic loop in the toxins comprises part of the membrane interaction surface (i-face, or interfacial binding site) [34]. Third, GDPD enzymes have a longer βα1 loop that interacts with GDPD-I [15].

PLD toxins in spiders and bacteria are single-domain enzymes, but some family members have additional modules. In our previous phylogenetic analysis [15], we identified three major subfamilies—Actinobacterial-toxin like (AT-like, including the bacterial toxins and many fungal homologs), Sicariid-toxin like (ST-like, including the recluse spider toxins and homologs from other chelicerates, corals and myriapods), and Aquatic (including ctenophore and other homologs)—that derived from a paraphyletic group of microbial homologs at the base of the tree (see Figure 1). Some of these "basal" homologs are fused to domains such as PLAT [40] or Bacterial Ig-like repeats [41] that may play roles in lipid interactions and cell surface adhesion. PLAT domains in particular, which are distantly related to C2 domains [42], occur in the domain architectures of some lipases and lipoxygenases [43-45], and PLAT and C2 domains commonly mediate membrane association [46-50].

In an extension of our previous phylogenetic study of this family[15], we show that the common ancestor of GDPD-like SMaseD/PLD enzymes was fused to a PLAT domain, and that the fusion also supplied the signature C-terminal motif of the PLD domain. Remarkably, the motif derives specifically from a segment of a second PLAT domain, indicating that the fusion event involved

a PLAT repeat protein and formally occurred inside the boundaries of a PLAT domain. The structure of the motif in spider toxins differs from the structure adopted by homologous PLAT segments. The PLAT repeat fusion likely both aided in membrane association of the nascent lipase and also stabilized its catalytic domain. A broader insight is that heterologous recombination can leave scraps of domains in the sequences of the resulting proteins; these scraps can be salvaged and repurposed, in this case to build an addition onto an existing domain.

Results and Discussion

Updates of our previous GDPD-like SMaseD/PLD database reinforce the conclusion [15] that GDPD-like SMaseD/PLD domains originally evolved from GDPD domains in bacteria and radiated widely through lateral gene transfer. A phylogenetic tree shows an expanded basal group of microbial sequences from an extremely diverse array of bacteria along with a few eukaryotes (fungi and diatoms) (Figure 1; Tables S1 and S2). There is little correlation between taxonomic assignment and the tree structure of the basal group, pointing to extensive ancient gene transfer between microbes. The three major derived subfamilies (AT-like, ST-like and Aquatic) were each consistently recovered as monophyletic independent of our sequence alignment approach, though placement of the AT-like group was highly unstable.

The early GDPD-like SMaseD/PLD proteins probably had multiple domains that aided these newly evolved lipases in adhesion to membranes and other molecules on cell surfaces. In the expanded database, 35 out of 45 of the basal group are multidomain proteins. In addition to the PLD domain, we detected the following domains directly via batch searches of the Conserved Domain Database: PLAT (cI00011), PUD1_2 (pfam18457), Pectate_lyase_3 (cI40625),

DUF5011 (HYR) (cI03620) and Ricin (cI23784) (Table S3). Using FFAS03, we further identified necrosis-inducing protein (pfam05630), thiol-activated cytolysin beta-sandwich (cI38715), Ig_like_Ice (cI14404), and carbohydrate-binding module family 6 (cI14880) domains (Table S4). By far the most common domain fusion observed involves PLAT domains, occurring in 22 members of the basal group.

The ancestral GDPD-like SMaseD/PLD was likely fused to a single PLAT domain and lacked a secretion signal. In a tree of the basal group (Figure 2A), the most recent ancestral node common to all PLAT-bearing PLDs is the root of the entire basal group. A reasonable evolutionary trace can be drawn (Figure 2A) involving an ancestral PLAT fusion that underwent several losses, several duplications, and one putative permutation moving the PLAT domain from C terminus to N terminus or *vice versa*. It is also conceivable that some of the PLD-PLAT fusions arose convergently, but trees of the PLAT-bearing basal group derived from PLAT sequence appear roughly congruent with those derived from the cognate PLD sequences (Figure 2C and 2B). The PLAT-derived tree also suggests that some if not all of the PLAT repeats observed in basal PLDs derive from PLAT duplications. These observations, coupled with the reasonableness of a membrane association domain in a newly evolved lipase, support a working hypothesis that fusion to a single PLAT domain was a foundational event in the evolutionary origin of the GDPD-like SMaseD/PLD family. Interestingly, although 10 of 45 basal family members bear N-terminal signal sequences as analyzed by SignalP 6.0 [51], none of the PLD-PLAT fusions do, except for two that have highly divergent or degraded PLAT domains (Figure 2A). If the first GDPD-like SMaseD/PLD was indeed fused to a PLAT domain, it may have been a cytoplasmic protein.

It then occurred to us that fusion of a PLAT-containing protein to the C terminus of a GDPD domain might have simultaneously supplied the signature C-terminal extension motif of GDPD-like SMaseD/PLD domains (see Figure 1). We wondered whether the PLD C-terminal motif might show recognizable similarity to regions flanking PLAT domains in known bacterial PLAT-containing proteins. BLAST searches initiated from a PLD-PLAT fusion in *Microcoleus asticus*, which lies closest to the root of the basal tree (see Figure 2), hit both single-domain and multidomain PLAT-containing proteins, including some containing PLAT repeats. Interestingly, BLAST alignments to several proteins with PLAT repeats not only spanned single PLAT domains but also extended across domain boundaries to align part of a second PLAT repeat to the PLD C-terminal motif (Figure 3). Moreover, the N-terminal boundaries of the BLAST alignments coincided approximately with the C-terminal boundary of typical bacterial GDPD domains (Figure 3). The C-terminal motif of GDPD-like SMaseD/PLD domains could thus plausibly derive from a PLAT domain segment that arrived *via* fusion of part of a PLAT repeat protein to a GDPD domain.

The C-terminal motif from the basal *Microcoleus* PLD domain is similar both to other PLD motifs and PLAT segments and is intermediate in sequence between them, consistent with an evolutionary link (Figure 4). It aligns well to C-terminal regions of bacterial PLAT domains used as outgroups in rooting the PLD-PLAT tree (Figure 4A), and to the C-terminal motifs of other basal GDPD-like SMaseD/PLD domains, which are similar to those in the derived toxins (Figure 4B). Sequence clustering at ≥40% identity also groups the bacterial PLAT fragments and most basal PLD motifs into two clusters, which are unified by multiple transitive similarities,

primarily involving the *Microcoleus* sequence (Figure 4C); qualitatively similar results (not shown) are obtained by clustering with CLANS[52] using matrix scoring.

For sequences this short, with ~40% sequence identity, it is challenging to establish firm statistical support for homology using the traditional benchmarks of conditional E-values based on searches of the entire nonredundant protein database. Instead, following an approach similar to that used in other recent studies of fragment sharing[53,54], we determined database-independent E-values (and approximate P-values) from pairwise alignments to evaluate support relative to the simplest alternative hypothesis of a random origin by piecemeal growth of the reading frame at the C terminus. Statistical support for pairwise alignments between the *Microcoleus* PLD motif and the other basal PLD motifs or bacterial PLAT C termini is significant for 15 of 39 PLAT sequences and 15 of 44 PLD motifs tested individually (Table S5; by comparison to scores from alignments of shuffled sequences, with normalized E-values < 0.01, corresponding approximately to $P < 0.01$).[55-57] For the PLAT C termini, the single best observed normalized E-value is 5e-04, which corresponds to an E-value of .015 for a database size of 39; overall we see 5 hits with database-adjusted E-values < 0.1. For the basal PLD motifs, the best normalized E-value is 7.5e-06, and the corresponding database E-value is 3e-04; overall we see 9 hits with database-adjusted E-values < 0.1. These statistics support the hypothesis that the *Microcoleus* PLD C-terminal motif derived from a PLAT fragment rather than from piecemeal growth, and that it is also homologous to the other basal PLD C-terminal motifs.

The case is further supported by independent evidence that a PLAT-repeat protein was the likeliest original source of the fragment. A C-terminal GDPD-PLAT fusion is a plausible

foundational event in the family (Figures 2 and 3), and the commonality of PLAT repeat-containing proteins [58] makes them good candidates for the PLAT donor. In fact, in a conserved domain search, 23 of 32 PLAT domains identified in basal GDPD-like SMaseD/PLDs belong to the PLAT_repeat family (cd01756) within the PLAT superfamily (see Table S3). We conclude that the signature C-terminal motif of GDPD-like SMaseD/PLD was repurposed from a scrap of a PLAT domain within a PLAT repeat, and used to build an addition onto the GDPD $(\beta/\alpha)_8$ barrel. The PLD domains, including extant toxins, are thus chimeras constructed from a full GDPD domain and part of a PLAT domain.

