J Mol Evol
DOI 10.1007/s00239-013-9552-5

ORIGINAL ARTICLE

Molecular Phylogeny and Evolution of the Proteins Encoded by Coleoid (Cuttlefish, Octopus, and Squid) Posterior Venom Glands

Tim Ruder · Kartik Sunagar · Eivind A. B. Undheim · Syed A. Ali · Tak-Cheung Wai · Dolyce H. W. Low · Timothy N. W. Jackson · Glenn F. King · Agostinho Antunes · Bryan G. Fry

Received: 18 November 2012 / Accepted: 12 February 2013

Abstract In this study, we report for the first time a detailed evaluation of the phylogenetic history and molecular evolution of the major coleoid toxins: CAP, carboxypeptidase, chitinase, metalloprotease GON-domain, hyaluronidase, pacifastin, PLA2, SE-cephalotoxin and serine proteases, with the carboxypeptidase and GON-domain documented for the first time in the coleoid venom arsenal. We show that although a majority of sites in these coleoid venom-encoding genes have evolved under the regime of negative selection, a very small proportion of sites are influenced by the transient selection pressures. Moreover, nearly 70 % of these episodically adapted sites are confined to the molecular surface, highlighting the importance of variation of the toxin surface chemistry. Coleoid venoms were revealed to be as complex as other venoms that have traditionally been the recipient of the bulk of research efforts. The presence of multiple peptide/protein types in coleoids similar to those present in other animal venoms identifies a convergent strategy, revealing new information as to what characteristics make a peptide/protein type amenable for recruitment into chemical arsenals. Coleoid venoms have significant potential not only for understanding fundamental aspects of venom evolution but also as an untapped source of novel toxins for use in drug design and discovery.

Keywords Molecular evolution · Coleoid · Cephalopod · Octopus · Squid · Cuttlefish · Venom · Positive selection

Introduction

Venoms are key evolutionary innovations under-pinning the explosive radiation of many clades. Research to date has been heavily taxonomically biased, with cone snails, scorpions, snakes and spiders receiving a disproportionate amount of attention. The cephalopods are a conspicuously

Tim Ruder, Kartik Sunagar, Eivind A. B. Undheim, Syed A. Ali are joint first authors.

Electronic supplementary material The online version of this article (doi:10.1007/s00239-013-9552-5) contains supplementary material, which is available to authorized users.

T. Ruder · E. A. B. Undheim · S. A. Ali ·
D. H. W. Low · T. N. W. Jackson · B. G. Fry (✉)
Venom Evolution Lab, School of Biological Sciences,
University of Queensland, St. Lucia, QLD 4072, Australia
e-mail: bgfry@uq.edu.au

K. Sunagar · A. Antunes
CIMAR/CIIMAR, Centro Interdisciplinar de Investigação
Marinha e Ambiental, Universidade do Porto,
Rua dos Bragas, 177, 4050-123 Porto, Portugal

K. Sunagar · A. Antunes
Departamento de Biologia, Faculdade de Ciências, Universidade
do Porto, Rua do Campo Alegre, 4169-007 Porto, Portugal

E. A. B. Undheim · G. F. King
Institute for Molecular Biosciences, University of Queensland,
St. Lucia, QLD 4072, Australia

S. A. Ali
HEJ Research Institute of Chemistry, International Center for
Chemical and Biological Sciences (ICCBS), University of
Karachi, Karachi 75270, Pakistan

T.-C. Wai
Department of Biology and Chemistry, State Key Laboratory in
Marine Pollution, City University of Hong Kong, Tat Chee
Avenue, Kowloon, Hong Kong SAR, China

Published online: 02 March 2013

neglected area of venom research. The group as a whole has been the subject of scant research and even so, strongly biased towards the octopuses, with little attention devoted to other coleoids such as cuttlefish and squid. This narrow taxonomical view complicates the further study of venom molecular evolution (Fry et al. 2003). Recently, we have shown that octopus and cuttlefish share a common, venomous ancestor (Fry et al. 2009), confirming that cephalopods add new protein scaffolds to their arsenals via the duplication of body regulatory proteins and subsequent selective overexpression in the venom gland (Fry 2005; Fry et al. 2009), yet the number of species examined in detail remains low.

Toxicity of octopus saliva from the posterior pair of salivary glands to invertebrates was established as early as 1888 (Lo Bianco 1888). Ghiretti subsequently succeeded in isolating the crab toxic fraction from *Sepia officinalis*, *Octopus vulgaris* and *Octopus macropus*, which he termed cephalotoxin (Ghiretti 1959 1960). Ghiretti also notes that the initial hyper excitability observed in crabs is due to the amides present in coleoid venom, whilst the lethal phase was a result of cephalotoxin. Cariello and Zanetti (1977) isolated five proteins toxic to crabs from *O. vulgaris* PSG homogenate. However due to impurities in their samples only two components were further described: alpha and beta cephalotoxin. Both were found to consist of approximately 50 % carbohydrate, indicating heavy glycosylation. Tachykinins such as Eledoisin (Anastasi and Erspamer 1962), isolated from *Eledone aldrovandi* and *Eledone moschata,* OctTK 1 and 2 isolated from *O. vulgaris* (Kanda et al. 2003) and an OctTK 1 homologue from *Octopus kaurna* (Fry et al. 2009) have been isolated from *octopus*. SE-cephalotoxin, was described by Udea et al. (2008). Fry et al. (2009) described six novel putative toxins with no homology to any known peptide type, in addition to CriSP, and phospholipase A_2, proteins from *Hapalochlaena maculosa*, *O. kaurna* and *Sepia latimanus* (Fry et al. 2009). Enzymes also play key roles in coleoid toxicity in addition to the small organic molecules, peptides, and non-enzymatic proteins. Indeed, large amounts of S_1 peptidase gene transcripts from both *H. maculosa* and *O. kaurna* have been recently identified (Fry et al. 2009). Other studies also associate proteolytic activity with PSG extract of both *Eledone cirrhosa* (Grisley 1993; Grisley and Boyle 1987) and *O. vulgaris* (Morishita 1974). Hyaluronidase and Chitinase have also been identified from octopus venom (Fry et al. 2009; Grisley and Boyle 1990; Romanini 1952). Further evidence of functional diversification is that enzymes from the venoms of octopus species living in Antarctica have sub-zero temperature optimum efficiency, with decreased activity at higher temperatures (Undheim et al. 2010).

Whilst some toxic secretions of coleoids have been previously investigated, the literature to date shows high toxin diversity and offers little insight into the evolutionary aspects of coleoid toxicity. This leaves questions regarding the genetic origin and the strategies of venom recruitment of coleoid toxins entirely up to speculation. In this study, we present for the first time a glimpse into the coleoid venom phylogenetic history and molecular evolution. We report a particularly detailed evaluation of the major coleoid toxins (PLA_2, CAP, pacifastin, and serine proteases) and highlight the prominent role of episodic diversifying selections in shaping some of these toxins.

