Domain loss enables evolution of novel functions in a gene superfamily, including snake 3-finger toxins

Ivan Koludarov [1,*,^], Michael Heinzinger [1,*], Timothy N. W. Jackson [2], Daniel Dashevsky [3], Steven D. Aird [4] & Burkhard Rost [1, 5]

[1]TUM (Technical University of Munich) Department of Informatics, Bioinformatics & Computational Biology - i12, Boltzmannstr. 3, 85748 Garching/Munich, Germany
[2]Australian Venom Research Unit, Department of Biochemistry and Pharmacology, University of Melbourne, Australia
[3]Australian National Insect Collection, Commonwealth Scientific & Industrial Research Organisation, Canberra, Australia
[4]7744-23 Hotaka Ariake, Azumino-shi, Nagano-ken, Japan 399-8301
[5]Institute for Advanced Study (TUM-IAS), Lichtenbergstr. 2a, 85748 Garching/Munich, Germany & TUM School of Life Sciences Weihenstephan (WZW), Alte Akademie 8, Freising, Germany
*Contributed equally
^ Corresponding author:koludarov@rostlab.org, http://www.rostlab.org/
Tel: +49-289-17-811 (email rost: assistant@rostlab.org)

Abstract

3-finger toxins (3FTxs) are a functionally diverse family of toxins, seemingly unique to the venoms of caenophidian snakes. Although the ancestral function of 3FTxs is antagonism of nicotinic acetylcholine receptors (nAChR), redundancy conferred by accumulation of duplicate genes has facilitated extensive neofunctionalization, such that derived members of the family interact with a range of targets. 3FTxs are members of the Ly6/uPAR family, but their precise non-toxin ancestor remains unknown. Using a combination of traditional phylogenetic approaches, manual synteny analysis, and cutting-edge machine learning techniques, we have reconstructed a detailed evolutionary history of 3FTxs. We identify their immediate ancestor as a non-secretory Ly6, unique to squamate reptiles, and propose that changes in molecular ecology resulting from loss of a transmembrane region and changes in gene expression, paved the way for evolution of one of the most important families of snake toxins.

Key words: gene family evolution, evolution of function, comparative genomic analysis, protein natural language models, protein structure prediction, transmembrane helices.

Introduction

Proteins evolve as 3D objects. The low cost of accumulating sequence data has permitted molecular evolutionary studies to focus on the central role of nucleic acid sequences in "storing and transmitting" biological information. Furthermore, the linear nature (writable as 1D-strings of letters) of such sequences makes them readily amenable to computational analyses that might increase our understanding of ways in which traits evolve and affect their functional roles in the life histories of the organisms that possess them. However, since functional traits, including those of individual molecules, are relational, we cannot understand their evolution by focusing on sequences alone [1, 2]. Gene-products act as three-dimensional (3D) protein structures and one-dimensional (1D) gene/protein sequence data alone are insufficient to explain molecular mechanisms or evolutionary history. Although nucleic acids "store" information, deployment of that information is selective and context-specific [3]. A multi-level approach is therefore required to understand phenotypic consequences of sequence variation upon which selection acts.

3FTx from venoms are an excellent model system. Venoms are excellent model systems to study the impact of gene sequence change upon protein function because a majority of proteinaceous toxins are encoded by single genes, adapted for specific functions when injected into target organisms [4]. Few other systems offer such a clear causal pathway from genetics to ecology. This tractability has greatly benefitted research investigating toxin-derived molecules as drugs or utilizing them as investigational ligands [5]. In this study we chronicle origins and innovations of an ancient gene superfamily that includes one of the most widespread and functionally diverse toxin families in snake venoms. Three-finger toxins (3FTxs) are derivatives of a protein superfamily called lymphocyte antigen 6 (Ly6) or urokinase-type plasminogen activator receptors (uPAR) [6]. Members of this superfamily are located on several chromosomes in vertebrates and are also present in many groups of invertebrates, indicating their ancient origin. Although 3FTxs have appeared relatively recently within this superfamily, they are its best characterized members, and as a result, the Ly6/uPAR superfamily is also sometimes referred to as "toxin-like proteins" (TOLIP), or the snake-toxin-like superfamily [7].

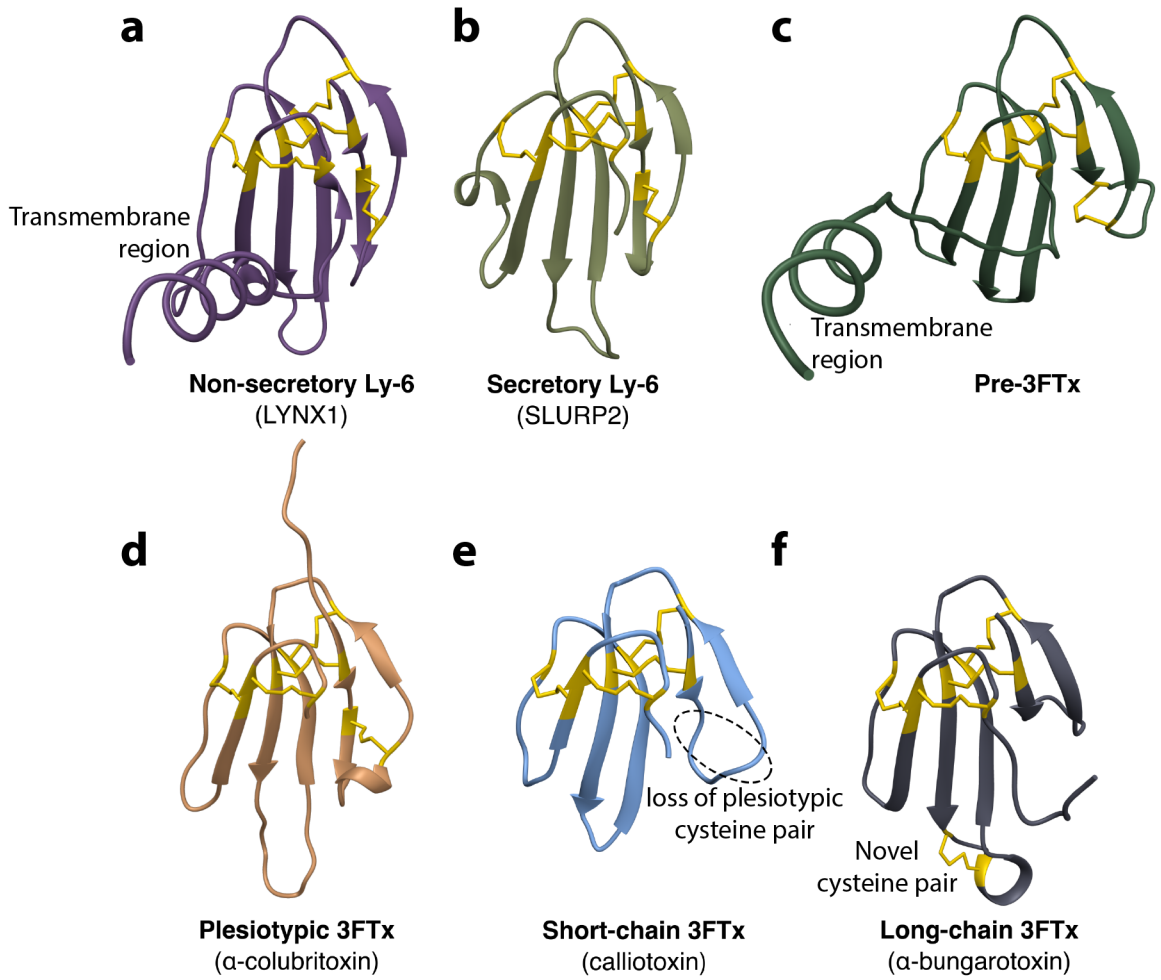

Fig. 1: Representative structural diversity in 3FTx/Ly6/uPAR family. Six representative structures with two Ly6 (non-secretory and secretory) and four 3FTx (pre-toxin form, plesiotypic/plesiomorphic ten-cysteine form, derived elapid eight-cysteine form, and the derived ten-cysteine form found in elapid venoms). All structures computed as a part of this study (see Materials and Methods and Supplementary Materials). Disulfide bonds are highlighted in yellow.