The repurposed PLAT fragment in the PLD C-terminal motif has been structurally remodeled (Figure 4D). We clustered PLAT fragments and PLD motifs from known structures with the other sequences analyzed in Figure 4C, and thereby found structures from each group that were similar enough in sequence to incorporate into our multiple alignments (see Figures 4A and 4B). We thus mapped the PLAT fragment to the two terminal β-strands (β7 and β8) and connecting loop from a domain in *Mus musculus* Rab6-IP1 (Figure 4A and 4D; see also Supplementary Figure S1B and Materials and Methods) [59]. Meanwhile, we readily mapped the PLD motif to the structures of spider PLD toxins (Figure 4B and 4D; see also Supplementary Figure S1A). The motif begins with an α-helix, then moves into a β-strand followed by a meandering but ordered loop (Figure 4B and Figure 4D). Given this extensive structural remodeling, it is not surprising that the sequence conservation patterns observed in basal PLD C-terminal motifs are significantly different from those of the related PLAT domain fragments (compare Figures 4A and 4B). The determinants of the structural change are unclear at present and might include

changes in the sequence itself, including an apparent deletion (see Figures 3 and 4A), as well as changes in the structural context.

The C-terminal extension motif of PLD toxins has been proposed to act as cap or plug that stabilizes the core $(\beta/\alpha)_8$ barrel.[13,37,38] To test this idea, we deleted the motif from a reconstructed Sicariid venom toxin ancestor with very high thermostability as well as enzymatic activity (Figure 5; see also Materials and Methods). We reasoned that the choice of background sequence for the deletion was somewhat arbitrary, and that a putative common venom toxin ancestor represented a generic choice; in addition, its high stability increased the chance of soluble recombinant expression and foldability of the truncated variant, which we anticipated might be dramatically destabilized. Removal of the motif did in fact lower the apparent thermal denaturation midpoint from 72 °C to 64 °C (Figure 5B). Unexpectedly, the truncated variant also showed drastically reduced activity (two orders of magnitude) toward both sphingomyelin and lysophosphatidylcholine substrates (Figure 5C). The C-terminal extension motif indeed appears to support structural stability, but may also play an important role in maintaining a catalytically active conformation.

Was the structural stabilization of any importance in the early evolution of this family? One possibility is that stabilization protected the enzyme against proteolytic degradation or oxidative damage in a hostile extracellular environment. As noted above, however, the basal PLD-PLAT fusions lack secretion signals (Figure 2A). Alternatively, expansion of the domain might have acted to compensate (potentially as a permissive mutation) for severe truncation of the GDPD-I domain (Figure 1). Although GDPD-I is technically regarded as a separate domain, its interface

with the main $(\beta/\alpha)_8$ barrel domain is extensive and may contribute to the overall folding stability. Truncation of GDPD-I would for example eliminate a hydrophobic interface with the βα1 loop (Figure 1), along with the surface buried within its own hydrophobic core.

Mixing and joining of short, heterologous polypeptide segments played a major role in the early evolution of protein domains and continued to happen. Domains big enough to fold independently were formed through duplication, fusion and recombination of smaller polypeptide modules, of the general size of the PLAT subdomain identified here [8,9]. Once a large enough domain repertoire existed, domains became the primary modules of proteins, and have been shuffled between proteins as whole units to yield diverse domain architectures [2,8]. Nonetheless, recombination at the subdomain level may continue to occur in the background. Similar segments are found within domains of very different structure and it is speculated that some may have been grafted from other domains[60]. Design and selection studies show that heterologous recombination has particularly high potential to yield novel chimeric domains [4-6] and this principle is central to several current protein design approaches including SEWING [61] and Protlego [62 63]. In the thematic variation described here, natural evolution has used a subdomain segment to build an important *addition* onto a completely different domain. Recombinations that result in structural *expansion* rather than *substitution* may be less disruptive and more likely to occur in evolution.

Segments exchanged between unrelated domains may undergo remodeling in surprising ways. The structure adopted by a polypeptide sequence is highly dependent on its context [64-66]. For example, in a chimeric CspA/S1 domain constructed by De Bono *et al.* [7], the CspA fragment

retained its structure, but the S1 fragment structure was significantly changed. Similarly, the PLAT domain segment studied here has diverged in structure in PLAT and GDPD-like SMaseD/PLD contexts, though it is not clear what precise combination of contextual change and sequence change led to this divergence. Remodeling and repurposing protein domain segments is akin to fashioning steel drums from oil barrels, in which the original piece of scrap metal is hammered into a new shape to fit its new context and purpose.

Materials and methods

Materials. Phospholipid substrates 14:0 lysophosphatidylcholine (LPC; 1-myristoyl-2-hydroxy-*sn*-glycero-3-phosphocholine) and sphingomyelin (SM; brain, porcine) were purchased from Avanti Polar Lipids (Alabaster, AL). QuikChange site-directed mutagenesis kit components were purchased from Agilent (Santa Clara, CA). QIAprep Spin Miniprep kits (250) and nickel-nitrilotriacetic acid resin were purchased from QIAgen (Hilden, Germany). BugBuster Protein Extraction Reagent and Benzonase nuclease were purchased from EMD Millipore (Burlington, MA). Amplex Red Sphingomyelinase Assay Kits were purchased from Invitrogen (Eugene, OR). All other reagents were obtained from standard sources.

Sequence database

A previously reported database [15] of 7 basal microbial sequences in the GDPD-like SMaseD/PLD family was expanded to 45 sequences using protein BLAST (blastp) searches of the NCBI nonredundant database, initiated from multiple diverse query sequences. BLAST Hits with E < 1e-10 and > 90% complete PLD domains were retained, pooled and filtered at 90% sequence identity using CD-HIT [67,68]. Subsequent to the primary analysis here, we verified that

deeper searches using PSI-BLAST did not substantially expand the basal group. To probe for homologous but unannotated eukaryotic genes, translated BLAST searches (tblastn) of NCBI whole genome shotgun databases were also conducted. These searches yielded intriguing hits, including two metazoan sequences putatively from arthropod genome assemblies. However, the new genome hits were not used in the current analysis due to uncertainties in gene modeling as well as the potential for contamination. A table of all sequences retained for analysis is found in Table S1. In addition to updating the basal group, the three major derived subfamilies (ST-like, Aquatic, and AT-like) were also updated in order to briefly confirm the placement of these subfamilies relative to the basal group (see below and Figure 1). These updates included new identifications of metazoan and microbial organisms that carry members of the ST-like and Aquatic groups, but these are not germane to the current study and will be described in detail elsewhere. Standard database sequence identifiers were tagged with a prefix consisting of a shorthand for the genus and species names ('Genu_sp') of the source organism. A taxonomic key for these shorthand names is found in Table S2, including information about the environmental sources of these species, where known. Domains were initially identified using Batch CD-Search of the full NCBI Conserved Domain Database at an E-value cutoff of 1e-03 [69] (Table S3), supplemented by FFAS03 [70] and PSI-BLAST searches that in certain cases produced clear annotations of domains missed by CD-search (Table S4).

Sequence alignment and phylogenetic analysis

For an initial global alignment and phylogeny of the GDPD-like SMaseD/PLD family (Figure 1), two limiting (and one intermediate) protocols were used. One protocol was to simply align all PLD domain sequences together *de novo* using MAFFT [71]. A second protocol was to align each

of the major known groups (basal, ST-like, AT-like and Aquatic) using MAFFT followed by alignment of the group profiles using Muscle [72,73]. An intermediate protocol was also employed in which pairs of groups were aligned *de novo* using MAFFT and these profiles aligned to each other using Muscle. Phylogenetic trees were then estimated using maximum likelihood in IQ-Tree [74,75], using a WAG+F+R7 model as selected by the model selection routine in IQ-Tree. Three independent runs were performed for each of three alignments, using a slow, thorough NNI search protocol. Ultrafast bootstrapping was performed with 1000-2000 replicates and with the -bnni option turned on to reduce overestimation of support. Regardless of alignment protocol, the three major derived subfamilies were always recovered as monophyletic. Placement of the ST-like and Aquatic groups in the tree was quite consistent across alignments and runs (see Figure 1), while placement of the AT-like group was unstable.

For analyses specifically focused on the basal group, and for the PLD-PLAT fusions within the basal group, best maximum likelihood trees were obtained using an LG+G4 model under IQ-Tree, based either on alignments of the PLD domain, or on alignments of the PLAT domains. These trees were subjected to rooting using outgroups identified and aligned as described below, with no restriction on ingroup topology. For the PLD domains, outgroups were identified and aligned using an approach similar to that described previously [15]. Specifically, six bacterial GDPD sequences of known structure were selected as outgroups, structurally aligned to a spider toxin structure (PDB ID 4Q6X) using Chimera, and the resulting alignment profile aligned to the ingroup profile using Muscle. Because certain portions of the sequence and structure match poorly between GDPD and GDPD-like SMaseD/PLD domains, the resulting profile-profile alignment was then reduced to retain only the blocks of sequence that matched well for all

sequences in the structural alignment (128 residues in total, about half the domain). The root position of the PLAT domain tree was estimated using outgroups derived from BLAST searches of bacterial proteins using all ingroup PLAT domain sequences as queries. Several outgroups were tried: 1) the two best nonredundant (90% ID) BLAST hits overall, which also showed E < 1e-20, 2) the six BLAST hits that gave E < 1e-15 to queries from at least two species (in this case *Ignavibacter* and *Microcoleus*), 3) a set of 39 bacterial PLAT sequences identified from secondary BLAST searches using the two best BLAST hits described above as queries, followed by pooling and filtering for redundancy at 90% ID as above. All three of these outgroups aligned straightforwardly to the ingroup using MAFFT. All three placed the root on the same branch.