Materials and Methods

Tissue Sampling and Taxon Selection

Posterior secretory glands were dissected from freshly euthanized specimens collected from tropical to polar waters, thus providing wide taxiconomical and ecological coverage: cuttlefish species included were *S. latimanus* (Osprey Reef, Coral Sea) and *Sepia pharaonis* (Hong Kong); octopus species include *Abdopus aculeatus* (Lizard Island, Queensland, Australia), *Adeleiledone polymorpha* [East Antarctica coastal waters off of George V's Land (1398E to 1458E)], *H. maculosa* (Mornington Peninsula, Victoria Australia), *O. cyanea* (Lizard Island, Queensland, Australia), *O. kaurna* (Mornington Peninsula, Victoria, Australia) and *Pareledone turqueti* [East Antarctica coastal waters off of George V's Land (1398E to 1458E)]; squid species included *Loliolus noctiluca* (Moreton Bay, Queensland, Australia) and *Sepioteuthis australis* (Moreton Bay, Queensland, Australia).

cDNA Library Construction and Analysis

Total RNA extracted using the standard TRIzol Plus method (Invitrogen). Extracts were enriched for mRNA using standard RNeasy mRNA mini kit (Qiagen) protocol. mRNA was reverse transcribed, fragmented, and ligated to a unique 10-base multiplex identifier (MID) tag prepared using standard protocols and applied to one PicoTitrePlate (PTP) for simultaneous amplification and sequencing on a Roche 454 GS FLX+ Titanium platform (Australian Genome Research Facility). 50,000 sequences for each sample were read. Automated grouping and analysis of sample-specific MID reads informatically separated sequences from the other transcriptomes on the plates, which were then post-processed to remove low quality sequences before de novo assembly into contiguous sequences (contigs) using v 3.4.0.1 of the MIRA software program. Assembled contigs were processed using CLC Main Work Bench (CLC-Bio) and Blast2GO bioinformatic suite (Gotz et al. 2011, 2008) to provide Gene Ontology, BLAST and domain/Interpro annotation. The above analyses assisted in the rationalisation of the large numbers of assembled contigs into phylogenetic 'groups' for the detailed phylogenetic analyses outlined below.

Sequences are available from Genbank with the Bioproject and Biosample retrieval numbers of: *A. aculeatus* PRJNA 188569 SAMN01911391, *A. polymorpha* PRJNA188570 SAMN01911392, *H. maculosa* PRJNA188571 SAMN 01911430, *L. noctiluca* PRJNA188572 SAMN01911444, *O. cyanea* PRJNA188574 SAMN01911445, *O. kaurna* PRJNA188658 SAMN01911449, *P. turqueti* PRJNA188 575 SAMN01911446, *S. latimanus* PRJNA188659 SAMN 01911447, *S. pharaonis* PRJNA188577 SAMN01911448 and *S. australis* PRJNA188576 SAMN01911450.

Bioinformatics

Phylogenetics

Phylogenetic analyses of the bioinformatically recovered transcripts were performed to allow reconstruction of the molecular evolutionary history of each toxin type. Toxin sequences were identified by comparison of the translated DNA sequences with previously characterised toxins using a BLAST search (Altschul et al. 1997) of the UniProtKB protein database. Molecular phylogenetic analyses of toxin transcripts were conducted using the translated amino acid sequences. Comparative sequences from physiological gene homologues identified from non-venom tissues were included in each dataset as outgroup sequences. To minimize confusion, all sequences obtained in this study are referred by their Genbank accession numbers (http://www.ncbi.nlm.nih.gov/sites/entrez?db=Nucleotide) and sequences from previous studies are labelled with their UniProtKB accession numbers (http://www.expasy.org/cgi-bin/sprot-search-ful). Resultant sequence sets were aligned using CLC Mainbench. When presented as sequence alignments, the leader sequence (as identified through use of SignalP) is shown in lowercase and cysteines are highlighted in black. > and < indicate incomplete N/5′ or C/3′ ends, respectively. Datasets were analysed using Bayesian inference implemented on MrBayes, version 3.0b4 (Ronquist and Huelsenbeck 2003). Two different run conditions were used to test for congruence: lset rates = invgamma with prset aamodelpr = fixed (WAG) and lset rates = gamma with prset aamodelpr = mixed. The analysis was performed by running a minimum of 1×10^7 generations in four chains, and saving every 100th tree. The log-likelihood score of each saved tree was plotted against the number of generations to establish the point at which the log-likelihood scores reached their asymptote, and the posterior probabilities for clades established by constructing a majority-rule consensus tree for all trees generated after completion of the burn-in phase. Trees shown are invgamma with WAG, which are identical in topology to gamma with mixed.

Test for Recombination

To overcome the effects of recombination on phylogenetic and evolutionary interpretations (Posada and Crandall 2002), we employed GARD and Single Breakpoint algorithms implemented in the HyPhy package and assessed recombination on all the toxin forms examined in this study (Delport et al. 2010; Kosakovsky Pond et al. 2006). When potential breakpoints were detected using the small sample Akaike information criterion (AICc), the sequences were compartmentalized before conducting the selection analyses.

Selection Analysis

We evaluated selection pressures using maximum-likelihood models (Goldman and Yang 1994; Yang 1998) implemented in CODEML of the PAML (Yang 2007). We first employed the one-ratio model that assumes a single ω for the entire phylogenetic tree. This model tends to be very conservative and can only detect positive selection if the ω ratio averaged over all the sites along the lineage is significantly >1. Because such lineage-specific models assume a single ω for the entire tree, they often fail to identify regions in proteins that might be affected by episodic selection pressures and ultimately, underestimate the strength of selection. Hence, we employed site-specific models which estimate positive selection statistically as a non-synonymous-to-synonymous nucleotide-substitution rate ratio (ω) significantly >1. We compared likelihood values for three pairs of models with different assumed ω distributions as no a priori expectation exists for the same: M0 (constant ω rates across all sites) versus M3 (allows the ω to vary across sites within 'n' discrete categories, $n \geq 3$); M1a (a model of neutral evolution) where all sites are assumed to be either under negative ($\omega < 1$) or neutral selection ($\omega = 1$) versus M2a (a model of positive selection) which in addition to the site classes mentioned for M1a, assumes a third category of sites; sites with $\omega > 1$ (positive selection) and M7 (Beta) versus M8 (Beta and ω), and models that mirror the evolutionary constraints of M1 and M2 but assume that ω values are drawn from a beta distribution (Nielsen and Yang 1998). Only if the alternative models (M3, M2a and M8: allow sites with $\omega > 1$) show a better fit in likelihood ratio test (LRT) relative to their null models (M0, M1a and M8: do not show allow sites $\omega > 1$), are their results considered significant. LRT is estimated as twice the difference in maximum-likelihood values between nested models and compared with the c^2 distribution with the appropriate degree of freedom—the difference in the number of parameters between the two models. The Bayes empirical Bayes (BEB) approach (Yang et al. 2005) was used to identify amino acids under positive selection by calculating the posterior probabilities that a