Structural theme and diversity of 3FTx/Ly6/uPAR. Ly6/uPAR proteins possess a characteristic "LU" or "3FF" ("three-finger fold"; Fig. 1). The group includes membrane-bound proteins (including multimeric uPARs), which have a region with transmembrane helices (TMHs), and secretory proteins lacking this TMH-region (e.g., secreted Ly-6/uPAR-related protein [SLURP] and 3FTxs) [6-8]. Many Ly6/uPAR proteins function as lymphocyte antigens, while others bind and modulate nicotinic acetylcholine receptors (nAChRs) [7], Table 1. The latter function appears to be mediated by the cysteine-rich 3FF region typically covering 60-74 residues. While this region does not seem to exhibit enzymatic activity, 3FTxs have co-opted their affinity for the nAChR as the primordial toxic function [9]. In snakes, these genes were recruited to serve as neurotoxins, leading to an explosive diversification of toxin sequences and activities [7, 10]. Plesiomorphic forms of the 3FTx family possess ten cysteines joined in five disulfide bonds (Fig. 1). These forms exert "α-neurotoxicity" by antagonizing nAChRs, with some forms exhibiting greater affinity for reptile/bird rather than mammalian receptors [8].

This primordial theme has been extensively varied, generating considerable evolutionary and ecological impact. For instance, dimeric 3FTx complexes in the colubrid genus *Boiga* (cat snakes) are roughly 100-times more potent than their monomeric toxin ancestors [11]. Following an accidental introduction of a brown tree snake *Boiga irregularis* on Guam, these toxins, which are the major component of its venom, facilitated extirpation of many bird species on the island [12]. Another 3FTx innovation was the loss of cysteine residues in Loop I (Fig. 1) in the common ancestor of the Elapidae (elapid snakes), one of the two families that contain most of dangerously venomous snakes (another being Viperidae, vipers). The resulting "short-chain" α-neurotoxins (Fig. 1E) exhibit increased potency, which likely contributed to explosive diversification of elapids [9]. The structural integrity conferred by simplification of these toxins facilitated an accumulation of mutations in their surface residues, with consequent diversification of their interaction partners [13]. Various short-chain 3FTxs bind and modulate a range of targets, including muscarinic acetylcholine receptors, adrenergic receptors, L-type calcium channels, and voltage-gated sodium channels. Additional activities of 3FTxs include acetylcholinesterase inhibition, platelet inhibition, and non-specific cytotoxicity [9, 10].

"Long-chain" neurotoxins are an innovation of the elapid snakes with a novel pair of bonded cysteines on Loop II (Fig. 1F). These toxins bind a broader range of nAChR subtypes than their short-chain counterparts and structural analyses of the interactions of long-chain toxins with nAChRs indicate that they utilize distinct binding sites [10]. One long-chain neurotoxin, α-bungarotoxin, from the venom of *Bungarus multicinctus* (many-banded krait), was the first to be characterized, and has since seen widespread use as an investigational ligand. Its potent, specific binding to nAChR was useful for investigations of the distribution and functional diversity of nAChR subtypes [14]. 3FTxs have also been investigated for their medical potential, especially as painkillers [10]. Toxins from venoms of *Dendroaspis polylepis* (black mamba), *Ophiophagus hannah* (king cobra), and *Micrurus lemniscatus* (South American coral snake), all possess pain inhibiting effects. Toxins from *Ophiophagus* and *Micrurus* are even active when administered orally [15]. Their small size makes 3FTxs easier to synthesize than large, enzymatic toxins and the structural integrity conferred by their disulphide-bonded cores may contribute to their oral activity and potential commercial utility. Extensive neofunctionalization of 3FTxs suggests that there may be therapeutic applications far beyond those that have already been discovered and initial experiments in directed evolution of 3FTxs show that they can be readily modified to acquire previously unknown activities, such as protease inhibition [10, 16]

Relation between plesiomorphic toxins and derived elapid forms remains elusive. Despite the research effort devoted to 3FTxs, the relation between plesiomorphic toxins and derived elapid forms has been difficult to discern. Molecular phylogenies of 3FTxs have generally only included plesiomorphic or derived forms, often due to a narrow focus of the research in question or because Bayesian phylogenies that include both types almost invariably collapse into extensive polytomies that are not very informative [17, 18]. Ancestral relations within the Ly6/uPAR superfamily remain unclear, although they have long been suspected to be orthologues of LYNX1 and its homologues on Chromosome 8 in humans [8]. We used public sequence data (including newly annotated and re-annotated genes) to examine the Ly6 genomic region in various vertebrate genomes to reconstruct the evolutionary history of 3FTxs and non-toxin family members. For simplicity, we refer to genes in this region as "Ly6", while using "TOLIP" to refer to the larger parental group of uPARs.

New AI perspective: AlphaFold2 prediction and protein Language Models (pLMs). While high-quality sequence data have been exploding for decades [19] much of these data remain

unannotated and inferring structure or activity from sequence has been challenging. Recent advances in artificial intelligence (AI) promise to ease these bottlenecks, e.g., Nature's Method of the Year 2021 [20], AlphaFold2 [21] has freely made accurate predictions of 3D protein structure for over 100 million proteins publicly available [22] (by summer 2022). This wealth of data unlocks the potential of existing methods to infer similarity in activity between proteins based on their structures using fast, accurate structure-comparison methods such as FoldSeek [23]. Independently, another breakthrough surfed on advances in natural language processing (NLP) by implementing the following two analogies: NLP words become amino acids and NLP sentences become proteins. Thereby protein Language Models (pLMs) learn to understand some aspects of the language of life as written in protein sequences [24-26]. These pLMs incorporate constraints put upon sequences through protein structure and function in an entirely different way than previous approaches. For instance, requiring only single protein sequences as input, they improve protein comparisons [27-32]. Akin to a new experimental technique, this provides a novel perspective on the relationship between proteins, orthogonal to established approaches such as homology-based inference. Given the correct representation geometry, protein language models can produce a sequence space that reveals previously obscured biological information [33], an invaluable feature to study gene evolution at the center of the present study.

In this study, we combined these computational developments, i.e., highly accurate 3D structure predictions from AlphaFold2 and representations derived from the pLM, ProtT5 with traditional bioinformatics analyses and extensive manual microsyntenic analyses. This approach [34] enabled reconstructing the evolutionary history of the Ly6 gene family at an unprecedented level of detail and cataloguing the emergence of novel structures and functions that led from an ancestral Ly6 to the radiation of deadly snake venom 3FTxs.

Table 1: Existing knowledge on human Ly6 genes – homologs of 3FTx.

Protein name	Sketch of known information
GPIHBP1 Glycosylphosphatidylinositol anchored high density lipoprotein binding protein 1	Capillary endothelial cell membrane protein that assists lipoprotein lipase in processing dietary lipoproteins.
Ly6H Lymphocyte Antigen 6 Family Member H	Believed to act as a modulator of nAChRs activity. Seems to inhibit alpha-7/CHRNA7 signaling in hippocampal neurons. Membrane-anchored.
Ly6L Lymphocyte Antigen 6 Family Member L	Membrane-anchored.
Ly6E Lymphocyte Antigen 6 Family Member E	Anchored cell surface protein that regulates T-lymphocytes proliferation, differentiation, and activation.
Ly6D Lymphocyte Antigen 6 Family Member D	Predicted to be involved in lymphocyte differentiation. Predicted to be membrane-anchored.
LYNX1 Ly6/Neurotoxin 1	Membrane-anchored. Modulates functional properties of nicotinic acetylcholine receptors (nAChRs) to prevent excessive excitation, and hence neurodegeneration.
SLURP2 Secreted Ly6/UPAR Related Protein 2	Secreted. Binds and may modulate the functional properties of nicotinic and muscarinic acetylcholine receptors. May regulate keratinocytes proliferation, differentiation and apoptosis.
LYPD2 Ly6/PLAUR Domain-Containing Protein 2	Predicted to be located in extracellular region and plasma membrane.
SLURP1 Secreted Ly6/UPAR Related Protein 1	Secreted. Was found to be a marker of late differentiation of the skin. In T cells may be involved in regulation of intracellular Ca(2+) signaling.
Ly6K Lymphocyte Antigen 6 Family Member K	Predicted to be located in cell surface; cytoplasm; and plasma membrane. Predicted to be active in acrosomal vesicle and to be involved in binding activity of sperm to zona pellucida.

Protein name	Sketch of known information
PSCA Prostate Stem Cell Antigen	Membrane-anchored. Highly expressed in the prostate, also expressed in the bladder, placenta, colon, kidney, and stomach. May be involved in the regulation of cell proliferation. In vitro inhibits nicotine-induced signaling.

Based on UniProt and Entrez databases. In order of chromosomal position downstream from TOP1MT gene.