Pairwise sequence alignments between the *Microcoleus asticus* motif and the 39 bacterial PLAT C termini and 44 other basal PLD C-terminal motifs were performed using SSEARCH in the FASTA package and statistically analyzed using PRSS program in FASTA with 1000 uniformly shuffled sequences. [55-57] PRSS compares each alignment score to an extreme-value distribution of alignments with shuffled sequences, and generates an E-value that can be normalized for a database size of 1. In cases where E < 0.01, the P-value for the alignment should be approximately equal to this normalized E-value.

Single linkage clusters based on sequence identity were generated and graphed using the epitope cluster analysis tool at tools.iedb.org [76]. In the linkage graphs, edges were drawn between nodes with ≥40% ID over a 20-residue span. An alternative clustering approach using CLANS [52] and all-against-all BLAST, gave qualitatively similar results to those of Figure 4C, with the *Microcoleus asticus* sequence forming the sole transitive connection between the remaining PLD

motifs and the PLAT fragments (conditional E-value threshold of <0.01). The set of sequences to be clustered derived from combining and curating multiple alignments of 45 basal PLD and 44 bacterial PLAT domain fragments (see Figures 4A and 4B for samplings from these alignments). The bacterial PLAT alignment derived from the set of 39 PLAT domain sequences used above as an outgroup for generating a tree of PLD-PLAT fusion; it also included the five sequences from ingroup PLAT domains that gave the strongest PLAT hits in the original BLAST searches. Six heavily gapped columns were removed, as was the single most C-terminal column, as this residue is absent in some spider toxin structures (4Q6X). This curation reduced the number of alignment columns to 20. Sequences that still contained gaps were then also removed, reducing the number of bacterial PLAT sequences from 44 to 32, and the number of basal GDPD-like SMaseD/PLD sequences from 45 to 30.

To assess whether the PLD and PLAT sequence fragments might be confidently aligned/mapped to their respective domain structures, clustering runs were also performed that included C-terminal regions from three sicariid spider toxin structures (from PDB IDs 3RLH, 4Q6X and 1XX1) as well as PLAT domain C-terminal fragments from known structures identified using BLAST searches of the PDB with bacterial PLAT domain sequences as queries (PDB IDs 3CWZ, 2FNQ, 3VF1 and 6A70). All three toxin and three of four (all except 6A70) PLAT sequences were incorporated into the main cluster (see Figure 4C), suggesting that PLD and PLAT domain sequence fragments could both be aligned/mapped to three-dimensional structures. Superpositions are shown in Supplementary Figure S1A (PLD toxin structures and S1B (PLAT domains), to illustrate that the C-terminal motif structure is strongly conserved among the Sicariid toxin structures, and is highly remodeled from the strand-loop-strand

structures formed by the representative C-terminal PLAT fragments. The three PLAT fragment structures did show some variability in conformation and secondary structure, with 3CWZ having a longer loop sequence than the other two (Figure S1B). 3CWZ was chosen as best representative for secondary structure comparison (Figures 4A and 4B) and for visualizing the remodeling (Figure 4D), for several reasons: 1) it was the only one of the three for which BLAST alignments to members of our bacterial PLAT family had extended all the way through the C terminus of the domain, 2) in the C-terminal clustering analysis, it exhibited 12 connections to members of the bacterial PLAT family, while 2FNQ showed only two, and 3VF1 showed connections only to 3CWZ and 2FNQ.

Ancestral state reconstruction. Multiple sequence alignments were constructed using MAFFT [71] of a database of 93 venom-expressed toxin sequences, with 35 non-venom expressed sequences from scorpions, spiders and Opiliones incorporated as an outgroup. Maximum likelihood trees were constructed with IQ-TREE[74], and an ancestral sequence was predicted for a common Sicariid venom toxin ancestral node using RAxML-NG[77], under a WAG+F+R5 model.

Cloning and mutagenesis. Synthetic genes were codon optimized for expression in *E. coli* using a NovoPro codon optimization tool and ordered as gBlocks gene fragments from Integrated DNA Technologies (Coralville, IA). Genes were cloned into a pHis8 expression vector as described previously. The C-terminal truncation (Δ260) was made by introducing a stop codon in place of the codon for residue Lys260, using the QuikChange site directed mutagenesis kit according to manufacturer protocol (Agilent; Santa Clara, CA).

Protein expression and purification. N-terminally His_8-tagged recombinant proteins were expressed from pHis8-toxin constructs in *E. coli* BL21(λDE3) strains, extracted, and purified by nickel affinity chromatography closely following published methods[26,78] with the modification of an 18-20 h induction period at 20 °C. After dialysis into TBS (0.1 M Tris [pH 8.0], 0.2 M NaCl), concentration of protein was determined from absorbance at 280 nm and extinction coefficients estimated from amino acid sequence.[79]

Thermal denaturation. Thermal denaturation curves were obtained by measuring the change in circular dichroism at 222 nm in an OLIS DSM-20 CD spectropolarimeter, at 0.2 mg/mL protein in 0.1 M Tris (pH 8.0), 0.2 M NaCl and a pathlength of 1 mm. Samples were heated from 30 °C to 90 °C in 2 °C increments with a 2 min equilibration time and integration time of 55 s. The apparent melting point (T_m) was estimated by fitting to a variant of the Gibbs-Helmholtz equation[80] using an estimated heat capacity of unfolding based on sequence length[81].

Choline release assay. An Amplex Red Sphingomyelinase Assay Kit was used to indirectly detect choline release from a lipid substrate in an enzyme-coupled assay using spectrophotometric absorbance of resorufin. In addition to standard concentrations of secondary reagents (horseradish peroxidase, choline oxidase, and Amplex Red reagent), samples included 200 μM SM or LPC substrate in 0.1% Triton X-100 micelles and 0.25-5 μg enzyme in 0.1 M Tris (pH 7.4), 10 mM $MgCl_2$, 37°C. The absorbance of each sample at 572 nm was measured every 3 min for a duration of 1 h. Parent and truncated enzymes were directly compared in paired runs using common reagent and substrate solutions. Rates were derived from fitting the absorbance curves to an equation that accounts for both the primary enzymatic reaction as well

as the secondary reactions converting choline into resorufin. An apparent first-order rate constant for the secondary reaction was independently measured in a reaction between the secondary enzymes/reagents and choline. Absorbance was related to amount of product by measuring the absorbance generated in control reactions containing known concentrations of hydrogen peroxide.

SUPPLEMENTARY MATERIAL DESCRIPTION

ACKNOWLEDGMENTS

This research was supported by NSF grant CHE-1808716 (to M.H.J.C. and G.J.B.). The authors have selected an Open Access license.

REFERENCES

1. Baron M, Norman DG, Campbell ID (1991) Protein modules. Trends Biochem Sci 16:13-17.
2. Ponting CP, Russell RR (2002) The natural history of protein domains. Annu Rev Biophys Biomol Struct 31:45-71.
3. Apic G, Russell RB (2010) Domain recombination: a workhorse for evolutionary innovation. Sci Signal 3:pe30.
4. Riechmann L, Winter G (2000) Novel folded protein domains generated by combinatorial shuffling of polypeptide segments. Proc Natl Acad Sci U S A 97:10068-10073.
5. Bharat TA, Eisenbeis S, Zeth K, Hocker B (2008) A beta alpha-barrel built by the combination of fragments from different folds. Proc Natl Acad Sci U S A 105:9942-9947.
6. Hocker B (2014) Design of proteins from smaller fragments-learning from evolution. Curr Opin Struct Biol 27:56-62.
7. de Bono S, Riechmann L, Girard E, Williams RL, Winter G (2005) A segment of cold shock protein directs the folding of a combinatorial protein. Proc Natl Acad Sci U S A 102:1396-1401.
8. Lupas AN, Ponting CP, Russell RB (2001) On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? J Struct Biol 134:191-203.
9. Soding J, Lupas AN (2003) More than the sum of their parts: on the evolution of proteins from peptides. Bioessays 25:837-846.
10. Yagi S, Padhi AK, Vucinic J, Barbe S, Schiex T, Nakagawa R, Simoncini D, Zhang KYJ, Tagami S (2021) Seven Amino Acid Types Suffice to Create the Core Fold of RNA Polymerase. J Am Chem Soc 143:15998-16006.
11. Seal M, Weil-Ktorza O, Despotovic D, Tawfik DS, Levy Y, Metanis N, Longo LM, Goldfarb D (2022) Peptide-RNA Coacervates as a Cradle for the Evolution of Folded Domains. J Am Chem Soc 144:14150-14160.
12. Binford GJ, Cordes MH, Wells MA (2005) Sphingomyelinase D from venoms of Loxosceles spiders: evolutionary insights from cDNA sequences and gene structure. Toxicon 45:547-560.
13. Cordes MH, Binford GJ (2006) Lateral gene transfer of a dermonecrotic toxin between spiders and bacteria. Bioinformatics 22:264-268.
14. Murakami MT, Fernandes-Pedrosa MF, de Andrade SA, Gabdoulkhakov A, Betzel C, Tambourgi DV, Arni RK (2006) Structural insights into the catalytic mechanism of sphingomyelinases D and evolutionary relationship to glycerophosphodiester phosphodiesterases. Biochemical and biophysical research communications 342:323-329.
15. Cordes MHJ, Binford GJ (2018) Evolutionary dynamics of origin and loss in the deep history of phospholipase D toxin genes. BMC Evol Biol 18:194.
16. Tambourgi DV, Magnoli FC, van den Berg CW, Morgan BP, de Araujo PS, Alves EW, Da Silva WD (1998) Sphingomyelinases in the venom of the spider Loxosceles intermedia are responsible for both dermonecrosis and complement-dependent hemolysis. Biochemical and biophysical research communications 251:366-373.