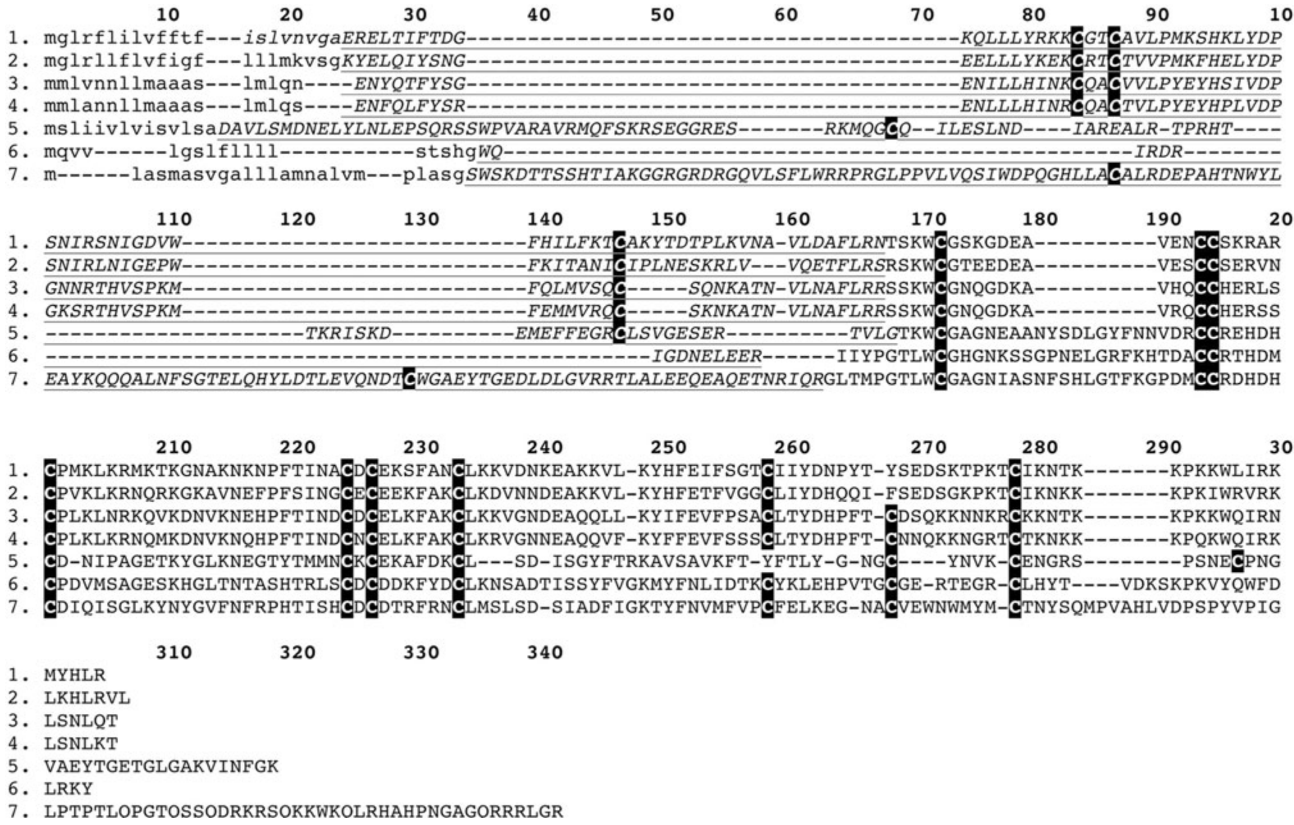
Table 1 Toxin types recovered from each species by transcriptome surveying

Species	CAP	CARB	CHIT	GON	HYAL	PACI	PLA2	SE-C	SP
Abdopus aculeatus			X	X		X			X
Adelieledone polymorpha			X						X
Hapalochlaena maculosa		X	X		X	X			X
Loliolus noctiluca	X	X	X	X		X	X		X
Octopus cyanea	X	X	X		X	X			X
Octopus kaurna	X	X	X			X			X
Pareledone turqueti			X						X
Sepia latimanus	X		X				X	X	X
Sepia pharaonis	X		X	X		X	X	X	X
Sepioteuthis australis	X		X	X		X	X	X	X

CAP CRiSP/Allergen/PR-1, *CARB* carboxypeptidease, *CHIT* chitinase, *GON* metalloprotease GON-domain, *HYAL* hyaluronidase, *PACI* pacifastin, *PLA2* phospholipase A2, *SE-C* SE-cephalotoxin and *SP* serine protease

particular amino acid belongs to a given selection class (neutral, conserved or highly variable). Sites with greater posterior probability (*PP* ≥ 95 %) of belonging to the 'ω > 1 class' were inferred to be positively selected.

Single Likelihood Ancestor Counting (SLAC), fixed-effects likelihood (FEL) and random-effects likelihood models (Kosakovsky Pond et al. 2005) implemented in HyPhy (Kosakovsky Pond et al. 2005) were employed to provide additional support to the aforementioned analyses and to detect sites evolving under the influence of positive and negative selection. Mixed Effects Model Evolution (MEME) (Kosakovsky Pond et al. 2011) was also used to detect episodic diversifying selection. Since, the three domains of the coleoid pacifastin gene are expressed as



```
              10        20        30        40        50        60        70        80        90       100
1. mglrflilvfftf---islvnvgaERELTIFTDG----------------------------------------KQLLLYRKKCGTCAVLPMKSHKLYDP
2. mglrllflvfigf---lllmkvsgKYELQIYSNG----------------------------------------EELLLYKEKCRTCTVVPMKFHELYDP
3. mmlvnnllmaaas---lmlqn----ENYQTFYSG----------------------------------------ENILLHINKCQACVVLPYEYHSIVDP
4. mmlannllmaaas---lmlqs----ENFQLFYSR----------------------------------------ENLLLHINRCQACTVLPYEYHPLVDP
5. msliivlvisvlsaDAVLSMDNELYLNLEPSQRSSWPVARAVRMQFSKRSEGGRES-------RKMQGCQ--ILESLND----IAREALR-TPRHT----
6. mqvv------lgslfllll----------stshgWQ----------------------------------------------------IRDR--------
7. m------lasmasvgalllamnalvm---plasgSWSKDTTSSHTIAKGGRGRDRGQVLSFLWRRPRGLPPVLVQSIWDPQGHLLACALRDEPAHTNWYL

             110       120       130       140       150       160       170       180       190       200
1. SNIRSNIGDVW----------------------------FHILFKTCAKYTDTPLKVNA-VLDAFLRNTSKWCGSKGDEA----------VENCCSKRAR
2. SNIRLNIGEPW----------------------------FKITANICIPLNESKRLV---VQETFLRSRSKWCGTEEDEA----------VESCCSERVN
3. GNNRTHVSPKM----------------------------FQLMVSQC-----SQNKATN-VLNAFLRRSSKWCGNQGDKA----------VHQCCHERLS
4. GKSRTHVSPKM----------------------------FEMMVRQC-----SKNKATN-VLNAFLRRSSKWCGNQGDKA----------VRQCCHERSS
5. ---------------------TKRISKD----------EMEFFEGRCLSVGESER----------TVLGTKWCGAGNEAANYSDLGYFNNVDRCCREHDH
6. -------------------------------------------------IGDNELEER------IIYPGTLWCGHGNKSSGPNELGRFKHTDACCRTHDM
7. EAYKQQQALNFSGTELQHYLDTLEVQNDTCWGAEYTGEDLDLGVRRTLALEEQEAQETNRIQRGLTMPGTLWCGAGNIASNFSHLGTFKGPDMCCRDHDH

             210       220       230       240       250       260       270       280       290       300
1. CPMKLKRMKTKGNAKNKNPFTINACDCEKSFANCLKKVDNKEAKKVL-KYHFEIFSGTCIIYDNPYT-YSEDSKTPKTCIKNTK-------KPKKWLIRK
2. CPVKLKRNQRKGKAVNEFPFSINGCECEEKFAKCLKDVNNDEAKKVL-KYHFETFVGGCLIYDHQQI-FSEDSGKPKTCIKNKK-------KPKIWRVRK
3. CPLKLNRKQVKDNVKNEHPFTINDCDCELKFAKCLKKVGNDEAQQLL-KYIFEVFPSACLTYDHPFT-CDSQKKNNKRCKKNTK-------KPKKWQIRN
4. CPLKLKRNQMKDNVKNQHPFTINDCNCELKFAKCLKRVGNNEAQQVF-KYFFEVFSSSCLTYDHPFT-CNNQKKNGRTCTKNKK-------KPQKWQIRK
5. CD-NIPAGETKYGLKNEGTYTMMNCKCEKAFDKCL---SD-ISGYFTRKAVSAVKFT-YFTLY-G-NGC----YNVK-CENGRS--------PSNECPNG
6. CPDVMSAGESKHGLTNTASHTRLSCDCDDKFYDCLKNSADTISSYFVGKMYFNLIDTKCYKLEHPVTGCGE-RTEGR-CLHYT----VDKSKPKVYQWFD
7. CDIQISGLKYNYGVFNFRPHTISHCDCDTRFRNCLMSLSD-SIADFIGKTYFNVMFVPCFELKEG-NACVEWNWMYM-CTNYSQMPVAHLVDPSPYVPIG

             310       320       330       340
1. MYHLR
2. LKHLRVL
3. LSNLQT
4. LSNLKT
5. VAEYTGETGLGAKVINFGK
6. LRKY
7. LPTPTLOPGTOSSODRKRSOKKWKOLRHAHPNGAGORRRLGR
```

