Toxicon 95 (2015) 84–92

Contents lists available at ScienceDirect

Toxicon

journal homepage: www.elsevier.com/locate/toxicon

Combinations of long peptide sequence blocks can be used to describe toxin diversification in venomous animals

Antonio Starcevic [a], Ana M. Moura-da-Silva [b], John Cullum [c], Daslav Hranueli [a], Paul F. Long [d, e, f, g, *]

[a] Section for Bioinformatics, Department of Biochemical Engineering, Faculty of Food Technology & Biotechnology, University of Zagreb, Pierottijeva 6, HR-10000 Zagreb, Croatia
[b] Laboratório de Imunopatologia, Instituto Butantan, Av. Vital Brasil 1500, 05503-900 São Paulo, SP, Brazil
[c] Department of Genetics, University of Kaiserslautern, Postfach 3049, 67653 Kaiserslautern, Germany
[d] Institute of Pharmaceutical Science, King's College London, United Kingdom
[e] Department of Chemistry, King's College London, United Kingdom
[f] Brazil Institute, King's College London, United Kingdom
[g] Faculdade de Ciências Farmacêuticas, Universidade de São Paulo, Av. Prof. Lineu Prestes, 580, B16, 05508-000 São Paulo, SP, Brazil

ARTICLE INFO

Article history:
Received 24 November 2014
Received in revised form
7 January 2015
Accepted 13 January 2015
Available online 14 January 2015

Keywords:
Toxin diversification
Evolution
Multiple alignments
Hidden Markov models

ABSTRACT

An important mechanism for the evolution of toxins in venomous animals is believed to be the acquisition of genes encoding proteins that switch from physiological to toxic roles following gene duplication. The 'reverse recruitment' hypothesis pertains that these genes can also revert back to physiological functions, although such events are thought to be rare. A non-supervised homology searching method was developed which allowed the peptide diversity of animal toxins to be described as combinations between limited numbers of amino-acid sequence blocks we called 'tox-bits'. Taking the phospholipase A2 (PLA2) protein family as an example, a Bernoulli Trial was used to test if 'tox-bits' were robust enough to distinguish between peptides with physiological or toxin functions. The analysis revealed that discrimination was indeed possible, and supports the very recent 'restriction' hypothesis whereby genes with the potential to encode toxic functions have likely been independently recruited into venom systems and therefore require few, if any, reverse recruitment events. The development of 'tox-bits' provides a novel bioinformatics tool to allow recognition of toxins from other proteins in genome sequences, facilitating the study of gene recruitment and duplication strategies in venom diversification. The 'tox-bits' library is freely available at http://bioserv.pbf.hr/blocks.zip.

1. Introduction

The venoms of animals are a cocktail of mainly proteins and peptides, colloquially referred to as 'toxins'. These toxins are delivered into a victim for the purposes of defence and/or predation by inflicting a wound using specialized apparatus such as fangs, pincers, barbs and harpoons. The diversity of animals that are venomous is staggering, for example, marine invertebrates (e.g. cnidarians such as jellyfish and sea anemones; cone snails and other gastropods; cephalopods and echinoderms), marine and freshwater fish, aquatic and terrestrial amphibians, reptiles especially snakes, a plethora of arachnid groups most notably scorpions and spiders; insects and even species of mammals including shrew, *Platypus* and a recently discovered slow loris primate *Nycticebus menagenis* (Fry et al., 2009; Whittington et al., 2010; Nekaris et al., 2013). Venom composition and mechanisms of delivery vary markedly, which often reflects the function of venom to the natural history of a given species, but also strongly implies that venoms have evolved independently across different phyla of the animal kingdom (Casewell et al., 2013). This complex chemical diversity is believed to have arisen by convergent recruitment of ancestral genes into different animals. These ancestral genes encoded proteins which then switch functions by duplication, followed by rapid hyper-mutation from physiological to toxic roles (Fry, 2005). Reverse recruitment of these toxin genes back to physiological functions in non-venomous tissues is also

* Corresponding author. Institute of Pharmaceutical Science, King's College London, 150 Stamford Street, London SE1 9NH, United Kingdom.
E-mail address: paul.long@kcl.ac.uk (P.F. Long).

http://dx.doi.org/10.1016/j.toxicon.2015.01.005
0041-0101/

thought to occur, but is rare (Casewell et al., 2012). Specific domains within a gene can also be duplicated or even lost, also generating toxin diversity. Extensive and continuous duplications of whole genes or gene domains give rise to larger multi-gene families (Nei et al., 1997; Kordiš and Gubenšek, 2000). These multi-gene families can encode for pre-propeptide precursors that are particularly cysteine-rich, forming disulphide bridges. Variable processing of these bridges can also generate the massive numbers and diversity of toxins that exist within a given species (Lu et al., 2014).