Results

Genomic analysis reveals an evolutionary stable Ly6 cluster. To distinguish distantly related TOLIP uPAR/PLAUR genes from Ly6s that are direct homologs of snake 3FTxs, we surveyed genomes of 14 species of Tetrapoda (four-limbed vertebrates, Table 2) for regions syntenic with ophidian 3FTxs (Fig. 2). We identified the mitochondrial topoisomerase, TOP1MT , as the best marker gene for this region. In all genomes surveyed, this gene was present in a single copy and either flanked the 3FTx/Ly6 genomic region or was in the middle of it (Fig. S3, Supporting Online Material, SOM). Another useful marker gene is THEM6 (Thioesterase Superfamily Member 6 – UniProt ID Q8WUY1 for a human form). In mammals, some reptiles (including snakes), and *Xenopus* (clawed frog), this gene is located near SLURP1 homologs, making its stable synteny rather ancient.

We located all Ly6/3FTx genes (both full-length and missing exons) using previously established methods of BLASTing exons across a genomic region [34-37]. After this, we could describe multiple genes, pseudogenes, and orphan exons in each species examined. In most cases, the Ly6/3FTx cluster was located downstream from TOP1MT. It spanned several megabase pairs (Mbp), contained 8-30 isoforms (10, on average, for non-venomous species) from the Ly6/3FTx family and rarely included genes from any other protein family (Fig. 2, Fig. S3). Syntenic arrangements were extremely conserved amongst all species observed, with the notable exception of snakes (Fig. 4), and most genes were shared by all members within each taxonomical class (i.e., Reptilia or Mammalia).

The 3FTx cluster is dramatically expanded in *Naja naja* (Indian cobra), the only elapid genome sequenced to-date at the chromosomal level [35]. This pattern conformed to the central role of this toxin family in the evolution of venoms in elapid snakes [9]. Another unique feature of elapid genomes was evidenced by the presence of two genes in the the Indian cobra (*N. naja*) located outside the TOP1MT neighborhood (on Chromosome 3). Those were single-copy Ly6/3FTx family genes located on Chromosomes 1 and 4. We could locate an ortholog only for the one on Chromosome 4 in the *Pseudonaja textilis* (Eastern brown snake) genome. The genome of no other organism analyzed had any related genes (or stray exons) in homologous genomic regions. Many 3FTx genes have a 4-exon structure distinct from the typical 3-exon structure (Fig. 4B). All additional exons are duplicated and pseudogenized copies of exon 3, inserted before the original exon 3. We did not find evidence involving these redundant exons in protein production.

Protein 3D structures analysis. We used the ColabFold [38] implementation of AlphaFold2 [21] to predict protein 3D structure for all 991 Ly6 family proteins in our dataset (Fig. S4). Whenever an experimental 3D structure was available, we validated the AlphaFold2/ColabFold predictions against it. As expected, all proteins in the dataset were observed and predicted in the same canonical 3FTx-domain architecture. They differed in their compactness (density of inter-residue contacts) and loop length (continuous regions with neither helix nor strand). Most Ly6 genes (~86%) possessed a lengthy N-terminal “tail” in some cases predicted to adopt an alpha helix

Phylogenetics (Bayesian, ML, DALI). We removed signal peptides from the sequences (to avoid using potentially mis-predicted N-termini) and used MAFFT [39] to align processed sequences. The resulting alignments seeded the Bayesian (exaBayes [40]) and Maximum Likelihood (iqTree [41]) phylogenetic trees. We also generated a phylogenetic tree based on DALI (Distance-matrix ALIgnment [42]) comparisons of predicted 3D structures. While smaller protein clades were consistent across the three trees, their interrelationships differed between the Bayesian tree and the ML/DALI pair (which appeared more congruent with each other

than either was with the Bayesian tree, S2). The Bayesian phylogeny recovered a single monophyletic 3FTx clade (albeit one that was plagued by extensive polytomy, but still more robust than any previously published phylogeny (S2). This clade was sister to Ly6E-like sequences, including a clade of toxicoferan Ly6 that we refer to as "pre-3FTx" (Fig. 2). In contrast, the DALI and ML phylogenies recovered the vast majority of 3FTxs in two separate clades, one composed primarily of *Boiga* and other colubrid toxins, and the other including plesiomorphic toxins from other snake families and more derived elapid toxins. 3FTxs were even more polyphyletic in the DALI topology than in the ML topology, because several 3FTxs were orphaned on single branches, sisters to a clade containing the vast majority of Ly6 sequences and the two main 3FTx clades. Secretory Ly6 proteins lacking the transmembrane region were distributed among various clades, indicating that this domain has been lost convergently on multiple occasions.

We consider results of Bayesian analysis to be the closest representation of the actual reality of the three methods used, and therefore where applicable rely on them in the present study.

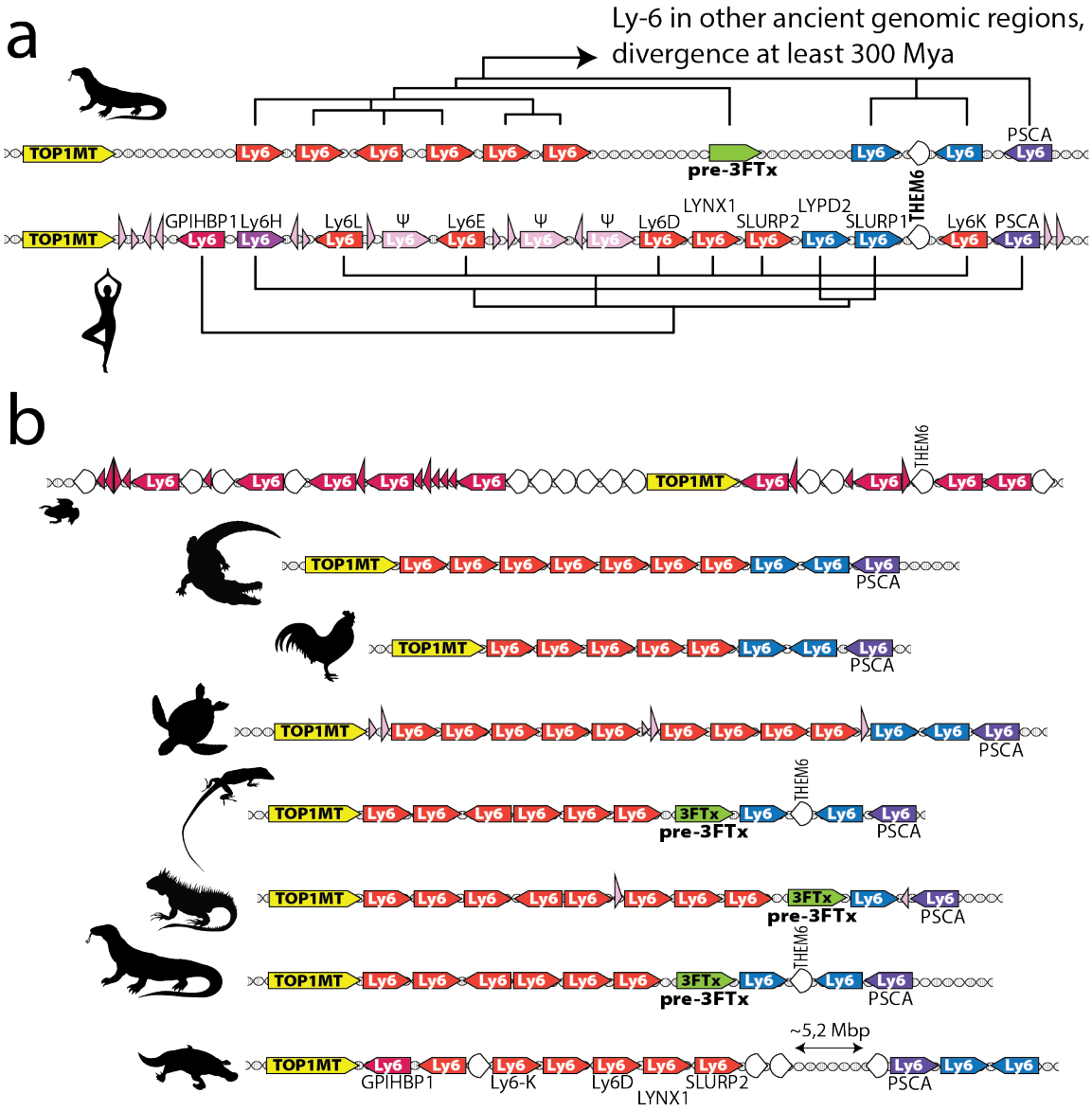

Fig. 2: Syntenic schema derived from combined phylogenetic and genomic analyses. Panel A compares Ly6/3FTx clusters from Komodo dragon and human. Gene names are from annotations. Unless marked with Ψ, gene relationships are based on our phylogenetic analyses. Panel B provides a syntenic map for representatives of tetrapod clades. Ly6 genes are colored blue (homologous to human SLURP1/LYPD2), red ("Ly6E-like"), maroon (GPIHBP in Platypus and GPIHBP-like in *Xenopus*) and violet (PSCA-like). Pre-3FTx genes are colored green. TOP1MT are yellow, and other genes are depicted as hollow semi-circular, white arrowheads, with narrow triangles representing orphan exons.