Fig. 1 Sequence alignment of venom type III PLA_2 precursors from the coleoids (*1*) *Sepioteuthis australis* PLA2-SepT-1, (*2*) *Loliolus noctiluca* PLA2-Lol-2, (*3*) *Sepia pharaonis* PLA2-Sepea-1, (*4*) *Sepia latimanus* (B6Z1Y5), (*5*) the scorpion *Hadrurus gertschi* (P0C8L9), (*6*) the bee *Apis mellifera* (P00630) and (*7*) the lizard *Abronia graminea* (RL8c9). Propeptide sequence is shown *underlined*

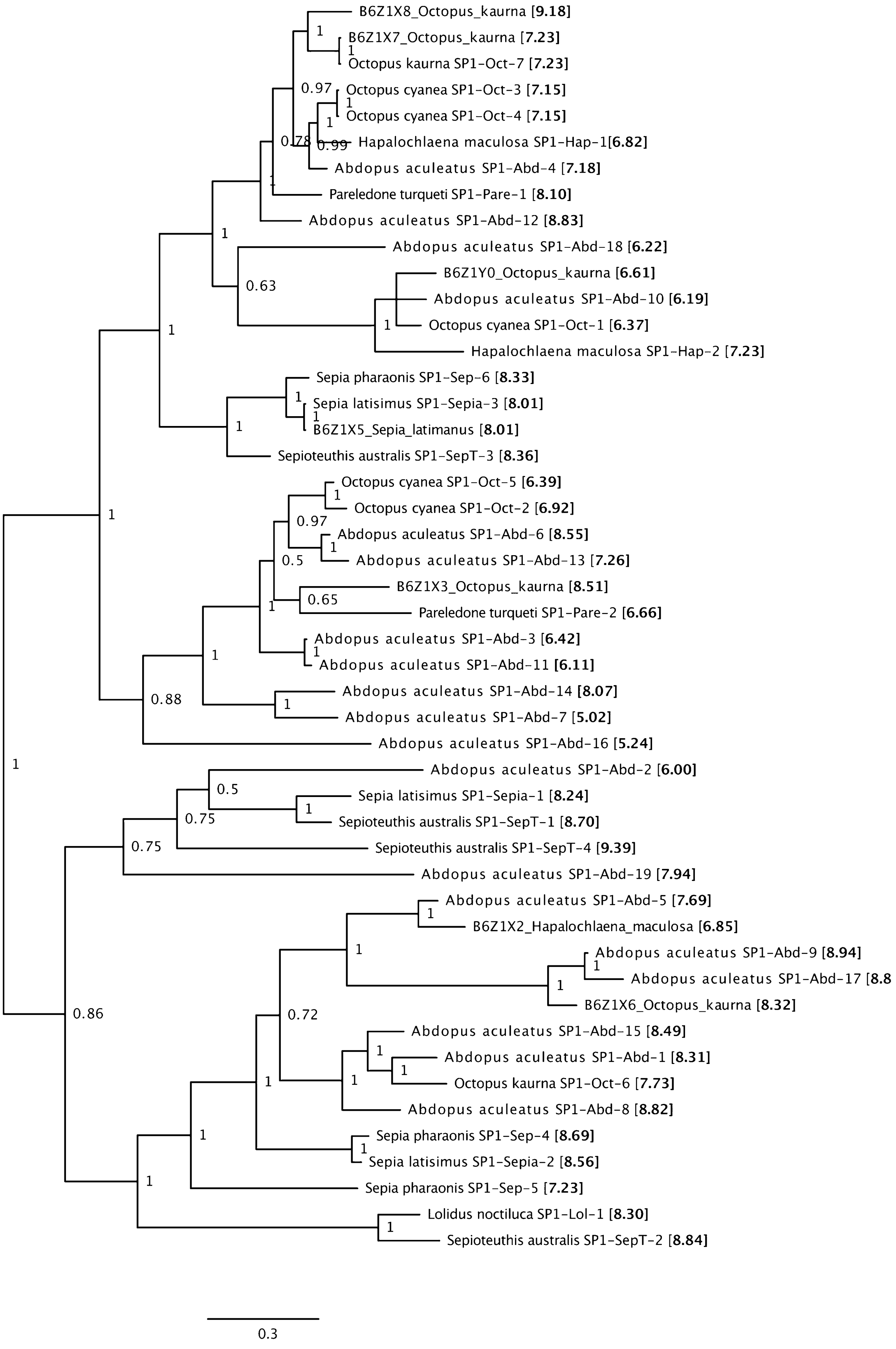

B6Z1X8_Octopus_kaurna [9.18]
B6Z1X7_Octopus_kaurna [7.23]
Octopus kaurna SP1-Oct-7 [7.23]
Octopus cyanea SP1-Oct-3 [7.15]
Octopus cyanea SP1-Oct-4 [7.15]
Hapalochlaena maculosa SP1-Hap-1[6.82]
Abdopus aculeatus SP1-Abd-4 [7.18]
Pareledone turqueti SP1-Pare-1 [8.10]
Abdopus aculeatus SP1-Abd-12 [8.83]
Abdopus aculeatus SP1-Abd-18 [6.22]
B6Z1Y0_Octopus_kaurna [6.61]
Abdopus aculeatus SP1-Abd-10 [6.19]
Octopus cyanea SP1-Oct-1 [6.37]
Hapalochlaena maculosa SP1-Hap-2 [7.23]
Sepia pharaonis SP1-Sep-6 [8.33]
Sepia latisimus SP1-Sepia-3 [8.01]
B6Z1X5_Sepia_latimanus [8.01]
Sepioteuthis australis SP1-SepT-3 [8.36]
Octopus cyanea SP1-Oct-5 [6.39]
Octopus cyanea SP1-Oct-2 [6.92]
Abdopus aculeatus SP1-Abd-6 [8.55]
Abdopus aculeatus SP1-Abd-13 [7.26]
B6Z1X3_Octopus_kaurna [8.51]
Pareledone turqueti SP1-Pare-2 [6.66]
Abdopus aculeatus SP1-Abd-3 [6.42]
Abdopus aculeatus SP1-Abd-11 [6.11]
Abdopus aculeatus SP1-Abd-14 [8.07]
Abdopus aculeatus SP1-Abd-7 [5.02]
Abdopus aculeatus SP1-Abd-16 [5.24]
Abdopus aculeatus SP1-Abd-2 [6.00]
Sepia latisimus SP1-Sepia-1 [8.24]
Sepioteuthis australis SP1-SepT-1 [8.70]
Sepioteuthis australis SP1-SepT-4 [9.39]
Abdopus aculeatus SP1-Abd-19 [7.94]
Abdopus aculeatus SP1-Abd-5 [7.69]
B6Z1X2_Hapalochlaena_maculosa [6.85]
Abdopus aculeatus SP1-Abd-9 [8.94]
Abdopus aculeatus SP1-Abd-17 [8.8
B6Z1X6_Octopus_kaurna [8.32]
Abdopus aculeatus SP1-Abd-15 [8.49]
Abdopus aculeatus SP1-Abd-1 [8.31]
Octopus kaurna SP1-Oct-6 [7.73]
Abdopus aculeatus SP1-Abd-8 [8.82]
Sepia pharaonis SP1-Sep-4 [8.69]
Sepia latisimus SP1-Sepia-2 [8.56]
Sepia pharaonis SP1-Sep-5 [7.23]
Lolidus noctiluca SP1-Lol-1 [8.30]
Sepioteuthis australis SP1-SepT-2 [8.84]
1
1
0.97
1
1
0.78
0.99
1
1
0.63
1
1
1
1
1
1
1
0.97
1
0.5
0.65
1
1
1
1
0.88
1
0.5
1
0.75
0.75
1
1
1
1
0.72
1
1
1
1
1
1
1
0.86
0.3

Fig. 2 Phylogenetic reconstruction of coleoid serine proteases. Bracket values indicate calculate pI