17. Borchani L, Sassi A, Shahbazzadeh D, Strub JM, Tounsi-Guetteti H, Boubaker MS, Akbari A, Van Dorsselaer A, El Ayeb M (2011) Heminecrolysin, the first hemolytic dermonecrotic toxin purified from scorpion venom. Toxicon 58:130-139.
18. Zobel-Thropp PA, Kerins AE, Binford GJ (2012) Sphingomyelinase D in sicariid spider venom is a potent insecticidal toxin. Toxicon 60:265-271.
19. McNamara PJ, Cuevas WA, Songer JG (1995) Toxic phospholipases D of Corynebacterium pseudotuberculosis, C. ulcerans and Arcanobacterium haemolyticum: cloning and sequence homology. Gene 156:113-118.
20. Kurpiewski G, Forrester LJ, Barrett JT, Campbell BJ (1981) Platelet aggregation and sphingomyelinase D activity of a purified toxin from the venom of Loxosceles reclusa. Biochim Biophys Acta 678:467-476.
21. Corda D, Mosca MG, Ohshima N, Grauso L, Yanaka N, Mariggio S (2014) The emerging physiological roles of the glycerophosphodiesterase family. FEBS J 281:998-1016.
22. van Meeteren LA, Frederiks F, Giepmans BN, Pedrosa MF, Billington SJ, Jost BH, Tambourgi DV, Moolenaar WH (2004) Spider and bacterial sphingomyelinases D target cellular lysophosphatidic acid receptors by hydrolyzing lysophosphatidylcholine. The Journal of biological chemistry 279:10833-10836.
23. Lee S, Lynch KR (2005) Brown recluse spider (Loxosceles reclusa) venom phospholipase D (PLD) generates lysophosphatidic acid (LPA). Biochem J 391:317-323.
24. Lajoie DM, Cordes MH (2015) Spider, bacterial and fungal phospholipase D toxins make cyclic phosphate products. Toxicon 108:176-180.
25. Ohshima N, Kudo T, Yamashita Y, Mariggio S, Araki M, Honda A, Nagano T, Isaji C, Kato N, Corda D, Izumi T, Yanaka N (2015) New members of the mammalian glycerophosphodiester phosphodiesterase family: GDE4 and GDE7 produce lysophosphatidic acid by lysophospholipase D activity. The Journal of biological chemistry 290:4260-4271.
26. Lajoie DM, Zobel-Thropp PA, Kumirov VK, Bandarian V, Binford GJ, Cordes MH (2013) Phospholipase D toxins of brown spider venom convert lysophosphatidylcholine and sphingomyelin to cyclic phosphates. PloS one 8:e72372.
27. Shi L, Liu JF, An XM, Liang DC (2008) Crystal structure of glycerophosphodiester phosphodiesterase (GDPD) from Thermoanaerobacter tengcongensis, a metal ion-dependent enzyme: insight into the catalytic mechanism. Proteins 72:280-288.
28. Songer JG, Libby SJ, Iandolo JJ, Cuevas WA (1990) Cloning and expression of the phospholipase D gene from Corynebacterium pseudotuberculosis in Escherichia coli. Infect Immun 58:131-136.
29. Cuevas WA, Songer JG (1993) Arcanobacterium haemolyticum phospholipase D is genetically and functionally similar to Corynebacterium pseudotuberculosis phospholipase D. Infect Immun 61:4310-4316.
30. Rees RS, Nanney LB, Yates RA, King LE, Jr. (1984) Interaction of brown recluse spider venom on cell membranes: the inciting mechanism? J Invest Dermatol 83:270-275.
31. Chaves-Moreira D, Chaim OM, Sade YB, Paludo KS, Gremski LH, Donatti L, de Moura J, Mangili OC, Gremski W, da Silveira RB, Senff-Ribeiro A, Veiga SS (2009) Identification of a direct hemolytic effect dependent on the catalytic activity induced by phospholipase-D (dermonecrotic toxin) from brown spider venom. J Cell Biochem 107:655-666.

32. Chaim OM, da Silveira RB, Trevisan-Silva D, Ferrer VP, Sade YB, Boia-Ferreira M, Gremski LH, Gremski W, Senff-Ribeiro A, Takahashi HK, Toledo MS, Nader HB, Veiga SS (2011) Phospholipase-D activity and inflammatory response induced by brown spider dermonecrotic toxin: endothelial cell membrane phospholipids as targets for toxicity. Biochim Biophys Acta 1811:84-96.
33. Wille AC, Chaves-Moreira D, Trevisan-Silva D, Magnoni MG, Boia-Ferreira M, Gremski LH, Gremski W, Chaim OM, Senff-Ribeiro A, Veiga SS (2013) Modulation of membrane phospholipids, the cytosolic calcium influx and cell proliferation following treatment of B16-F10 cells with recombinant phospholipase-D from Loxosceles intermedia (brown spider) venom. Toxicon 67:17-30.
34. Moutoussamy EE, Waheed Q, Binford GJ, Khan HM, Moran SM, Eitel AR, Cordes MHJ, Reuter N (2022) Specificity of Loxosceles alpha clade phospholipase D enzymes for choline-containing lipids: Role of a conserved aromatic cage. PLoS Comput Biol 18:e1009871.
35. Lucas EA, Billington SJ, Carlson P, McGee DJ, Jost BH (2010) Phospholipase D promotes Arcanobacterium haemolyticum adhesion via lipid raft remodeling and host cell death following bacterial invasion. BMC microbiology 10:270.
36. Stock RP, Brewer J, Wagner K, Ramos-Cerrillo B, Duelund L, Jernshoj KD, Olsen LF, Bagatolli LA (2012) Sphingomyelinase D activity in model membranes: structural effects of in situ generation of ceramide-1-phosphate. PloS one 7:e36003.
37. Binford GJ, Bodner MR, Cordes MH, Baldwin KL, Rynerson MR, Burns SN, Zobel-Thropp PA (2009) Molecular evolution, functional variation, and proposed nomenclature of the gene family that includes sphingomyelinase D in sicariid spider venoms. Mol Biol Evol 26:547-566.
38. Dias-Lopes C, Neshich IA, Neshich G, Ortega JM, Granier C, Chavez-Olortegui C, Molina F, Felicori L (2013) Identification of new sphingomyelinases D in pathogenic fungi and other pathogenic organisms. PloS one 8:e79240.
39. Santelli E, Schwarzenbacher R, McMullan D, Biorac T, Brinen LS, Canaves JM, Cambell J, Dai X, Deacon AM, Elsliger MA, Eshagi S, Floyd R, Godzik A, Grittini C, Grzechnik SK, Jaroszewski L, Karlak C, Klock HE, Koesema E, Kovarik JS, Kreusch A, Kuhn P, Lesley SA, McPhillips TM, Miller MD, Morse A, Moy K, Ouyang J, Page R, Quijano K, Rezezadeh F, Robb A, Sims E, Spraggon G, Stevens RC, van den Bedem H, Velasquez J, Vincent J, von Delft F, Wang X, West B, Wolf G, Xu Q, Hodgson KO, Wooley J, Wilson IA (2004) Crystal structure of a glycerophosphodiester phosphodiesterase (GDPD) from Thermotoga maritima (TM1621) at 1.60 A resolution. Proteins 56:167-170.
40. Bateman A, Sandford R (1999) The PLAT domain: a new piece in the PKD1 puzzle. Curr Biol 9:R588-590.
41. Chatterjee S, Basak AJ, Nair AV, Duraivelan K, Samanta D (2021) Immunoglobulin-fold containing bacterial adhesins: molecular and structural perspectives in host tissue colonization and infection. FEMS Microbiol Lett 368.
42. Serrano P, Geralt M, Mohanty B, Wuthrich K (2013) Structural representative of the protein family PF14466 has a new fold and establishes links with the C2 and PLAT domains from the widely distant Pfams PF00168 and PF01477. Protein Sci 22:1000-1007.