Embedding maps. We also used the protein Language Model (pLM) ProtT5 to generate multidimensional (1024d) vector representations (embeddings) from single protein sequences. Essentially, these vectors resemble the simplicity of 20d vectors obtained by computing amino acid composition, i.e. the fraction of each of the 20 amino acids in a protein. However, while amino acid composition vectors essentially ignore any influence of any pair of residues i and j on each other, pLM embeddings reflect such effects, on average. We computed embedding vectors for all proteins in two sets: the core Ly6/3FTx dataset (991 sequences), and the entire InterPro [43] TOLIP family (2,095 sequences at June 30 2022, identical sequences removed). To facilitate visual inspection, resultant embeddings were projected from 1024-dimensions to 3D via UMAP [44] (Fig. 3, http://3ftx.pythonanywhere.com for an interactive plot). The projected embedding space distribution for Ly6s from the TOP1MT region (including 3FTxs) was similar between the two sets. All Ly6 with a transmembrane regions occupied a compact area of the projected sequence space. Secretory Ly6s and 3FTxs formed a separate, more diffuse 'supercluster.' Plesiomorphic 3FTxs formed a distinct group along with Ly6s lacking the transmembrane domain, e.g., SLURPs. Toxins of colubrid snakes were found along the length of this elongated cluster while plesiomorphic toxins of elapid snakes were located together at one end (Fig. 3 and interactive plot). Near this end we found another long cluster that included all short-chain derived 3FTxs. Once again, relations within this cluster appeared relevant: cytotoxic 3FTxs were gathered at the end of one extended cluster whereas another small subcluster contained most toxins from hydrophiine snakes. There were also three outliers in this cluster: two sequences from *Helicops leopardinus* (leopard keelback) which have convergently lost the same disulfide bond as elapid short-chain toxins and one sequence from *Causus rhombeatus* (night adder) that appears to be a misattributed elapid sequence, possibly due to cross-contamination between sequencing samples. Long-chain toxins formed their own distinct cluster, which is far much more compact than those of plesiomorphic or short-chain 3FTxs.

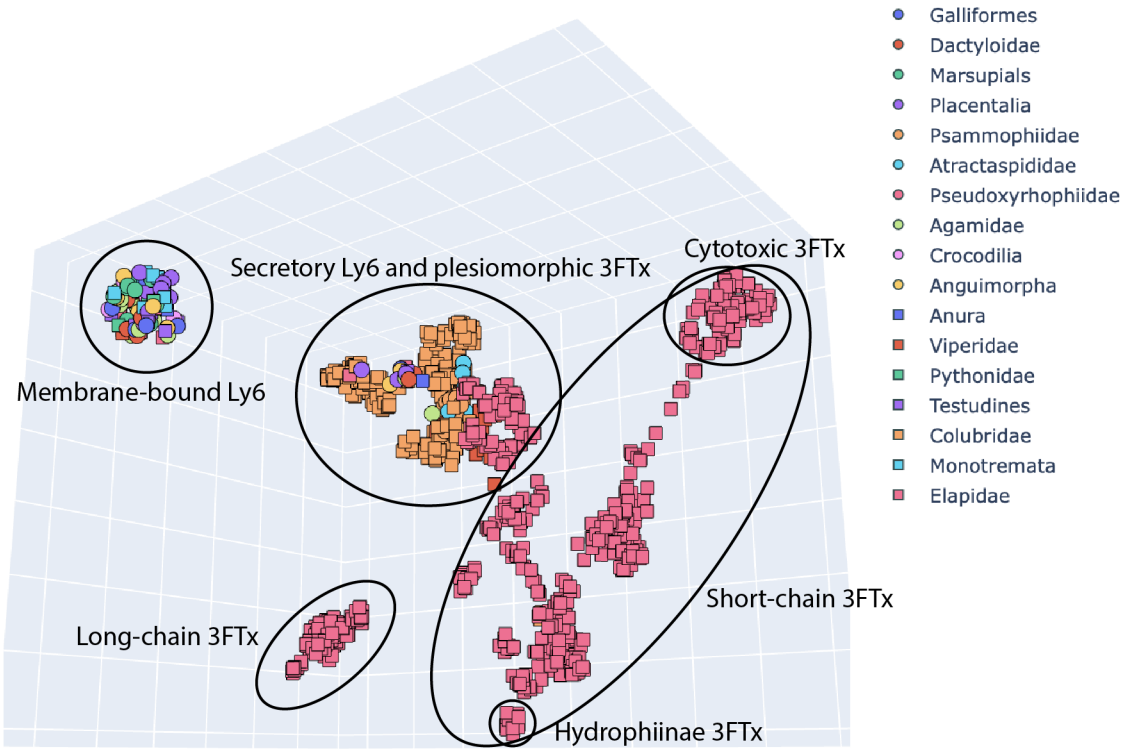

Fig. 3: Rendering of 3D UMAP projection embedding space for Ly6/3FTx. Key structural-functional clusters are circled and named. Each protein in our set was represented by a 1024-dimensional vector giving the ProtT5 pLM embeddings [26]. These 1024d vectors were projected by UMAP [44] onto 3D. The resulting embedding map seeded hypotheses on the hypothetical evolutionary trajectory of this family in protein space that were compared to existing sequence- or structure-based approaches. Interactive plots available at: http://3ftx.pythonanywhere.com .

Discussion

Embeddings boost understanding of the origins of three-finger toxins. Our results vividly illustrate why reconstructing the early evolution of 3FTxs has proven difficult [8]. Each traditional phylogenetic method that we employed provided conflicting results and the Bayesian phylogeny exhibited unresolved polytomies similar to those that have plagued previous efforts to incorporate plesiomorphic and derived toxin sequences into the same trees (Fig. S2). However, despite the polytomous substructure of the 3FTx clade in the Bayesian phylogeny, our synteny- and PLM-based methods are concordant with a single origin of the whole toxin clade, as opposed to multiple origin hypotheses implied by the ML and DALI phylogenies.

The embedding model differentiated remarkably well the major molecular and functional forms of 3FTxs that have been identified by decades of structure-function research. Impressively, endophysiological Ly6, plesiomorphic, short-chain, and long-chain toxins all form distinct clusters, toxins of shared activity, such as cytotoxins, even group together within these clusters (Fig. 3). Importantly, this grouping emerged intrinsically without having to invest any expertise other than choosing the entire InterPro [43] TOLIP family. Thus, using embeddings from pLMs clearly constituted a progress toward predicting molecular activity from sequence data by analogy with prior knowledge and enabled the annotation of uncharacterized proteins extracted from the genomic data with greater confidence than possible when only using phylogenies.

3FTx evolve from a single-copy gene unique to Toxicofera reptiles. Identifying a common, endophysiological ancestor of exophysiological 3FTxs has proven challenging, although previous studies have suggested monomeric neuromodulatory Ly6s, such as LYNXs and SLURPs, to be the most likely candidates [8]. Our embedding results put LYNXs in the main cluster of Ly6 sequences while SLURP2s and LY6Ks clustered with plesiomorphic 3FTxs as the closest human homologs. This makes intuitive sense, as secretory SLURPs lack the transmembrane region of 3FTxs. A small clade of Ly6 genes in lizards and snakes was revealed as a sister group to 3FTxs in our Bayesian phylogeny. These sequences are from a single-copy gene group that we located exclusively in toxicoferan genomes and dubbed “pre-3FTx” (Fig. 4). Interestingly, members of this clade have a transmembrane region, so that they cluster with the main Ly6 group in embedding space. In turn, this indicated that our embedding space analyses did not constitute a phylogenetic method, as it groups proteins according to structure and function phenotypes that may evolve convergently.