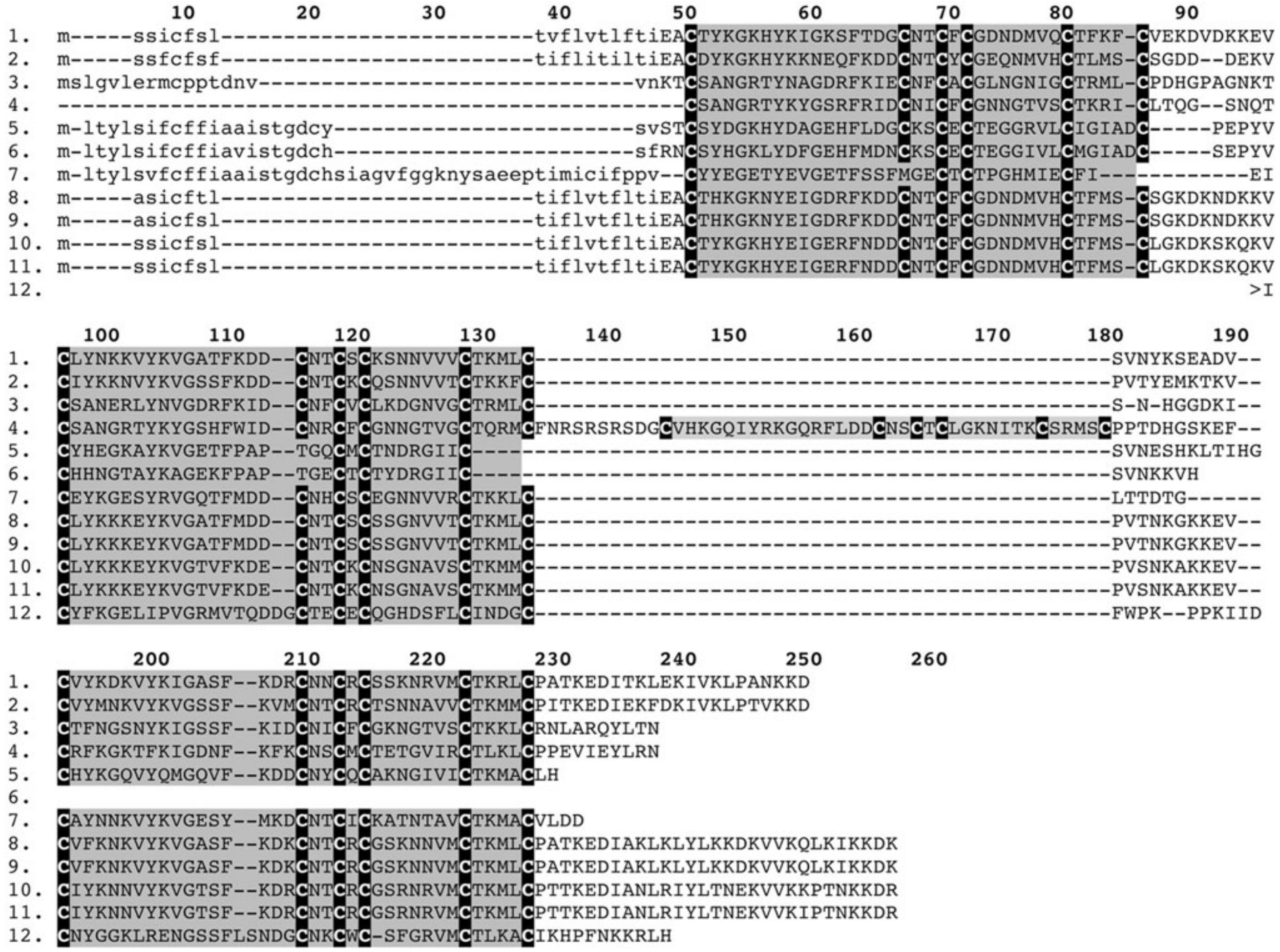
```
           10        20        30        40        50        60        70        80        90       100
1. mkvli-aflfclslaesFNYTN-------YYLLKVKPQTSKGFDFLKNLEAKHPFDYDFWIPPSKLRKNAEVLMPQSAYNNIKDYLKESDVQVT-ILSKN
2. mqpfiwllligvaqvgaKTYSN-------YHLFRVTPKNQQQLTGLKDMFEDEDSEVDFWIAPSSMNRTTEFLVPPHFVSSVKDLLDSLEAETL-VLHED
3. mkplletlyllgmlvpgglgYDRSLAQH------------------------------------------------------------------------
4. mr----ffllmaviyttlaIAPVHFDREKVFRVKLQNEKHASVLKNLTQSIELDFWYPDAIHDIAVNMTVDFRVSEKESQTIQSTLEQHKIHYE-ILIHD

          110       120       130       140       150       160       170       180       190       200
1. LQSDMDKEKDA----DRSV---ATTSTTDKYLFIHSINRLIKS-FAEAYDYVSVKNYGKSHEKRDLYAVKFSVGPGRK---TIIVETGMHGRDWIAIAST
2. IQTIIDEESTSPSKNDSQFTSYSTGSITDKYASYEEIVNWMKV-LAKTYDHAKLIEFGYTFERRKLYAIKFSTGPNKK---GIFVESGMHSREWISIASS
3. RQEIVDKSVSPWSLE----T-----YSYNIYHPMGEIYEWMREISEKYKEVVTQHFLGVTYETHPMYYLKI-SQPSGNPKKIIWMDCGIHAREWIAPAFC
4. LQEEIEKQF--DVKD--EIAGR---HSYAKYNDWDKIVSWTEKMLEKHPEMVSRIKIGSTVEDNPLYVLKI-GKKDG-ERKAIFMDCGIHAREWISPAFC

          210       220       230       240       250       260       270       280       290       300
1. LKLIEYMATKYKTDLDVKIMLNNFDWLFIPVANPDGYVVTYSQDRLWKKNMKRDPSGTGCIGVDLNRNFNAKWGKEGSSGDPCNRAYHGRNAFSEPETIA
2. VKIIERMATQYKSDKEIDELLALYDWTFVPYSNPDGYYYTQHFDRMWRKN--RRPINLFCTGTDINRNFASGWGGPGASRNPCNPTFRGTSVFSEPESRA
3. QWFVKEILQNHKDNSSIRKLLRNLDFYVLPVLNIDGYIYTWTTDRLWRKSRSPHNN-GTCFGTDLNRNFNASWCSIGASRNCQDQTFCGTGPVSEPETKA
4. QWFVYQATKSYGKNKIMTKLLDRMNFYVLPVFNVDGYIWSWTQDRMWRKNRSRNQN-STCIGTDLNRNFDVSWDSSPNTNKPCLNVYRGPAPESEKETKA

          310       320       330       340       350       360       370       380       390       400
1. LSRLVRSTRDK--LAYFSVHAYSQYILTPYACKSRKPKNSEHLMEVAEKVKEAIYEENESRYFVGSPPDILYPFTGSSYDWAKMEMDIKYAYTLKLGPPA
2. LAQLVKQSAPL--VGFFSVHAYSQLILTPYGYTYEKPSDSPELTQLAIKAATAMKKVDGKQFTVGTPPDILYIATGGAYDWAKLKMNIKYAYAFELRPSQ
3. VASFIESKKDDIL-CFLTMHSYGQLILTPYGYTKNKSSNHPEMIQVGQKAANALKAKYGTNYRVGSSADILYASSGSSRDWARDI-GIPFSYTFELRDSG
4. VTNFIRSHLNSIK-AYITFHSYSQMLLIPYGYTFKLPPNHQDLLKVARIATDALSTRYETRYIYGPIASTIYKTSGSSLDWVYDL-GIKHTFAFELRDKG

          410       420       430       440       450       460
1. FNGKGYIVPESQIEANFRELYIALKTFAGNLEK
2. NGRNGFIIPKRNIKPSSLELFAALKTFAKGF-R
3. TY--GFVLPEAQIQPTCEETMEAVLSVLDDVYAKHWHSDSAGRVTSATMLLGLLVSCMSLL
4. KS--GFLLPESRIKPTCKETMLSVKFIAKYILKNTS
```