43. van Tilbeurgh H, Egloff MP, Martinez C, Rugani N, Verger R, Cambillau C (1993) Interfacial activation of the lipase-procolipase complex by mixed micelles revealed by X-ray crystallography. Nature 362:814-820.
44. Gillmor SA, Villasenor A, Fletterick R, Sigal E, Browner MF (1997) The structure of mammalian 15-lipoxygenase reveals similarity to the lipases and the determinants of substrate specificity. Nat Struct Biol 4:1003-1009.
45. Naylor CE, Eaton JT, Howells A, Justin N, Moss DS, Titball RW, Basak AK (1998) Structure of the key toxin in gas gangrene. Nat Struct Biol 5:738-746.
46. Kulkarni S, Das S, Funk CD, Murray D, Cho W (2002) Molecular basis of the specific subcellular localization of the C2-like domain of 5-lipoxygenase. The Journal of biological chemistry 277:13167-13174.
47. Walther M, Wiesner R, Kuhn H (2004) Investigations into calcium-dependent membrane association of 15-lipoxygenase-1. Mechanistic roles of surface-exposed hydrophobic amino acids and calcium. J Biol Chem 279:3717-3725.
48. Oldham ML, Brash AR, Newcomer ME (2005) Insights from the X-ray crystal structure of coral 8R-lipoxygenase: calcium activation via a C2-like domain and a structural basis of product chirality. J Biol Chem 280:39545-39552.
49. Xu Y, Streets AJ, Hounslow AM, Tran U, Jean-Alphonse F, Needham AJ, Vilardaga JP, Wessely O, Williamson MP, Ong AC (2016) The Polycystin-1, Lipoxygenase, and alpha-Toxin Domain Regulates Polycystin-1 Trafficking. J Am Soc Nephrol 27:1159-1173.
50. Guillouard I, Alzari PM, Saliou B, Cole ST (1997) The carboxy-terminal C2-like domain of the alpha-toxin from Clostridium perfringens mediates calcium-dependent membrane recognition. Mol Microbiol 26:867-876.
51. Teufel F, Almagro Armenteros JJ, Johansen AR, Gislason MH, Pihl SI, Tsirigos KD, Winther O, Brunak S, von Heijne G, Nielsen H (2022) SignalP 6.0 predicts all five types of signal peptides using protein language models. Nat Biotechnol.
52. Frickey T, Lupas A (2004) CLANS: a Java application for visualizing protein families based on pairwise similarity. Bioinformatics 20:3702-3704.
53. Longo LM, Jablonska J, Vyas P, Kanade M, Kolodny R, Ben-Tal N, Tawfik DS (2020) On the emergence of P-Loop NTPase and Rossmann enzymes from a Beta-Alpha-Beta ancestral fragment. Elife 9.
54. Longo LM, Kolodny R, McGlynn SE (2022) Evidence for the emergence of beta-trefoils by 'Peptide Budding' from an IgG-like beta-sandwich. PLoS Comput Biol 18:e1009833.
55. Pearson WR (1996) Effective protein sequence comparison. Methods Enzymol 266:227-258.
56. - (1998) Empirical statistical estimates for sequence similarity searches. J Mol Biol 276:71-84.
57. - (2000) Flexible sequence similarity searching with the FASTA3 program package. Methods Mol Biol 132:185-219.
58. Grillet N, Schwander M, Hildebrand MS, Sczaniecka A, Kolatkar A, Velasco J, Webster JA, Kahrizi K, Najmabadi H, Kimberling WJ, Stephan D, Bahlo M, Wiltshire T, Tarantino LM, Kuhn P, Smith RJ, Muller U (2009) Mutations in LOXHD1, an evolutionarily conserved stereociliary protein, disrupt hair cell function in mice and cause progressive hearing loss in humans. Am J Hum Genet 85:328-337.

59. Recacha R, Boulet A, Jollivet F, Monier S, Houdusse A, Goud B, Khan AR (2009) Structural basis for recruitment of Rab6-interacting protein 1 to Golgi via a RUN domain. Structure 17:21-30.
60. Kolodny R, Nepomnyachiy S, Tawfik DS, Ben-Tal N (2021) Bridging Themes: Short Protein Segments Found in Different Architectures. Mol Biol Evol 38:2191-2208.
61. Jacobs TM, Williams B, Williams T, Xu X, Eletsky A, Federizon JF, Szyperski T, Kuhlman B (2016) Design of structurally distinct proteins using strategies inspired by evolution. Science 352:687-690.
62. Ferruz N, Noske J, Hocker B (2021) Protlego: A Python package for the analysis and design of chimeric proteins. Bioinformatics.
63. Ferruz N, Lobos F, Lemm D, Toledo-Patino S, Farias-Rico JA, Schmidt S, Hocker B (2020) Identification and Analysis of Natural Building Blocks for Evolution-Guided Fragment-Based Protein Design. J Mol Biol 432:3898-3914.
64. Minor DL, Jr., Kim PS (1996) Context-dependent secondary structure formation of a designed protein sequence. Nature 380:730-734.
65. Schellenberg MJ, Ritchie DB, Wu T, Markin CJ, Spyracopoulos L, MacMillan AM (2010) Context-dependent remodeling of structure in two large protein fragments. J Mol Biol 402:720-730.
66. Kumirov VK, Dykstra EM, Hall BM, Anderson WJ, Szyszka TN, Cordes MHJ (2018) Multistep mutational transformation of a protein fold through structural intermediates. Protein Sci 27:1767-1779.
67. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658-1659.
68. Huang Y, Niu B, Gao Y, Fu L, Li W (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics 26:680-682.
69. Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Lu F, Marchler GH, Mullokandov M, Omelchenko MV, Robertson CL, Song JS, Thanki N, Yamashita RA, Zhang D, Zhang N, Zheng C, Bryant SH (2011) CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res 39:D225-229.
70. Jaroszewski L, Rychlewski L, Li Z, Li W, Godzik A (2005) FFAS03: a server for profile--profile sequence alignments. Nucleic Acids Res 33:W284-288.
71. Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30:3059-3066.
72. Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113.
73. - (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792-1797.
74. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32:268-274.
75. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R (2020) IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol 37:1530-1534.

76. Dhanda SK, Vaughan K, Schulten V, Grifoni A, Weiskopf D, Sidney J, Peters B, Sette A (2018) Development of a novel clustering tool for linear peptide sequences. Immunology 155:331-345.
77. Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A (2019) RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics 35:4453-4455.
78. Lajoie DM, Roberts SA, Zobel-Thropp PA, Delahaye JL, Bandarian V, Binford GJ, Cordes MH (2015) Variable Substrate Preference Among Phospholipase D Toxins From Sicariid Spiders. The Journal of biological chemistry 290:10994-11007.
79. Pace CN, Vajdos F, Fee L, Grimsley G, Gray T (1995) How to measure and predict the molar absorption coefficient of a protein. Protein Sci 4:2411-2423.
80. Becktel WJ, Schellman JA (1987) Protein stability curves. Biopolymers 26:1859-1877.
81. Myers JK, Pace CN, Scholtz JM (1995) Denaturant m values and heat capacity changes: relation to changes in accessible surface areas of protein unfolding. Protein Sci 4:2138-2148.

FIGURE LEGENDS

Figure 1. The GDPD-like SMaseD/PLD family. This family is distinguished in sequence and structure from its GDPD ancestors by truncation of the GDPD-I domain and βα1 loop (cyan and green, respectively) and expansion of the domain at the C terminus (orange). Three major subfamilies of GDPD-like SMaseD/PLD diverged from a sparse basal group of microbial sequences that originally evolved from the GDPD family in bacteria. The color-coded taxonomic distribution among bacterial groups and microbial eukaryotes indicates extensive lateral gene transfer, with the caveat that sequences from metagenome assembled genomes (indicated by _ba_) may be of less certain taxonomic assignment. Sequence alignments for each group were generated using MAFFT, aligned to each other using profile alignment in Muscle, and a maximum-likelihood tree generated using IQ-Tree under a WAG+F+R7 model (see Materials and Methods). The tree was rooted using 6 bacterial GDPD sequences of known structure as described in reference 15 (see also Materials and Methods). Most features of the tree, including recovery of the major subfamilies as monophyletic, are robust to alignment method, but the position of the highly diverged Actinobacterial-toxin like group is highly unstable.