Syntenic analyses indicated that all reptiles have a genomic cluster (of varying copy number) of Ly6-like genes between TOP1MT and the more conserved SLURP1-like uPAR genes. Given the stability and age of these syntenic groups, it is likely that each of the proteins encoded by these genes possesses a conserved endophysiological function, including (in some cases) regulation of acetylcholinergic pathways. The copy farthest downstream in this cluster became the “pre-3FTxs” in the Toxicofera. After losing the transmembrane region (convergently with other secretory Ly6s), and gaining increased affinity for nAChRs, this form gave rise to snake venom 3FTxs. This led to a dramatic expansion of the cluster eventually resulting in genomes such as *Naja naja* (Indian cobra) with more than 30 3FTx genes and many orphan exons and pseudogenes that testify to a long history of “birth-and-death” evolution (Fig. S3). Likewise, the taxonomic distribution of leaves and branches of Ly6/3FTx protein phylogenies clearly indicates constant emergence of new forms and copies in the genomic cluster.

These genomic remnants discovered in our study are consistent with past research, which has repeatedly found 3FTx genes to be subject to high levels of diversifying selection and extremely fast rates of molecular evolution [9, 45, 46]. Such rapid rates of evolution have been attributed, partially, to relaxation of selective constraints on individual members of multigene arrays possessing a degree of functional redundancy [46]. Accumulation of 3FTx genes in the Indian cobra (*N. naja*) genome (the only elapid genome available on the chromosome-level) provides a vivid example of this process. Unique syntenic arrangements of snake 3FTx and Ly6 genes also enable speculation that some of the initial "Ly6E-like" genes were lost to recombination with newly evolved 3FTxs. If so, this would indicate the possibility of "reverse-recruitment" [47], or "moonlighting" (adaptation of separate function) [48], in which 3FTxs, derived for envenomation functions, acquired regulatory functions of their endophysiological ancestors. Although details of such molecular ecology are beyond the scope of this paper, expression of "3FTx-like" proteins in diverse tissues [49] and various mutations in nAChRs that confer some resistance to 3FTxs [50] suggest this possibility.

3FTx explosively evolve in snakes following the loss of a TM region. Over a period of ~25-38 million years [51, 52] Ly6, a previously stable cluster of genes, erupted into a massive array of specialized proteins with a novel exophysiological function: neurotoxic subjugation of prey organisms. An initial small but significant change in a membrane-bound protein in lizards, visible both in terms of embeddings and in structure- and sequence-based trees, was followed by a swap of its transmembrane region for a signal sequence. This swap may have resulted from partial recombination with secreted members of the Ly6 cluster located farther away on the genome, i.e., SLURP1/LYPD2. The novel secretory protein, perhaps constitutively expressed in oral glands of early snakes [53, 54] may have facilitated chemical subversion of neurotransmission in prey. If, as some have conjectured [53, 54], an incipient venom system was one of the key adaptations behind the explosive radiation of snakes, 3FTxs may have served as one of the first major toxin families. Viperids, which do not express appreciable quantities of 3FTxs in their venoms, with the possible exception of *Azemiops feae* (Fea's viper) [55], have "fully functional", though dormant, 3FTx genes (Fig. 4B). The reasons for this extreme conservation remain obscure, but could include regulatory spillover from molecular mechanisms responsible for conservation of Ly6 genes, or could indicate an endophysiological role for 3FTxs in viperids, at least some of which possess 3FTx-resistant nAChRs [50].

A recent analysis of the genome of *Bungarus multicinctus* (published too late for inclusion in our database) concluded that 3FTx may have originated via neofunctionalization of "Ly6E" [56]. Our results, garnered utlizing a broader range of methods and genomic data, support the identification of Ly6E as a close mammalian relative to 3FTx, however we identified a group of genes that are quite likely direct descendants of transitional forms between Ly6E and 3FTx. The initial expansion of Ly6E-like genes was the result of multiplication, or mutation followed by recombination throughout the cluster, of ancestral GPIHBP1-like genes, the only members of the gene family present in the *Xenopus* (clawed frog) genome. PSCA and other Amniote (branch of tetrapods containing reptiles and mammals) genes (including SLURP1/LYPD2 and Ly6H) are also likely derived from an ancestral GPIHBP1-like pool of genes. Of nine genes in the frog, eight display a complex α-helix transmembrane loop, indicating the ancestrality of this feature, which is shared by almost all family members except for those in the subfamilies of SLURP1/LYPD2 and 3FTx genes. This distinction between non-secretory, membrane bound proteins and secretory forms lacking the transmembrane region also represents the primary disjunction, or "leap" in protein configuration space (Fig. 3). The loss of the transmembrane region is thus relatively common in evolution of the Ly6 family,

having occurred convergently on multiple occasions, including in the origin of 3FTxs from "pre-3FTxs".

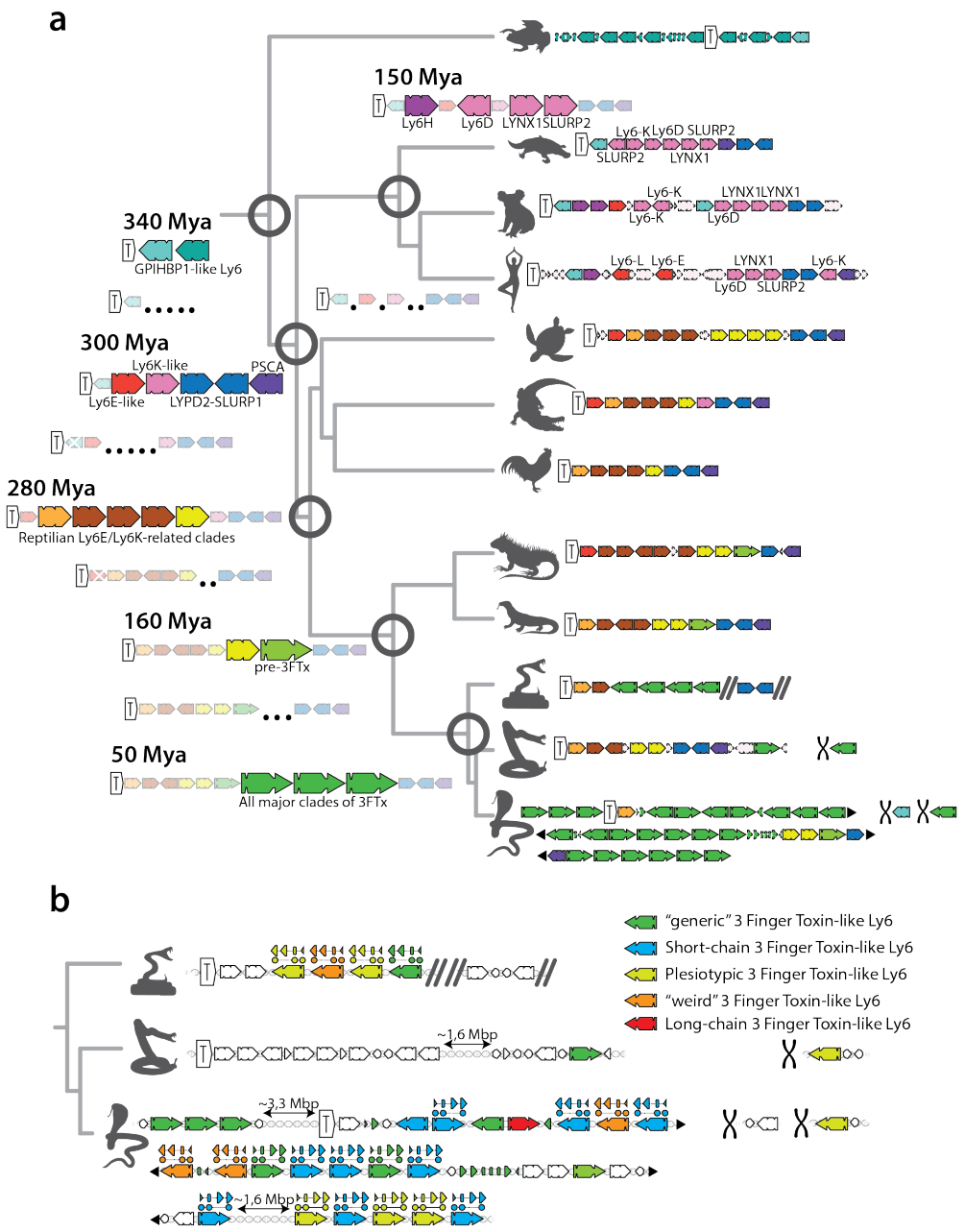

Fig. 4: Synteny map of Ly6/3FTx in tetrapods and an inferred evolutionary scenario. Panel A: Genomic regions around the TOP1MT gene (white arrow marked with "T") in tetrapod genomes and a proposed evolutionary scenario. Genes unrelated to the Ly6/3FTx family have been left out (exception: TOP1MT). Groups of Ly6 are colored by homology: cyan for GPIHBP1-like, blue for LYPD2-SLURP1-like, purple for PSCA, pink for Ly6K-like, bright red for Ly6E-like, violet for Ly6H-like, orange, brown and yellow for reptilian relatives of Ly6E/Ly6K, and green for 3FTx. Double diagonal bold lines signify breaks in a bioinformatic scaffold. Black triangular arrows show continuation of the chromosomal region, folded for the sake of space. The black chromosome icon specifies chromosomes different from those that house the TOP1MT region. Dotted outlines show pseudogenes and orphan exons. Panel B: 3FTx genomic regions in snakes. Non-3FTx Ly6 are in white. Genes unrelated to the Ly6/3FTx family are indicated by white semi-circular arrowheads. Exonic shapes above genes show non-canonical exonic structure with duplicated Exon 3. In all cases, only a single copy of exon 3 is used for a mature sequence, while another one is pseudogenized.