Fig. 3 Sequence alignment of carboxypeptidase precursors from coleoid venom glands (*1*) *Octopus cyanea* Carb-Oct-1 and (*2*) *Sepioteuthis australis* Carb-SepT-1 and related non-venom sequences from (*3*) *Homo sapiens* (Q8IVL8) and (*4*) *Mus musculus* (P15089). Propeptide sequence is shown *underlined*

```
           10        20        30        40        50        60        70        80        90       100       110       120
1. mkwlvsllvl-------lnsvrlqdaGALEGALRKFVQV-NSAALNTVKESWGGDHVTEEQCAEECLRRPDCASFEMQDTN--CFLSDKDSTGRATVRSNPNRNFFQRIRPNQEVIILKK
2. mk--vflvlfpliwsattgaGDHGSVRTSDSGVLSYFVHIESSTINDPNTKT-KDHVSLKKCAKLCLDDVTCVSFEHDKEKQICFLS-RNSANKYKIQFDANRDYYQAIIPDEALIGMKG
3. mkmkvflvlfpliwsvitnaDDTGXVRXXXT-----------------------------------------------------------------------------------------
4. m----ilvflpliwsgtvnAEDTVLANT--------------------------------------------------------------------------------------------
5. mvaianvlmlgtlysflanpvqcFTPSDVQS---HFVKIRNTILVLSTKAMWGDENVAPEDCAAKCWSDKRCLSFEITKNKGRCFLSKENAATFREVRRAKGRDYYQRVKPNEKLIIFKK

          130       140       150       160       170       180       190       200       210       220       230       240
1. VTIPAEDHLGLVKRVSVSECTKKCRENPHCTSFEFKETARKCDLSNVTHHTHDLHPSKWGWDIYIFNPDYKPIDCGKPPEIIHATVEFEHTDYLANATYKCETGETYTVVCEENNKWSHL
2. VGVRDTIKLERYRKLSLDECSAKCRANPKCKSLAYKPRTKLCNLNGATYLNQNLDLGKWGWDMYFVNQDQKEINCGSPPRLHYKSVTYVATIYKSRAEYRCGLGETFISTCGIDGKWS--
3. --------------------------------------------------------------------DQKEVNCGSPPRLGFKFVTYETTIYKSQAEYRCGMGETYISTCGIDGKWS--
4. --------------------------------------------------------------------DKKEVNCGSPPRLDFKMVSYITTNYKSQAEYRCGLGETFLSTCGIDGKWS--
5. GSIQGFDNLGHYKNVSLKKCKQNCRKHPECNSFEYKEETKLCDLSNVTHLTNEIKSNRWGWDIYIFNPGYRKIDCGTPKDLARALKVYSSTTYESEVTYICPMGEKLTSTCEMNNEW---

          250       260       270       280       290       300       310       320       330       340       350       360
1. VDVCKKPLDCGRPPLLGYAGVDYKETSLKSTATYSCLMGEKYTVVCGANGKWTMPKGRCRNPRSCEEFKKCGRHIPDGEYWIYPKILNGERVKVFCQGMATHHPKEYITLKKENFALASP
2. -----------------------------------------------------KPNGTCKRPTTCKEFGKCGEFVPDGEYWIYPAGLYGKKLKVYCSGMQTANPKEYVSLLSENYSLAPA
3. -----------------------------------------------------KPNGTCKKPSTCDEFRRCGEWPTDGEYWIYPAALHGQKMKVYCSGMYTAKPKEYISLFQENYSFAPA
4. -----------------------------------------------------KPNGTCKRPTTCKEFKKCGEIVTDGEYWIYPTGLHGKKLKVYCSGMQTEHPKEYISLMRENFSFYPA
5. -------------------------TSVEST---------------------------CKEPVSCKDIKMCGRDATDGEYWIYPSPLSHKRLKVYCHDMETDKPKEFITLPHENFAFAPA

          370       380       390       400       410       420       430       440       450       460       470       480
1. LKSIGIFCLSRTALGRGRPGKSIYHKVRLLPERMVLQNTDYKYVNLVDGFGSFKHDFATAEDCLFRNDDRCNSQGAFQVDLIGTGLAIDKSVKWESYGYYSRVQSLNRSHNDQKIRGVCG
2. KKVL-FFCFTRTSLDSYQPGKTIFYKIRINPENLKVNRTDFTFAKLAPDSGNTKWAYGVAADCLYDKESSCESEGKMKIDLTGTGLAVDRRGSWTNFGTNSRMVNFKRSENGQQVSATCG
3. EKVW-FFCYTRSALNVHQPGKTVFSKIRIDPTNLKVNRSDFTFAKLLPKSGNTTWSYGTGGDCLYDKDSRCDSQGKFKIDLTGTGLAVDRRVSWTNFGTNSRMVDFKRSENGQQVTATCG
4. KKPW-FFCVTRSNLEWQHPGKTLFKKIRINPENLKVNRTDFQFAELVPKSGKARWSYGVAADCLFEKGSSCKSQGEMRIDLTGTGLAVDRRVSWTNFGTNSRIAEFKRLKNGQVVTARCG
5. LKP-SYLCMTRFDEEISKPGKTIFKKVRIYPEEMVLDRFDFTFTELEKGKGKHKRGFGAAEDCLYGKSSYCKSKGEMKVDLTHTGLIIGDSVKWVTYGDDSRISDLTYSENKKVVTATCG

          490       500       510       520
1. GTCGGCRPNGPLKVNIFSDDQ-PNTISAEFCQEM
2. GKCGGCKP---LCIKLVATRTGPPEESAKTCENV
3. GKCGGCKP---XCLKLVASKTVPPEETAGTCESV
4. GHCGGCKP---LCIKLIAAKTGFVENPARECEEL
5. GKCGGCKPEGKFTVQL-SSDE-PSSSGVKTCEFLRQKHCSKSAQAQ
```