Figure 2. The ancestral GDPD-like SMaseD/PLD was likely fused to a PLAT domain and lacked a secretion signal. (A) Overall phylogeny of basal group PLD domains, with those fused to PLAT domains indicated by black bold typeface. The *Microcoleus* sequence central to our analysis is indicated with an arrow. Colored tracing illustrates the hypothesis of an ancestral PLD-PLAT fusion that persisted in about half of the basal group and underwent multiple duplications (cyan), permutation plus duplication (green), divergence/degradation (red), and loss (grey). The two putatively divergent/degraded PLAT domains (red) were not detected in a

Conserved Domain Database search but were obvious in multiple sequence alignments. Sequences with a recognizable signal sequence as identified by SignalP 6.0 are marked with asterisks (*). Note that sequences with clear PLAT domains always lack signal sequences and that there is some correlation between PLAT loss/degradation/divergence and signal sequence acquisition, (B) Phylogeny of PLD domains from basal PLD-PLAT fusions, (C) Phylogeny of PLAT domains from basal PLD-PLAT fusions. Of a minimum of four duplications necessary to generate the observed PLAT repeats (two in the Igna_ba proteins, one each in Micr_as and Meth_YR), two are clearly evident in the tree (black circles). Taxa and clades in panels B and C are color-coded to highlight the approximate congruence of the two phylogenies. All trees shown here were best ML trees under LG+G4 models selected using IQ-Tree. PLD trees were rooted using GDPD domains of known structure as described in reference 15. The PLAT tree was rooted using bacterial PLAT homologs recovered from exhaustive BLAST searches (see Materials and Methods).

Figure 3. The C-terminal extension of GDPD-like SMaseD/PLD domains derives from fusion of a PLAT repeat fragment to a GDPD domain. This multiple alignment focuses on the region immediately on either side of the PLD domain C terminus. In BLAST searches with a PLD-PLAT fusion from *Microcoleus asticus* (cyan), several strong hits (E < 1e-10) were found to bacterial proteins containing PLAT repeats (gold). The local sequence alignments for these hits are approximated here as a multiple sequence alignment realigned using Muscle, with the PLAT repeat hits truncated to show the consensus local alignment boundary. These alignments included an entire PLAT repeat (far right; with the alignment truncated) but also extended toward the N terminus across the domain boundary, aligning the last 20+ residues of another

PLAT repeat to the C-terminal region of the *Microcoleus* PLD domain. This region of possible homology between the PLD and PLAT domains contains the previously described signature C-terminal plug motif (boxed) that is absent in bacterial GDPD ancestors of known structure (green), but conserved among GDPD-like SMaseD/PLD, including basal/ancestral (cyan) but originally described in recluse spider toxins (orange). The basal PLDs putatively originated from fusion of ~1.2 PLAT repeats (gold, second repeat truncated) to a GDPD domain (green), with incorporation of a PLAT fragment as a C-terminal PLD domain extension, along with a full PLAT domain (cyan). In toxins (orange), the full PLAT domain has been lost, but a vestige of the PLAT fragment remains.

Figure 4. Repurposing and remodeling of a PLAT fragment as a C-terminal PLD motif. (A) Alignment of *Microcoleus asticus* PLD domain C terminus (cyan label) with C termini of bacterial PLAT domains (gold labels) as well as the closest PLAT homologs of known structure (box) annotated with secondary structure (h=helix, e=strand; assignments come from 3CWZ, but a general strand-loop-strand conformation is seen in all three; see Figure S1B). The region of alignment is expanded slightly relative to the BLAST alignments shown in Figure 3. (B) Alignment of *Microcoleus asticus* PLD domain C terminus with other basal PLD C termini and with C termini from spider toxins of known structure, annotated with secondary structure (box; ; assignments come from 3RLH, but all three conformations are highly similar; see Figure S1A). In both panels A and B, only the 12 PLAT or PLD sequence fragments with highest sequence identity are shown, in order of decreasing percent identity. (C) Sequence identity clustering of basal GDPD-like SMaseD/PLD and bacterial PLAT sequence fragments, as well as sequences with known structure shown in panels A and B. Edges are drawn between nodes with ≥40% ID

over 20 residues, after removal of the heavily gapped columns shown in panels A and B, and removal of any sequences containing additional gaps. Six of 30 gapless basal PLD sequences are singletons at this cutoff (not shown), but all other sequences are unified. Note the central position of the *Microcoleus* PLD motif as an intermediate sequence (transitive connection) that is similar to multiple PLD- and PLAT-derived segments. (D) Mapping of the homologous PLAT (gold) and PLD (cyan) subdomains onto known structures. Bacterial GDPD domains (represented here by PDB ID 2PZ0) lack a C-terminal extension on the barrel. The GDPD-like SMaseD/PLD domain family has acquired a C-terminal extension, which in spider toxins (PDB ID 3RLH) has a conserved structure consisting of a helix, strand and complex ordered loop. The extension is a remodeled, repurposed fragment of a PLAT domain (PDB ID 3CWZ) consisting of strands β7 and β8.

Figure 5. Truncation of the C-terminal extension decreases both thermostability and catalytic activity. (A) sequence of reconstructed Sicariid venom ancestral PLD domain and C-terminally truncated variant (Δ260). (B) thermal denaturation monitored by change in circular dichroism at 222 nm, at 0.2 mg/mL protein and a pathlength of 1 mm, in 0.1 M Tris (pH 8.0), 0.2 M sodium chloride. (B) Choline head-group release assay of enzymatic activity using sphingomyelin (SM) or lysophosphatidylcholine (LPC) as substrate (200 μM) in mixed micelles with 0.1% Triton X-100, in reaction buffer (0.1 M Tris [pH 7.4], 10 mM $MgCl_2$, 37 °C. Error bars represent the standard error of the mean of three measurements.

Figure 1

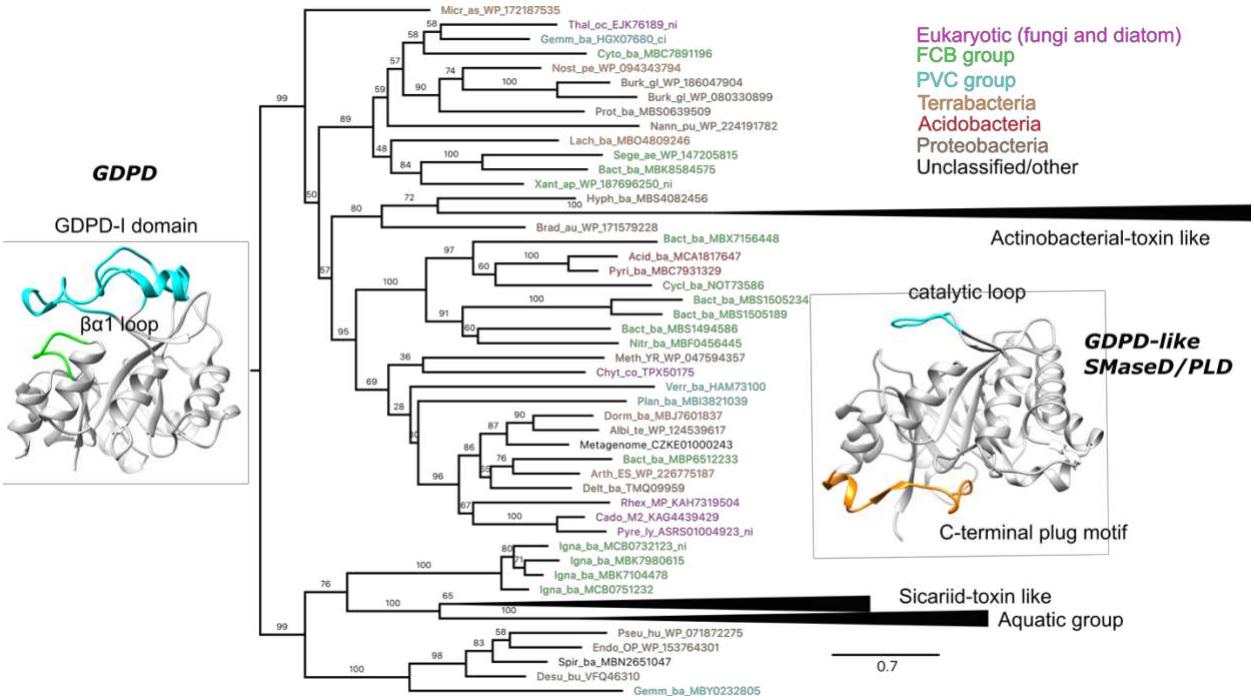

Figure 1. The GDPD-like SMaseD/PLD family. This family is distinguished in sequence and structure from its GDPD ancestors by truncation of the GDPD-I domain and βα1 loop (cyan and green, respectively) and expansion of the domain at the C terminus (orange). Three major subfamilies of GDPD-like SMaseD/PLD diverged from a sparse basal group of microbial sequences that originally evolved from the GDPD family in bacteria. The color-coded taxonomic distribution among bacterial groups and microbial eukaryotes indicates extensive lateral gene transfer, with the caveat that sequences from metagenome assembled genomes (indicated by _ba_) may be of less certain taxonomic assignment. Sequence alignments for each group were generated using MAFFT, aligned to each other using profile alignment in Muscle, and a maximum-likelihood tree generated using IQ-Tree under a WAG+F+R7 model (see Materials and Methods). The tree was rooted using 6 bacterial GDPD sequences of known structure as described in reference 15 (see also Materials and Methods). Most features of the tree, including recovery of the major subfamilies as monophyletic, are robust to alignment method, but the position of the highly diverged Actinobacterial-toxin like group is highly unstable.