Snake toxin genes suggest convergent evolution. The evolutionary history of 3FTxs is marked by many unique and intriguing details. Nevertheless, its general contours are very similar to scenarios we have previously described for snake venom serine proteases (SVSP) and viperid venom phospholipases A_2 (PLA_2) [1, 36, 37]. A more-or-less conserved group of physiologically important genes from the same protein family occupy a tight cluster on a chromosome with a genomic architecture shared between representatives of extant clades of tetrapods. Among mammals and toxicoferan reptiles, a particular clade of genes is exapted for new functions, usually immune functions in mammals, and toxicity in snakes; however, the change always includes toxicoferan "lizards." In each of these cases, a single gene that mutated into a form distinct from the rest of the cluster has founded a whole sub-family of genes and functions, with its members being so numerous and clinically relevant that research efforts directed at them overshadow the older and supposedly more important parental clade.

As in the pattern described for $PLAs_2$ [36] and SVSPs [37], the major "neofunctionalization" (in a broad sense) events in the Ly6 family follow an apparently "arbitrary" change in molecular ecology. The major structural innovation within the family concerns the loss of the transmembrane region and the acquisition of a signal peptide, i.e., a shift from a non-secretory to a secretory role. Such mutational changes, which, according to a classic Neo-Darwinian view, are "random", or more adequately "arbitrary", as they are highly non-random and context-specific [1, 34, 57], result in a change in context, or molecular ecology, in which the secretory form is exposed to a novel milieu and the opportunity to engage novel interaction partners. The interactions of a secretory protein are distinct from those of a membrane-bound protein. Exposure to novel biochemical environments facilitates novel relations that originate in chance encounters between molecules and may subsequently be stabilized by selection [1]. "Recruitment" of an endophysiological protein, such as a secretory Ly6, as a venom toxin with an exophysiological target, follows a further arbitrary change in molecular ecology, likely facilitated by stochastic gene expression. Thus, the transition from membrane-bound "pre-3FTx" to functional 3FTx likely involved a two-stage process of transitioning between molecular ecologies, from non-secretory to secretory and from endophysiological to exophysiological. However, the possibility of 3FTx moonlighting in endophysiological roles is intriguing, and further complicates linear conceptions of neofunctionalization. Investigation of these scenarios will be further illuminated by additional high-quality caenophidian snake genomes.

Conclusion

Utilizing recent advances in machine learning/artificial intelligence, including reliable predictions of protein 3D structure by AlphaFold2 and embeddings from protein Language Models (pLMs), to complement traditional phylogenetic and genomic tools, we unveiled the complex evolutionary history of a fascinating and diverse gene family (Fig. 4). Paired with manual synteny analyses (Fig. 1), this approach provides an unprecedented window on protein evolution, enabling us to demonstrate that the major division within the Ly6/uPAR family is not phylogenetic, but structural, resulting from loss of the transmembrane region (Fig. 1), which has apparently occurred multiple times in evolution. Most significantly, we could chart the evolutionary trajectory of 3FTxs from an ancestral Ly6 via an intermediate, Toxicoferan, membrane-bound “pre-3FTx” (Fig. 3). The functionally diverse clade of snake venom 3FTxs emerges following another independent loss of the transmembrane domain; thus, contrary to prior hypotheses, they cannot be said to have descended directly from either LYNXs, Ly6Es or SLURPs.

Author Contributions

IK, SA, TNWJ and BR conceived the study. IK, DD and MH performed the experiments. All authors contributed to writing the manuscript.

Acknowledgements

IK is funded by an Alexander von Humboldt Foundation Fellowship.

Thanks primarily to Tim Karl for invaluable help with hardware and software; to Inga Weise (TUM) for support with many other aspects of this work. Also thanks to all those who maintain public databases in particular Steven Burley (PDB, Rutgers), Alan Bridge (Swiss-Prot, SIB, Lausanne), Alex Bateman (UniProt, EBI Hinxton) and their crews, and to all experimentalists who enabled this analysis by making their data publicly available.

We further thank the Bavarian Ministry of Education through funding to the TUM and by a grant from the Alexander von Humboldt foundation through the German Ministry for Research and Education (BMBF: Bundesministerium für Bildung und Forschung); BMBF [031L0168 and program 'Software Campus 2.0 (TUM) 2.0' 01IS17049]; Deutsche Forschungsgemeinschaft [DFG-GZ: RO1320/4-1].

Materials and Methods

Data genomes. All data sets were downloaded from NCBI (Table 2).

Table 2: Genomes used.

Species	ID
Alligator sinensis	GCA_000455745.1
Anolis carolinensis	GCA_000090745.2
Chelonia mydas	GCA_015237465.2
Gallus gallus	GCA_000002315.5
Homo sapiens	GCA_000001405.29
Naja naja	GCA_009733165.1
Ornithorhynchus anatinus	GCA_004115215.4
Phascolarctos cinereus	GCA_002099425.1
Pogona vitticeps	GCA_900067755.1
Protobothrops mucrosquamatus	GCA_001527695.3
Pseudonaja textilis	GCA_900518735.1

Python bivittatus	GCA_000186305.2
Varanus komodoensis	GCA_004798865.1
Xenopus tropicalis	GCA_000004195.4

Genomic and synteny analysis. As in previous studies, including one on 3FTxs, we used publicly available vertebrate genomes of good quality to establish locations and synteny of the Ly6 cluster. We used genomes for which verified RNA-seq genomic annotations were available as reference points and created an extensive map of genes that populate Ly6/3FTx clusters in those genomes. These include TOP1MT, THEM6, cytP450, KCNMB2 and others (Sup. Fig. 3). That allowed us to establish syntenic relationships of those regions in various genomes. We then used those flanking genes as a database to BLAST (NCBI-BLAST v.2.7.1+ suite, blastn, e-value cutoff of 0.05, default restrictions on word count and gaps) the genomes if they were less well annotated. That gave us several hundreds of genomic scaffolds that potentially contained Ly6 genes. We used those scaffolds for a second round of BLAST (tblastx, e-value cutoff of 0.01) against a database of sequences from well-annotated mammalian and reptilian TOLIP genes. Positive hits were checked by eye in Geneious (https://www.geneious.com), and complete exons were manually annotated and later merged into CDS of newly annotated genes if the exon order and count accorded with existing reliable Ly6 annotations. We included all resulting genes that produced viable mature peptides were then used for the phylogenetic analysis.

Dataset assembly. We extracted all complete gene coding regions from the genomic regions surveyed, combined them with previously compiled 3FTx datasets [17, 18] and then supplemented it with all non-redundant 3FTx sequences from UniProt. For broad comparison, we used all sequences in the TOLIP protein family in the InterProt database. All sequences had their signal peptides and propeptide regions removed, using well-annotated sequences from human and snakes as a guide and SignalP5 for confirmation [58].

Phylogenetic analysis. We translated all viable genes located in the previous step into proteins and aligned these with selected publicly available sequences of interest using the L-INS-i method of MAFFT software v7.305 [39] with 1000 iterations (--localpair --maxiterate 1000). These parameters were used for all subsequent alignments.