Fig. 4 Sequence alignment of GON-domain sequences from coleoid venom glands (*1*) *Sepioteuthis australis* GON-SepT-1, (*2*) *Sepioteuthis australis* GON-SepT-2, (*3*) *Loliolus noctiluca* GON-Lol-1, (*4*) *Sepia pharaonis* GON-Sepea-1 and (*5*) *Abdopus aculeatus* GON-Abd-1

multiple products, we also assessed the selection pressures influencing them simultaneously using Mgene (4) and option G test (Yang 1996) from Codeml. Further support for the results of the selection analyses was obtained using a complementary protein-level approach implemented in TreeSAAP (Woolley et al. 2003).



```
            10        20        30        40        50        60        70        80        90
1.  m-----ssicfsl------------------------tvflvtlftiEACTYKGKHYKIGKSFTDGCNTCFCGDNDMVQCTFKF-CVEKDVDKKEV
2.  m-----ssfcfsf------------------------tiflitiltiEACDYKGKHYKKNEQFKDDCNTCYCGEQNMVHCTLMS-CSGDD--DEKV
3.  mslgvlermcpptdnv-----------------------------vnKTCSANGRTYNAGDRFKIECNFCACGLNGNIGCTRML-CPDHGPAGNKT
4.  -------------------------------------------------CSANGRTYKYGSRFRIDCNICFCGNNGTVSCTKRI-CLTQG--SNQT
5.  m-ltylsifcffiaaistgdcy-----------------------svSTCSYDGKHYDAGEHFLDGCKSCECTEGGRVLCIGIADC-----PEPYV
6.  m-ltylsifcffiavistgdch-----------------------sfRNCSYHGKLYDFGEHFMDNCKSCECTEGGIVLCMGIADC-----SEPYV
7.  m-ltylsvfcffiaaistgdchsiagvfggknysaeeptimicifppv--CYYEGETYEVGETFSSFMGECTCTPGHMIECFI------------EI
8.  m-----asicftl------------------------tiflvtfltiEACTHKGKNYEIGDRFKDDCNTCFCGDNDMVHCTFMS-CSGKDKNDKKV
9.  m-----asicfsl------------------------tiflvtfltiEACTHKGKNYEIGDRFKDDCNTCFCGDNNMVHCTFMS-CSGKDKNDKKV
10. m-----ssicfsl------------------------tiflvtfltiEACTYKGKHYEIGERFNDDCNTCFCGDNDMVHCTFMS-CLGKDKSKQKV
11. m-----ssicfsl------------------------tiflvtfltiEACTYKGKHYEIGERFNDDCNTCFCGDNDMVHCTFMS-CLGKDKSKQKV
12.                                                                                               >I

     100       110       120       130       140       150       160       170       180       190
1.  CLYNKKVYKVGATFKDD--CNTCSCKSNNVVVCTKMLC----------------------------------------------SVNYKSEADV--
2.  CIYKKNVYKVGSSFKDD--CNTCKCQSNNVVTCTKKFC----------------------------------------------PVTYEMKTKV--
3.  CSANERLYNVGDRFKID--CNFCVCLKDGNVGCTRMLC----------------------------------------------S-N-HGGDKI--
4.  CSANGRTYKYGSHFWID--CNRCFCGNNGTVGCTQRMCFNRSRSRSDGCVHKGQIYRKGQRFLDDCNSCTCLGKNITKCSRMSCPPTDHGSKEF--
5.  CYHEGKAYKVGETFPAP--TGQCMCTNDRGIIC---------------------------------------------------SVNESHKLTIHG
6.  CHHNGTAYKAGEKFPAP--TGECTCTYDRGIIC---------------------------------------------------SVNKKVH
7.  CEYKGESYRVGQTFMDD--CNHCSCEGNNVVRCTKKLC----------------------------------------------LTTDTG------
8.  CLYKKKEYKVGATFMDD--CNTCSCSSGNVVTCTKMLC----------------------------------------------PVTNKGKKEV--
9.  CLYKKKEYKVGATFMDD--CNTCSCSSGNVVTCTKMLC----------------------------------------------PVTNKGKKEV--
10. CLYKKKEYKVGTVFKDE--CNTCKCNSGNAVSCTKMMC----------------------------------------------PVSNKAKKEV--
11. CLYKKKEYKVGTVFKDE--CNTCKCNSGNAVSCTKMMC----------------------------------------------PVSNKAKKEV--
12. CYFKGELIPVGRMVTQDDGCTECECQGHDSFLCINDGC----------------------------------------------FWPK--PPKIID

         200       210       220       230       240       250       260
1.  CVYKDKVYKIGASF--KDRCNNCRCSSKNRVMCTKRLCPATKEDITKLEKIVKLPANKKD
2.  CVYMNKVYKVGSSF--KVMCNTCRCTSNNAVVCTKMMCPITKEDIEKFDKIVKLPTVKKD
3.  CTFNGSNYKIGSSF--KIDCNICFCGKNGTVSCTKKLCRNLARQYLTN
4.  CRFKGKTFKIGDNF--KFKCNSCMCTETGVIRCTLKLCPPEVIEYLRN
5.  CHYKGQVYQMGQVF--KDDCNYCQCAKNGIVICTKMACLH
6.
7.  CAYNNKVYKVGESY--MKDCNTCICKATNTAVCTKMACVLDD
8.  CVFKNKVYKVGASF--KDKCNTCRCGSKNNVMCTKMLCPATKEDIAKLKLYLKKDKVVKQLKIKKDK
9.  CVFKNKVYKVGASF--KDKCNTCRCGSKNNVMCTKMLCPATKEDIAKLKLYLKKDKVVKQLKIKKDK
10. CIYKNNVYKVGTSF--KDRCNTCRCGSRNRVMCTKMLCPTTKEDIANLRIYLTNEKVVKKPTNKKDR
11. CIYKNNVYKVGTSF--KDRCNTCRCGSRNRVMCTKMLCPTTKEDIANLRIYLTNEKVVKIPTNKKDR
12. CNYGGKLRENGSSFLSNDGCNKCWC-SFGRVMCTLKACIKHPFNKKRLH
```