Figure 2

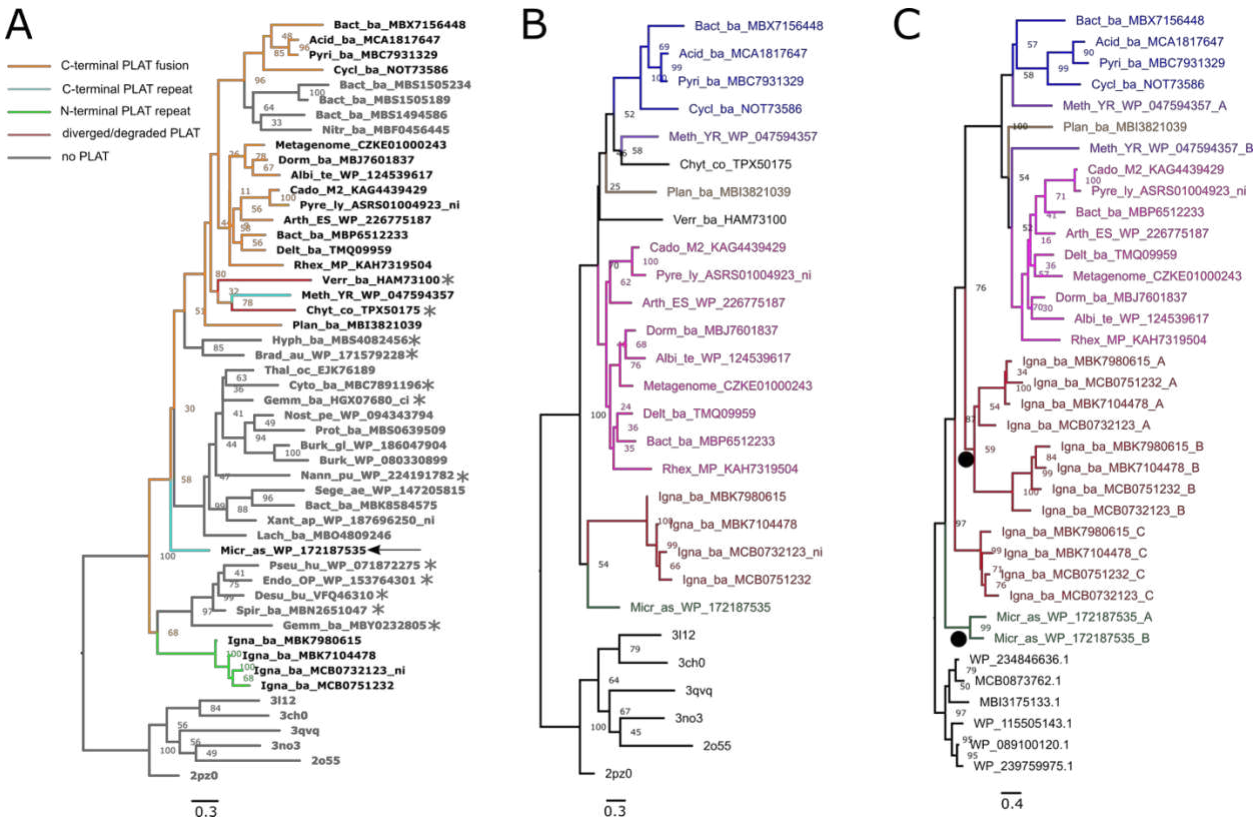

Figure 2. The ancestral GDPD-like SMaseD/PLD was likely fused to a PLAT domain and lacked a secretion signal. (A) Overall phylogeny of basal group PLD domains, with those fused to PLAT domains indicated by black bold typeface. The *Microcoleus* sequence central to our analysis is indicated with an arrow. Colored tracing illustrates the hypothesis of an ancestral PLD-PLAT fusion that persisted in about half of the basal group and underwent multiple duplications (cyan), permutation plus duplication (green), divergence/degradation (red), and loss (grey). The two putatively divergent/degraded PLAT domains (red) were not detected in a Conserved Domain Database search but were obvious in multiple sequence alignments. Sequences with a recognizable signal sequence as identified by SignalP 6.0 are marked with asterisks (*). Note that sequences with clear PLAT domains always lack signal sequences and that there is some correlation between PLAT loss/degradation/divergence and signal sequence acquisition, (B) Phylogeny of PLD domains from basal PLD-PLAT fusions, (C) Phylogeny of PLAT domains from basal PLD-PLAT fusions. Of a minimum of four duplications necessary to generate the observed PLAT repeats (two in the Igna_ba proteins, one each in Micr_as and Meth_YR), two are clearly evident in the tree (black circles). Taxa and clades in panels B and C are color-coded to highlight the approximate congruence of the two phylogenies. All trees shown here were best ML trees under LG+G4 models selected using IQ-Tree. PLD trees were rooted using GDPD domains of known structure as described in reference 15. The PLAT tree was rooted using bacterial PLAT homologs recovered from exhaustive BLAST searches (see Materials and Methods).

Figure 3

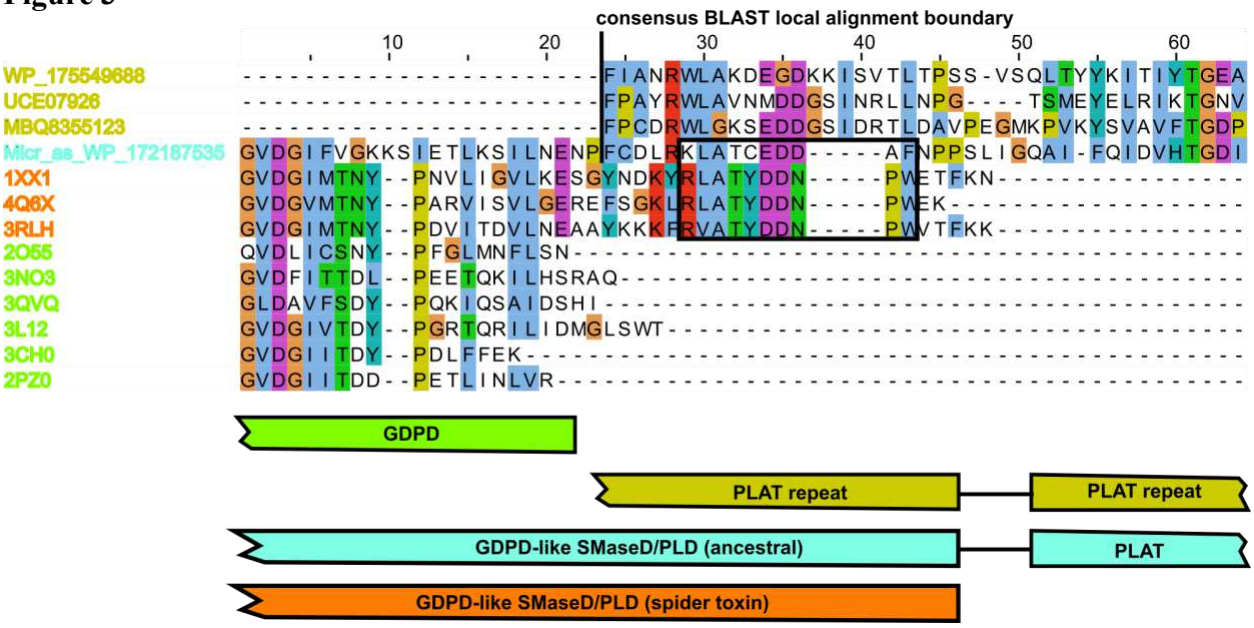

Figure 3. The C-terminal extension of GDPD-like SMaseD/PLD domains derives from fusion of a PLAT repeat fragment to a GDPD domain. This multiple alignment focuses on the region immediately on either side of the PLD domain C terminus. In BLAST searches with a PLD-PLAT fusion from *Microcoleus asticus* (cyan), several strong hits (E < 1e-10) were found to bacterial proteins containing PLAT repeats (gold). The local sequence alignments for these hits are approximated here as a multiple sequence alignment realigned using Muscle, with the PLAT repeat hits truncated to show the consensus local alignment boundary. These alignments included an entire PLAT repeat (far right; with the alignment truncated) but also extended toward the N terminus across the domain boundary, aligning the last 20+ residues of another PLAT repeat to the C-terminal region of the *Microcoleus* PLD domain. This region of possible homology between the PLD and PLAT domains contains the previously described signature C-terminal plug motif (boxed) that is absent in bacterial GDPD ancestors of known structure (green), but conserved among GDPD-like SMaseD/PLD, including basal/ancestral (cyan) but originally described in recluse spider toxins (orange). The basal PLDs putatively originated from fusion of ~1.2 PLAT repeats (gold, second repeat truncated) to a GDPD domain (green), with incorporation of a PLAT fragment as a C-terminal PLD domain extension, along with a full PLAT domain (cyan). In toxins (orange), the full PLAT domain has been lost, but a vestige of the PLAT fragment remains.