A naming convention was established to differentiate between genomic sequences (first two letters of both generic and specific epithets, followed by a number to differentiate between sequences from the same scaffold). Sequences from previously published studies and homologues from UniProt and SwissProt databases used to provide outgroups and fill gaps in sequence space (kept as UniProt or SwissProt IDs). Alignment was manually inspected for overt errors, e.g., proper alignment of the cysteine backbone, and used to construct phylogenetic trees in ExaBayes [40] (eight parallel runs of four chains each, and 6,000,000 generations) and iqTree [41] (100 boostrap), Sup. Mat. 2.

Embedding space analysis. Over the last few years, there has been a steady improvement of algorithms in natural language processing (NLP), resulting in breakthroughs in tasks such as automated translation or question answering. This was achieved in particular by training Language Models (LMs) on large, but unlabeled text corpora [59, 60]. Thanks to the sequential nature of both natural language and protein sequences, these advances have been transferred readily to protein sequences. In both cases, masked or missing tokens (words in

NLP, amino acids in computational biology) are predicted from the remaining, uncorrupted sequence context (a sentence or a protein sequence, respectively). As this objective does not require labeled data, it allows us to tap an abundance of unlabeled data [24, 26, 61, 62]. In particular, these methods leverage the wealth of information in exponentially growing, but unlabeled protein sequence databases by relying solely on sequential patterns in the input. Processing the information learned by such protein LMs (pLMs), e.g., by feeding protein sequences as inputs to the network and constructing vectors from activation in the network's last layers, yields a representation of protein sequences known as embeddings, e.g., Fig. 1 in Elnaggar et al. [26]. This allows transfer of features learned by the pLM to any predictive task requiring numerical protein representations (transfer learning), which has already been shown for various applications ranging from protein structure [31, 63] to protein function [64] Further, it was shown that distance in embedding space correlates with protein function and can be used as an orthogonal signal for clustering proteins into functional families [64].

In this work, we generated embeddings for each protein sequence using the pLM ProtT5-XL-UniRef50 [26] (ProtT5), which has been built in analogy to the NLP model T5 [60]. However, instead of being trained on natural language, ProtT5 was trained solely on unlabeled protein sequences from BFD (Big Fantastic Database; 2.5 billion sequences including meta-genomic sequences) [65] and UniRef50. Ultimately this allowed ProtT5 to learn some of the constraints of protein sequences. As ProtT5 was only trained on unlabeled protein sequences and no supervised training or fine-tuning was performed, there is no risk of information leakage or overfitting to a certain class or label. In order to transfer the knowledge or constraints that ProtT5 had learned to other tasks (transfer learning), we first created individual vector representations for each residue in a protein. In order to derive fixed-length vector representations for single proteins irrespective of protein length (per-protein embedding), we then averaged all residue embeddings in a protein (Fig. 1 in Elnaggar et al. [26]). As a result, every protein was represented as 1024-dimensional, per-protein embedding. Those high-dimensional representations were projected to 3-d using UMAP [66] (n_neighbors=10, min_dist=0.3, random_state=42, n_components=3) and colored according to their taxonomic group to allow visual analysis. Embeddings and 3D plots were created using the bio_embeddings package [67].

Protein 3D structures. For fast protein structure prediction, 3D structures for all proteins in our set were generated using default parameters of the ColabFold implementation (v1.3.0, commit: 7ebcbe62e8d88400b0e75aa0878dce2ff3a6c71f) of AlphaFold 2 (AF2), i.e., no early-stopping, no templates and no amber-relaxation were used [38, 68]. Input MSAs were generated using the search-script provided by ColabFold with highest sensitivity. As AF2 is an ensemble of five models each outputting its own 3D structure, only the best 3D structure for each protein was used for further processing. For this filtering we used the structure of the model with the highest predicted reliability, i.e., the highest predicted local distance difference test - pLDDT.

3D structure-based trees. 3D structures predicted by AF2 were used as input for a local version of DALI v5 (distance matrix alignment) [42] to generate trees based on structural similarity between all proteins in our set. All parameters were left at default, but computation was executed in parallel on 20 cores using the parameter, mpirun.

Protein property predictions. Besides protein structures, we used various other predicted protein properties to enrich our analysis. More specifically, we used a recently published pipeline that combines a variety of protein property predictors that all rely on the pLM ProtT5 to encode protein sequences. This enables efficient processing as the heavy-lifting of generating embeddings needs to be performed only once while the predictors trained add

only insignificant computational overhead allowing to predict thousands of proteins within few minutes [69]. Here, we mainly focused on predicting whether a protein is secreted or not using a method called LightAttention and on predicting transmembrane regions using the method TMbed [70].

Supplementary Materials

1. Dataset in a spreadsheet format, including aligned sequences, grouping and localisation predictions
2. Phylogenetic trees
3. TOP1MT region gene maps
4. Structures in .pdb format

References

1. Jackson, T.N.W. and I. Koludarov, *How the Toxin got its Toxicity.* Frontiers in Pharmacology, 2020. **11**.
2. Guttinger, S., *A Process Ontology for Macromolecular Biology*. 2018.
3. Gardner, K.E., C.D. Allis, and B.D. Strahl, *Operating on chromatin, a colorful language where context matters.* J. Mol. Biol., 2011. **409**(1): p. 36-46.
4. Fry, B.G., et al., *The toxicogenomic multiverse: convergent recruitment of proteins into animal venoms.* Annu. Rev. Genomics Hum. Genet., 2009. **10**: p. 483-511.
5. Fry, B.G., et al., *Seeing the Woods for the Trees: Understanding Venom Evolution as a Guide for Biodiscovery*, in *Venomous Reptiles and Their Toxins: Evolution, Pathophysiology and Biodiscovery*. 2015. p. 1-36.
6. Loughner, C.L., et al., *Organization, evolution and functions of the human and mouse Ly6/uPAR family genes.* Hum. Genomics, 2016. **10**: p. 10.
7. Kessler, P., et al., *The three-finger toxin fold: a multifunctional structural scaffold able to modulate cholinergic functions.* J. Neurochem., 2017. **142 Suppl 2**: p. 7-18.
8. Utkin, Y.N., et al., *Three-finger toxins (3FTxs)*, in *Venomous Reptiles and Their Toxins: Evolution, Pathophysiology and Biodiscovery*. 2015.
9. Sunagar, K., et al., *Three-fingered RAVERs: Rapid Accumulation of Variations in Exposed Residues of snake venom toxins.* Toxins, 2013. **5**(11): p. 2172-2208.
10. Utkin, Y.N., *Last decade update for three-finger toxins: Newly emerging structures and biological activities.* World J. Biol. Chem., 2019. **10**(1): p. 17-27.
11. Pawlak, J., et al., *Irditoxin, a novel covalently linked heterodimeric three-finger toxin with high taxon-specific neurotoxicity.* The FASEB Journal, 2009. **23**(2): p. 534-545.
12. Fritts, T.H. and G.H. Rodda, *The Role of Introduced Species in the Degradation of Island Ecosystems: A Case History of Guam.* Annu. Rev. Ecol. Syst., 1998. **29**: p. 113-140.
13. Nys, M., et al., *The molecular mechanism of snake short-chain alpha-neurotoxin binding to muscle-type nicotinic acetylcholine receptors.* Nat Commun, 2022. **13**(1): p. 4543.
14. Clementi, F., C. Gotti, and E. Sher, *Nicotinic Acetylcholine Receptors in the Nervous System*. 2013: Springer Science & Business Media. 424.
15. Gazerani, P. and B.E. Cairns, *Venom-based biotoxins as potential analgesics.* Expert Rev. Neurother., 2014. **14**(11): p. 1261-1274.
16. Makarova, Y.V., et al., *The First Recombinant Viper Three-Finger Toxins: Inhibition of Muscle and Neuronal Nicotinic Acetylcholine Receptors.* Dokl. Biochem. Biophys., 2018. **479**(1): p. 127-130.
17. Xie, B., et al., *Dynamic genetic differentiation drives the widespread structural and functional convergent evolution of snake venom proteinaceous toxins.* BMC Biol., 2022. **20**(1): p. 4.
18. Dashevsky, D. and B.G. Fry, *Ancient Diversification of Three-Finger Toxins in Micrurus Coral Snakes.* J. Mol. Evol., 2018. **86**(1): p. 58-67.
19. Mathé, E., et al., *The Omics Revolution Continues: The Maturation of High-Throughput Biological Data Sources.* Yearb. Med. Inform., 2018. **27**(1): p. 211-222.
20. Marx, V., *Method of the Year: protein structure prediction.* Nat Methods, 2022. **19**(1): p. 5-10.
21. Jumper, J., et al., *High accuracy protein structure prediction using deep learning*, in *Fourteenth Critical Assessment of Protein Structure Prediction (CASP14)*. 2021.