Eukaryotic genes usually duplicate slowly, and although gene duplication in cone snail (Chang and Duda, 2012) and snake toxins (Doley et al., 2009) may have occurred arguably at a much faster rate, convergent recruitment of sufficient ancestral genes from a single common venomous ancestor to account for the chemical diversity of venom by the gene duplication theory alone, seems unlikely unless this occurred very deep in early metazoan evolution (Starcevic and Long, 2013). Evidence from the few genomes of venomous animals currently available is at best equivocal with regard to the role that gene duplication might play in toxin diversification. For example, sequencing of the *Platypus* genome combined with transcriptomic data have found that gene duplication does not play a large role in the evolution of venom in this poisonous mammal (Wong et al., 2012). However, this is at odds with the large venom gene expansions reported in other animals including the king cobra *Ophiophagus hannah*, the genome and transcriptome sequences of which have recently been published and clearly demonstrate that gene duplication is a major driver for toxin diversification in this snake (Vonk et al., 2013). The genome sequence of the Burmese python *Python molurus bivittatus* is also now in the public domain (Castoe et al., 2013). Tissue specific gene expression profiling has shown that venom gene homologues are expressed at different levels in many different tissues outside of the python salivary glands. This observation forms a model to explain how gene recruitment and subsequent duplication might become a major mechanism of toxin diversification in venomous xenophidian snakes such as *O. hannah* (Reyes-Velasco et al., 2014). It has been suggested that gene homologues with particular expression profiles are preferentially recruited into gland tissues of venomous snakes where, following duplication the genes undergo subsequent conversion (neo-functionalization) and adopt toxic roles (Reyes-Velasco et al., 2014). In direct contrast, however, comparative transcriptomics between venomous and non-venomous reptiles has cast doubt on the extent of recruitment, duplication and subsequent conversion (neo-functionalization) of genes from physiological tissues into toxic roles when expressed in venom gland tissues (Hargreaves et al., 2014). These comparative studies form the 'restriction' hypothesis which suggests that genes that encode potentially toxic functions may also be expressed in many other tissues, including in the salivary glands of non-venomous reptiles. Hence, in this 'restriction' hypothesis, genes with the potential to encode toxin functions might not be recruited, but indeed already exist in venom gland tissue. Following duplication, these genes can then evolve by a process of sub-functionalization whereby expression of one copy, now encoding a toxic function, is restricted to the venom gland; whilst the other copy with the potential to be toxic is expressed, but in a physiological role in other tissues (Hargreaves et al., 2014).

Combinations of other molecular strategies might also exist that lead to toxin diversification without the need for gene duplication. For example, unequal crossing-over during homologous recombination of DNA, or exon shuffling of primary mRNA transcripts have both been cited, but never experimentally proven, to explain diversification of peptide sequences at the active sites of serine proteases in venom gland transcripts of the snake *Macrovipera schweizeri* (ex. *Vipera lebetina* (Siigur et al., 2001)). Recombination at the DNA or RNA levels has also been suggested as a possible explanation for sequence variation in venom gland transcripts of Class P–I and P-II snake venom metalloproteinases (SVMP) from *Bothrops neuwiedi* (Moura-da-Silva et al., 2011). This explanation was based upon mapping possible recombination sites in *B. neuwiedi* transcripts to regions close to known exon positions in a previously reported complete gene sequence for a Class P-III SVMP from *Echis ocellatus* (Sanz et al., 2012). Although these sequence comparisons were made between different classes of SVMPs taken from phlyogentically distinct snakes, when taken together with the previous example (Siigur et al., 2001), these studies highlight a possible reason as to why alternative molecular mechanisms that might give rise to toxin diversification remain largely unexplored is probably because there are so few genome sequences of venomous animal onto which RNA and peptide sequences can be mapped.

Using high throughput proteomics, we have described the proteome of a symbiont enriched fraction of the coral *Stylophora pistillata* as a complex mixture of putative toxins which was attributed to the venom content of contaminating stinging organelles (nematocysts). Many of these toxins were homologous to diverse toxins from very dissimilar animal phyla such as spiders, scorpions and snakes (Weston et al., 2012). Similarly, a large number of peptides with homology to toxins from a wide range of different animals were also reported in the proteome of isolated nematocysts of *Hydra magnipapillata* (Balasubramanian et al., 2012) and the hydrozoan jellyfish *Olindias sambaquiensis* (Weston et al., 2013). A comparison between the putative toxins of these three basal metazoans and the known toxins of higher animals suggests that toxins might be described using a limited set of smaller peptide sequences. To test this possibility we herein present a non-supervised homology searching method and demonstrate how combinations of limited numbers of sequences can be used to describe toxin diversification.

2. Materials and methods

The peptide sequences for the approximately 5000 toxins deposited in the UniProtKB/Swiss-Prot Tox-Prot dataset were downloaded in FASTA format (www.uniprot.org/program/Toxins (Jungo et al., 2012)). To assess the shared local similarities between these sequences, an 'all *versus* all' BLAST search was performed using the BLASTp program in version 2.2.27+ of BLAST+ from NCBI (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/2.2.27/ (Altschul et al., 1990)). The output from over 12.5 million comparisons was parsed and filtered using a customised program written in Python (www.python.org/) to select for high-scoring segment pairs (HSPs) with e-values below a cut-off value of 1e-05. Transitional grouping of these HSPs used an algorithm called the 'BlocksGenerator' we designed and implemented, again using the Python programming language which is freely available at http://bioserv.pbf.hr/blocks.zip. Constraint values for transitional grouping were a minimum of 70% identity for each HSP and no less than 50% overlap of corresponding protein coordinates for each shared region of homology identified. Successive cycles may start to produce transitional groups with loose cohesion which is an unavoidable trade-off between gain in sensitivity and loss in specificity. To solve this issue version 4.9.0 of the MEME motif discovery tool (http://meme.nbcr.net/meme/ (Bailey et al., 2006)) was used to identify conserved motifs in the grouped HSPs. Since the MEME tool is constrained by motif width and numbers of motifs expected, default parameters of up to 3 motifs in transitional groups of between 6 and 100 amino acids were chosen. Multiple alignments of the motifs were constructed using ClustalW version 2.1 compiled for the Debian based Linux OS (ftp.ebi.ac.uk/pub/software/clustalw2/2.1/ (Higgins et al., 1996)). Shared motif profiles from these