Toxicon 107 (2015) 344–358

Contents lists available at ScienceDirect

Toxicon

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

The importance of being genomic: Non-coding and coding sequences suggest different models of toxin multi-gene family evolution

Anita Malhotra [a, *], Simon Creer [a], John B. Harris [b], Roger S. Thorpe [a]

[a] School of Biological Sciences, College of Natural Sciences, Bangor University, Bangor LL57 2UW, UK
[b] Medical Toxicology Centre, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne NE2 4HH, UK

ARTICLE INFO

Article history:
Received 10 June 2015
Received in revised form
31 July 2015
Accepted 6 August 2015
Available online 7 September 2015

Keywords:
Gene turnover
Gene duplication
Selection
Phospholipase A_2
Crotalinae
Reconciliation methods
Functional inference

ABSTRACT

Studies of multi-gene protein families, including many toxins, are crucial for understanding the role of gene duplication in generating protein diversity in general. However, many evolutionary analyses of gene families are based on coding sequences, and do not take into account many potentially confounding evolutionary factors, such as recombination and convergence due to selection. We illustrate this using snake venom gene sequences from the Phospholipase A_2 (PLA_2) subfamily. Novel gene sequences from 20 species of understudied Asian pitvipers were analyzed alongside available genomic PLA_2 sequences from another four crotaline and several viperine species. In contrast to previous analyses of this toxin family based on cDNA sequences, we find that duplication events are concentrated at the tips of the tree, suggesting that major functions such as presynaptic neurotoxicity have evolved convergently multiple times in pitvipers. We provide evidence that this discrepancy is due to differing evolutionary patterns between introns and exons. The effects of several well-known sources of bias on the phylogeny were small, compared to the effect of analyses based on different partitions of the gene (whole gene sequence, non-coding regions, cDNA sequence). Switches of function were found to be largely associated with strong selection, and with duplication events. Use of coding sequences for phylogeny estimation potentially produces incorrect inferences about the action of selection on individual lineages and sites. Our results have major implications for phylogenomic methods of functional inference as well as for our understanding of the evolution of multigene families.

1. Introduction

Gene duplication has been recognised as an important source of evolutionary innovation in eukaryotes for decades (Haldane 1933; Ohno, 1970), and recent work (Zhou et al., 2010) suggests that it may have been fundamental to the successful radiation of early eukaryotes. More recently, studies on the scale, pattern and process of gene duplication (Cotton, 2005; Hughes and Liberles, 2007; Lynch and Conery, 2000, 2003; Mulley et al., 2006) have led to a proliferation of models (reviewed by Innan and Kondrashov, 2010). These differ in aspects such as the type and timing of selection acting on one or both of the duplicated genes during the process of spread and fixation in the population, although sometimes rather subtle differences separate alternative models. Innan and Kondrashov (2010) pointed out that most of the critical information that would allow one model to be favoured over others comes from the early stages of the process, when distinguishing orthologs from paralogs can be difficult (Ezawa et al., 2010). While phylogenetic methods (e.g. Han et al., 2009) may provide greater power to distinguish recent paralogs, lineage-specific duplications may still represent relatively old events if the lineages diverged a long time ago and duplication rates are high. Moreover, phylogenetic methods depend quite heavily on the accuracy of the phylogenetic hypothesis used.

Studies on single gene families or subfamilies that emphasize gene or genomic biodiversity (sensu Pollock et al., 2000) often rely on analyzing protein or cDNA sequence rather than gene sequences (Fry et al., 2010; Huang et al., 2012) as this is still more readily available than genomic data. However, if protein products are exposed to strong positive selection, phylogenetic analysis on coding regions may well give misleading results about the

Abbreviations used: PLA_2, Phospholipase A_2; NEOF, neofunctionalization; SUBF, subfunctionalization; EAC, Escape from Adaptive Conflict; PP, posterior probability; GTR, general time reversible model; SYM, symmetrical model; BF, Bayes factors; ESS, effective sample size; HPD, highest probability density.

* Corresponding author.
E-mail address: a.malhotra@bangor.ac.uk (A. Malhotra).

http://dx.doi.org/10.1016/j.toxicon.2015.08.009
0041-0101/

relationships of the genes themselves. The large size of such datasets also frequently leads to simplification of phylogeny reconstruction methods. As a consequence, evolutionary phenomena such as recombination (Arenas and Posada, 2010) and rate variation (among-site and/or among-lineage) (Moran et al., 2015) may not be adequately controlled. The incorporation of biologically reasonable variation in processes among sites may often account for the apparent derived trends predicted by simpler methods (Goldstein and Pollock, 2006).

In this study, we explore these themes in relation to the study of snake venom Phospholipase A_2 (PLA_2) toxins, which form a moderately sized multi-gene family, and have been well studied at the protein level due to their potential biomedical importance (Mackessey, 2010; Menez, 2002). Secretory PLA_2s, of low molecular weight (13–15 kDa), are present in all organisms except Archaea (Nevalainen et al., 2012) and are expressed in a wide variety of animal venoms, but are especially abundant and varied in viperid snakes (Kordiš, 2011). PLA_2s are the most variable of all major protein families in the venom, both intra- and inter-specifically (Creer et al., 2003; Sanz et al., 2006; Tsai et al., 2003). The substitution of amino acid residues in different isoforms has led to the proliferation of different activities, including pre- and post-synaptic neurotoxicity, myotoxicity, cardiotoxicity, anticoagulant and haemolytic activity (Doley et al., 2010). In some cases, individual PLA_2 enzymes exhibit several pharmacological effects, and a single amino acid difference can bring about significant changes in activity (Ohno et al., 1998).