22. Tunyasuvunakool, K., et al., *Highly accurate protein structure prediction for the human proteome.* Nature, 2021.
23. van Kempen, M., et al., *Foldseek: fast and accurate protein structure search.* bioRxiv, 2022.
24. Heinzinger, M., et al., *Modeling aspects of the language of life through transfer-learning protein sequences.* BMC Bioinformatics, 2019. **20**(1): p. 723.
25. Ofer, D., N. Brandes, and M. Linial, *The language of proteins: NLP, machine learning & protein sequences.* Comput Struct Biotechnol J, 2021. **19**: p. 1750-1758.
26. Elnaggar, A., et al., *ProtTrans: Towards Cracking the Language of Lifes Code Through Self-Supervised Deep Learning and High Performance Computing.* IEEE Trans Pattern Anal Mach Intell, 2021. **PP**.
27. Heinzinger, M., et al., *Contrastive learning on protein embeddings enlightens midnight zone.* NAR Genomics and Bioinformatics, 2022. **4**(2): p. lqac043.
28. Littmann, M., et al., *Protein embeddings and deep learning predict binding residues for various ligand types.* Scientific Reports, 2021. **11**: p. 23916.
29. Meier, J., et al., *Language models enable zero-shot prediction of the effects of mutations on protein function.* bioRxiv, 2021: p. 2021.07.09.450648.
30. Schütze, K., et al., *Nearest neighbor search on embeddings rapidly identifies distant protein relations.* bioRxiv, 2022.
31. Weissenow, K., M. Heinzinger, and B. Rost, *Protein language-model embeddings for fast, accurate, and alignment-free protein structure prediction.* Structure, 2022. **30**(8): p. 1169-1177.e4.
32. Dunham, A.S., P. Beltrao, and M. AlQuraishi, *High-throughput deep learning variant effect prediction with Sequence UNET.* bioRxiv, 2022.
33. Detlefsen, N.S., S. Hauberg, and W. Boomsma, *Learning meaningful representations of protein sequences.* Nat Commun, 2022. **13**(1): p. 1914.
34. Koludarov, I., et al., *Bee core venom genes predominantly originated before aculeate stingers evolved.* bioRxiv, 2022.
35. Suryamohan, K., et al., *The Indian cobra reference genome and transcriptome enables comprehensive identification of venom toxins.* Nat. Genet., 2020. **52**(1): p. 106-117.
36. Koludarov, I., et al., *Family saga: reconstructing the evolutionary history of a functionally diverse gene family reveals complexity at the genetic origins of novelty.* 2019.
37. Barua, A., I. Koludarov, and A.S. Mikheyev, *Co-option of the same ancestral gene family gave rise to mammalian and reptilian toxins.* BMC Biol., 2021. **19**(1): p. 268.
38. Mirdita, M., et al., *ColabFold: making protein folding accessible to all.* Nat. Methods, 2022. **19**(6): p. 679-682.
39. Katoh, K. and D.M. Standley, *MAFFT multiple sequence alignment software version 7: improvements in performance and usability.* Mol. Biol. Evol., 2013. **30**(4): p. 772-780.
40. Aberer, A.J., K. Kobert, and A. Stamatakis, *ExaBayes: massively parallel bayesian tree inference for the whole-genome era.* Mol. Biol. Evol., 2014. **31**(10): p. 2553-2556.
41. Minh, B.Q., et al., *IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era.* Mol. Biol. Evol., 2020. **37**(5): p. 1530-1534.
42. Holm, L., *Using Dali for Protein Structure Comparison.* Methods Mol. Biol., 2020. **2112**: p. 29-42.
43. Paysan-Lafosse, T., et al., *InterPro in 2022.* Nucleic Acids Res, 2022.

44. Narayan, A., B. Berger, and H. Cho, *Assessing single-cell transcriptomic variability through density-preserving data visualization.* Nat Biotechnol, 2021.
45. Doley, R., S.P. Mackessy, and R.M. Kini, *Role of accelerated segment switch in exons to alter targeting (ASSET) in the molecular evolution of snake venom proteins.* BMC Evol. Biol., 2009. **9**: p. 146.
46. Jackson, T.N.W., et al., *Rapid Radiations and the Race to Redundancy: An Investigation of the Evolution of Australian Elapid Snake Venoms.* Toxins, 2016. **8**(11).
47. Casewell, N.R., G.A. Huttley, and W. Wüster, *Dynamic evolution of venom proteins in squamate reptiles.* Nat. Commun., 2012. **3**: p. 1066.
48. Jeffery, C.J., *Moonlighting proteins: old proteins learning new tricks.* Trends Genet, 2003. **19**(8): p. 415-7.
49. Hargreaves, A.D., et al., *Testing the Toxicofera: comparative transcriptomics casts doubt on the single, early evolution of the reptile venom system.* Toxicon, 2014. **92**: p. 140-156.
50. Khan, M.A., et al., *Widespread Evolution of Molecular Resistance to Snake Venom α-Neurotoxins in Vertebrates.* Toxins, 2020. **12**(10).
51. Zaher, H., et al., *Large-scale molecular phylogeny, morphology, divergence-time estimation, and the fossil record of advanced caenophidian snakes (Squamata: Serpentes).* PLoS One, 2019. **14**(5): p. e0216148.
52. Lee, M.S.Y., et al., *Diversification rates and phenotypic evolution in venomous snakes (Elapidae).* R Soc Open Sci, 2016. **3**(1): p. 150277.
53. Jackson, T.N.W., et al., *Endless forms most beautiful: the evolution of ophidian oral glands, including the venom system, and the use of appropriate terminology for homologous structures.* Zoomorphology, 2017. **136**(1): p. 107-130.
54. Fry, B.G., et al., *Squeezers and leaf-cutters: differential diversification and degeneration of the venom system in toxicoferan reptiles.* Mol. Cell. Proteomics, 2013. **12**(7): p. 1881-1899.
55. Babenko, V.V., et al., *Novel Bradykinin-Potentiating Peptides and Three-Finger Toxins from Viper Venom: Combined NGS Venom Gland Transcriptomics and Quantitative Venom Proteomics of the Azemiops feae Viper.* Biomedicines, 2020. **8**(8).
56. Zhang, Z.Y., et al., *The structural and functional divergence of a neglected three-finger toxin subfamily in lethal elapids.* Cell Rep, 2022. **40**(2): p. 111079.
57. Jackson, T.N.W., H. Jouanne, and N. Vidal, *Snake Venom in Context: Neglected Clades and Concepts.* Frontiers in Ecology and Evolution, 2019. **7**.
58. Almagro Armenteros, J.J., et al., *SignalP 5.0 improves signal peptide predictions using deep neural networks.* Nat Biotechnol, 2019. **37**(4): p. 420-423.
59. Devlin, J., et al., *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* arXiv [cs.CL], 2018.
60. Raffel, C., et al., *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.* arXiv [cs.LG], 2019.
61. Alley, E.C., et al., *Unified rational protein engineering with sequence-based deep representation learning.* Nature Methods, 2019. **16**(12): p. 1315-1322.
62. Bepler, T. and B. Berger, *Learning the protein language: Evolution, structure, and function.* Cell Syst, 2021. **12**(6): p. 654-669.e3.
63. Rao, R., et al., *Transformer protein language models are unsupervised structure learners.* 2020.
64. Littmann, M., et al., *Embeddings from deep learning transfer GO annotations beyond homology.* Sci. Rep., 2021. **11**(1): p. 1160.

65. Steinegger, M. and J. Söding, *Clustering huge protein sequence sets in linear time.* Nature Communications, 2018. **9**(1).
66. Leland McInnes, J.H., James Melville, *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.* arXiv [cs.CL], 2018.
67. Dallago, C., et al., *Learned Embeddings from Deep Learning to Visualize and Predict Protein Sets.* Curr Protoc, 2021. **1**(5): p. e113.
68. Skolnick, J., et al., *AlphaFold 2: Why It Works and Its Implications for Understanding the Relationships of Protein Sequence, Structure, and Function.* J. Chem. Inf. Model., 2021. **61**(10): p. 4827-4831.
69. Ferruz, N., et al., *From sequence to function through structure: deep learning for protein design.* bioRxiv, 2022: p. 2022.08.31.505981.
70. Bernhofer, M. and B. Rost, *TMbed: transmembrane proteins predicted through language model embeddings.* BMC bioinformatics, 2022. **23**(1): p. 1-19.