Figure 4

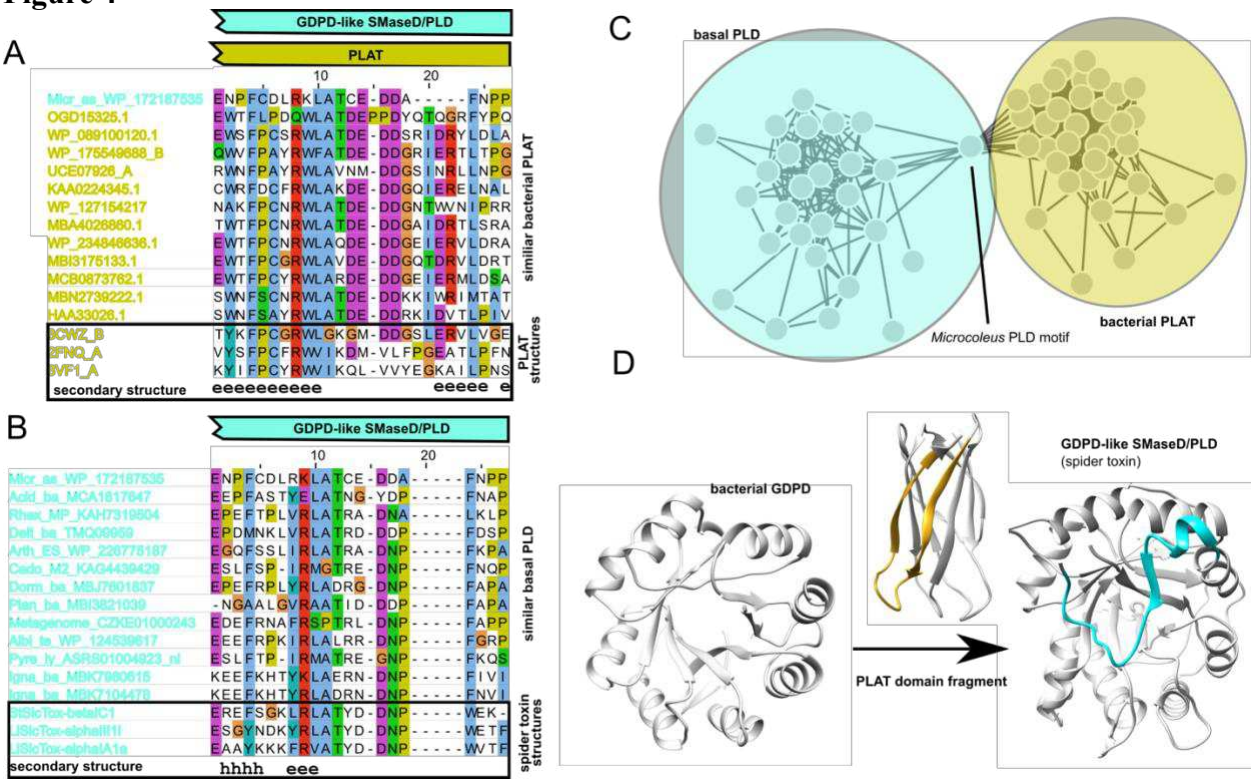

Figure 4. Repurposing and remodeling of a PLAT fragment as a C-terminal PLD motif. (A) Alignment of *Microcoleus asticus* PLD domain C terminus (cyan label) with C termini of bacterial PLAT domains (gold labels) as well as the closest PLAT homologs of known structure (box) annotated with secondary structure (h=helix, e=strand; assignments come from 3CWZ, but a general strand-loop-strand conformation is seen in all three; see Figure S1B). The region of alignment is expanded slightly relative to the BLAST alignments shown in Figure 3. (B) Alignment of *Microcoleus asticus* PLD domain C terminus with other basal PLD C termini and with C termini from spider toxins of known structure, annotated with secondary structure (box; assignments come from 3RLH, but all three conformations are highly similar; see Figure S1A). In both panels A and B, only the 12 PLAT or PLD sequence fragments with highest sequence identity are shown, in order of decreasing percent identity. (C) Sequence identity clustering of basal GDPD-like SMaseD/PLD and bacterial PLAT sequence fragments, as well as sequences with known structure shown in panels A and B. Edges are drawn between nodes with ≥40% ID over 20 residues, after removal of the heavily gapped columns shown in panels A and B, and removal of any sequences containing additional gaps. Six of 30 gapless basal PLD sequences are singletons at this cutoff (not shown), but all other sequences are unified. Note the central position of the *Microcoleus* PLD motif as an intermediate sequence (transitive connection) that is similar to multiple PLD- and PLAT-derived segments. (D) Mapping of the homologous PLAT (gold) and PLD (cyan) subdomains onto known structures. Bacterial GDPD domains (represented here by PDB ID 2PZ0) lack a C-terminal extension on the barrel. The GDPD-like SMaseD/PLD domain family has acquired a C-terminal extension, which in spider toxins (PDB ID 3RLH) has a conserved structure consisting of a helix, strand and complex ordered loop. The extension is a remodeled, repurposed fragment of a PLAT domain (PDB ID 3CWZ) consisting of strands β7 and β8.

Figure 5

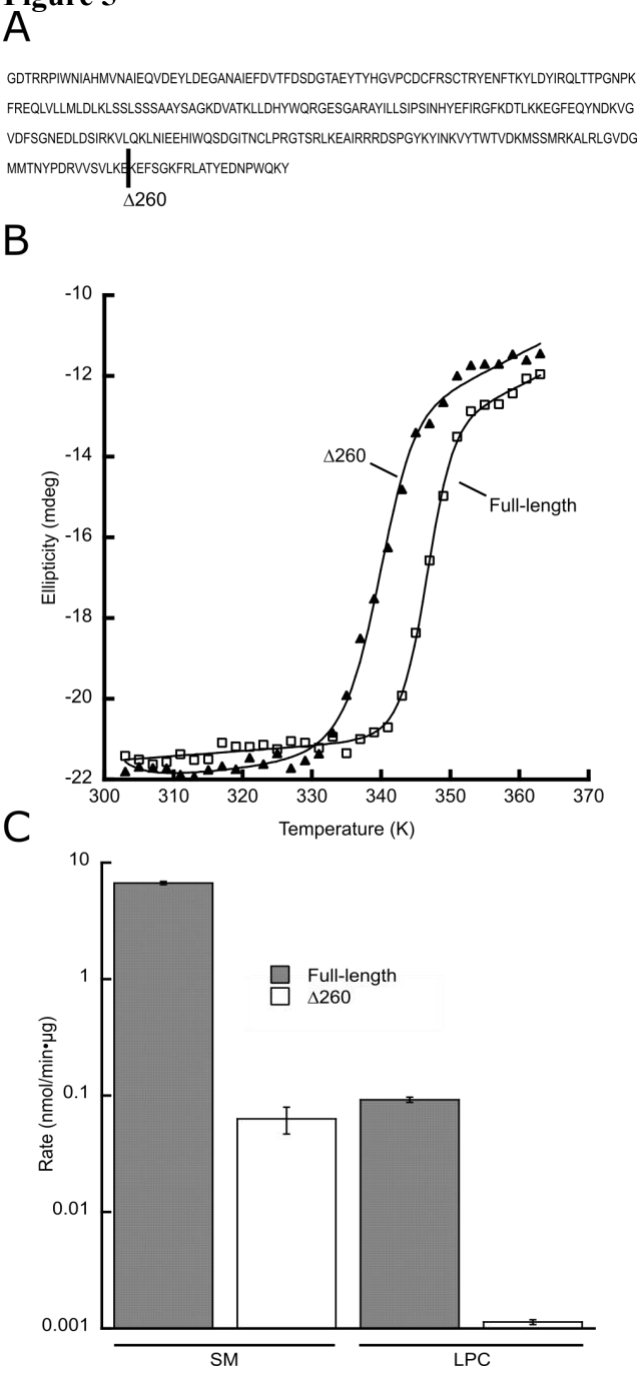

Figure 5. Truncation of the C-terminal extension decreases both thermostability and catalytic activity. (A) sequence of reconstructed Sicariid venom ancestral PLD domain and C-terminally truncated variant (Δ260). (B) thermal denaturation monitored by change in circular dichroism at 222 nm, at 0.2 mg/mL protein and a pathlength of 1 mm, in 0.1 M Tris (pH 8.0), 0.2 M sodium chloride. (C) Choline head-group release assay of enzymatic activity using sphingomyelin (SM) or lysophosphatidylcholine (LPC) as substrate (200 µM) in mixed micelles with 0.1% Triton X-100, in reaction buffer (0.1 M Tris [pH 7.4], 10 mM $MgCl_2$) at 37 °C. Error bars represent the standard error of the mean of three measurements.