Many studies of sequence variation have implicated positive selection as an evolutionary force driving the maintenance of hypervariability in these and other toxin genes (Conticello et al., 2001; Gibbs and Rossiter, 2008; Lynch, 2007; Nakashima et al., 1995). However, a thorough evolutionary analysis of PLA_2 toxins is challenged by uneven sampling of species, sampling bias towards expressed gene products, lack of availability of pre-mRNA regions of gene sequences, and, until relatively recently, computational limitations associated with analysing complex molecular evolutionary phenomena in large datasets. In this study, we aimed to remedy the paucity of PLA_2-encoding gene sequences and subject these to rigorous analysis to compare inferences about the evolution of the subfamily with those derived from coding sequence. We examine the contribution that various potential sources of bias make to the understanding of the evolution of Phospholipase A_2 (PLA_2) encoding genes expressed in the venom of viperid snakes. We also specifically address the extent to which the evolution of functional traits can be reconstructed using the transcribed or translated parts of the gene only compared to using information from the entire gene sequence, using newly characterised PLA_2 gene sequences from a phylogenetically diverse set of pitviper taxa.

2. Materials and methods

2.1. Sequencing, alignment and data screening

Samples were collected between 1992 and 2002 as part of a systematic study on Asian pitvipers (Malhotra and Thorpe, 2004; Malhotra et al., 2010) in the form of blood samples or scale clips. For this study, we amplified directly from genomic extracts of representatives of all major genera of Old World pitvipers, and where available, multiple populations per species, using conserved primers located in the untranslated regions of the PLA_2 genes. Individual PCR products were cloned, multiple clones sequenced and similar sequences from individual samples were grouped for detection of PCR errors and construction of consensus sequences as described in Dawson et al. (2010). However, we modified the acceptance criterion in the light of information available from gene expression (proteomic) studies, described below. We reasoned that the minimum number of differences separating two sequences that had confirmed translation products in the venom should set the threshold for determination of those sequences as "real" alleles rather than PCR artefacts, where this was less than the threshold value determined by the binomial function. Note that we did not attempt to determine whether we had completely sampled the alleles present in the individual genomes, as done by Duda and Remigio (2008) and Gibbs and Rossiter (2008). The focus of these two studies was on a much smaller group of species, and repeated sequencing of clones is required to ensure that the rate of discovery of new variants has reached an asymptote. Given limited resources, the focus in this study was to obtain a diversity of gene sequences from a wide range of species across the phylogenetic tree of Asian pitvipers, few of which have been studied previously.

Genomic PLA_2 sequences from four North American crotalines and several viperine outgroups available in GenBank were also included (see Accession numbers in Table S1). The data was partitioned according to the conserved splicing signal (GT/AG at the 5′/3′ ends of the introns respectively) and separated into nine regions corresponding to 3′ and 5′ UTR, exons 1–4, and introns 1–3. Each of these regions was independently aligned using CLUSTALW (Thompson et al., 1994) for non-coding regions (with gap penalties set as in the optimal strategy determined in Creer et al. (2005)) and MUSCLE (Edgar, 2004) for the protein-coding sequences. The aligned amino-acid sequences were subsequently reverse-translated into nucleotide sequence using the RevTrans 1.4 server (http://www.cbs.dtu.dk/services/RevTrans/). Once the aligned regions had been reassembled, they were checked by eye in Jalview v2 (Waterhouse et al., 2009) and corrected manually, since it was clear that similar (or even identical) stretches of sequence in different copies had been aligned differently (particularly in CLUSTALW). The final alignment has been deposited in the Dryad data depository (http://10.5061/dryad.jm7v0).

The aligned data was then searched for signals of recombination. Recombination is a major evolutionary force, which can obscure ancestral-descendant relationships and lead to lower resolution in phylogenetic trees, and can confound evolutionary analyses of selection pressure (Arenas and Posada, 2010; Schierup and Hine, 2000). A number of methods for the detection of recombinant sequences in an alignment have been proposed, but all have their weaknesses. We therefore used a number of methods (RDP, Geneconv, Chimaera, 3SEQ), all implemented in RDP3 (Martin et al., 2010). We began by searching within gene sequences from a single individual, as in this case PCR artefacts could have been produced. Those showing clear evidence of within-species recombination were removed. We then repeated the analysis on the entire dataset, and again weeded out those showing a signal of recombination in at least three of the above methods. Finally, we screened the entire dataset, with obvious recombinants removed, in GARD (Kosakovsky Pond et al., 2006), which relies on detecting phylogenetic congruence between different fragments of the alignment, on the Datamonkey webserver (Delport et al., 2010) of the HyPhy package (Kosakovsky Pond et al., 2005). GARD identifies a number of breakpoints in the alignment; individual recombinant sequences can be identified by comparing their position in the phylogenetic trees from the analysis of different fragments. The appropriate substitution model for each subset was identified using the model selection tool available within HyPhy. The identification of significant breakpoints suggests that the data would be better analysed as separate partitions.

The final aligned dataset contained many indels, which could contain useful phylogenetic information (Creer et al., 2006). However, the presence of gaps in any particular position cannot be treated as independent data (Simmons and Ochoterena, 2000) and