Fig. 5 Sequence alignment of pacifastin peptide precursors from coleoid venom glands (*1*) *Hapalochlaena maculosa* Paci-Hap-1, (*2*) *Octopus kaurna* Paci-Oct-1, (*3*) *Loliolus noctiluca* Paci-Lol-1, (*4*) *Sepia pharaonis* Paci-Sepea-1, (*5*) *Abdopus aculeatus* Paci-Abd-1, (*6*) *H. maculosa* (B6Z1Z0), (*7*) *Octopus cyanea* Paci-Oct-2, (*8*) *Abdopus aculeatus* Paci-Abd-2, (*9*) *Abdopus aculeatus* Paci-Abd-3, (*10*) *Octopus cyanea* Paci-Oct-3, (*11*) *Octopus cyanea* Paci-Oct-4 and (*12*) *Sepioteuthis australis* Paci-SepT-1. Pacifastin domains are shown in *grey*

Structural Analyses

In order to depict the selection pressures influencing the evolution of venom components, we mapped the sites under positive selection on the homology models created using Phyre 2 webserver (Kelley and Sternberg 2009). Pymol 1.3 (DeLano 2002) was used to visualize and generate the images of homology models. Consurf webserver (Armon et al. 2001) was used for mapping the evolutionary selection pressures on the three-dimensional homology models. GETAREA (Fraczkiewicz and Braun 1998) was used to calculate the Accessible Surface Area (ASA)/solvent exposure of amino acid side chains. It uses the atom co-ordinates of the PDB file and indicates if a residue is buried or exposed to the surrounding medium by comparing the ratio between side-chain ASA and the 'random coil' values per residue. An amino acid is considered to be buried if it has an ASA <20 % and exposed if the ASA is more than or equal to 50 %.

Results

Analysis of coleoid posterior gland cDNA libraries, recovered sequences previously known from one or more coleoid lineages (Fry et al. 2009): CAP (CRiSP/Allergein/PR-1 protein family) (previously sequenced only from cuttlefish and octopus), chitinase (previously sequenced only from octopus), hyaluronidase (known only from its activity in octopus venom but never sequenced), PLA2 (previously sequenced only from cuttlefish), SE-cephalotoxin (previously sequenced only cuttlefish) and serine protease (previously sequenced only from cuttlefish and octopus) (Table 1). The phylogenetic history of these toxin

Table 2 Molecular weights and isoelectric points of pacifastin peptides encoded by the multi-product precursors

Sequence	Domain I	Domain II	Domain III
Coleoid venom	MW/pI		
Hapalochlaena maculosa Paci-Hap-1	8.30/4,095	8.86/4,025	9.59/4,222
Octopus kaurna Paci-Oct-1	8.30/4,124	8.86/4,058	9.10/4,048
Loliolus noctiluca Paci-Lol-1	8.34/3,877	7.77/4,062	8.89/3,886
Sepia pharaonis Paci-Sepea-1	9.15/4,012	8.68/4,055	9.30/4,125
Abdopus aculeatus Paci-Abd-1	4.63/3,956	6.72/3,394	8.30/4,095
		(4 cysteines)	
B6Z1Z0 Hapalochlaena maculosa	4.57/4,081	6.89/3,345	
		(4 cysteines)	
Octopus cyanea Paci-Oct-2	3.89/3,765	6.72/4,122	8.62/3,974
Abdopus aculeatus Paci-Abd-2	5.34/4,140	8.30/3,976	9.42/4,053
Abdopus aculeatus Paci-Abd-3	6.01/4,139	8.30/3,976	9.42/4,053
Octopus cyanea Paci-Oct-3	4.86/4,189	8.84/4,059	9.50/4,197
Octopus cyanea Paci-Oct-4	4.86/4,189	8.84/4,059	9.50/4,197
Sepioteuthis australis Paci-SepT-1	-/-	4.14/4,187	8.66/4,053
Non-venom Proteins			
Q8WQ22	8.67/3,107	8.33/3,184	8.70/3,157
P80060	8.34/3,077	8.65/3,111	
O46162	8.67/3,106	8.33/3,100	
Q8WQ21	7.77/3,014	7.77/3,157	8.73/3,083
Q8MYK4	4.56/3,075	8.33/3,171	8.92/3,183
Q8MYK3	4.56/3,075	8.33/3,171	8.92/3,183
Q95PM3	8.33/3,164	8.96/3,117	
Q4GZT5	8.33/3,164	8.96/3,087	
Q5K4F7	8.33/3,124	7.77/3,089	8.94/3,129

types was previously unclear due to limited taxonomical sampling that had been done in earlier studies. Serine protease was shown to be present in the common coleoid ancestor whilst PLA_2 was recovered only from decapodiforme lineages (cuttlefish and squid). In contrast, chitinase and hyaluronidase transcripts were only identified in octopodiformes. For each protein type, our phylogenetic analyses resolved all coleoid venom gland sequences into a monophyletic group to the exclusion of non-venom gland related sequences, thus demonstrating a shared history of the venom gland forms.

Variable degree of sequence conservation was evident between the different protein types. CAP cysteines were highly conserved whilst the prolines and charged residues were more variable. However, octopus CAP had a basic pI ($\sim$8.44) whilst squid and cuttlefish were acidic to neutral/slightly-acidic pI (5.72–7.40) (calculated using the Expasy pI/MW online service). The globular enzymatic hyaluronidase showed the least variation of all the recovered sequences. Like other type III PLA_2, the coleoid sequences have a lengthy propeptide region, followed by a 10 cysteine arrangement in the processed final form (Fig. 1) The squid sequences, however, lacked the 9th ancestral cysteine, resulting in an odd number of cysteines, which may promote dimerization. The serine peptidase transcripts recovered were the most diverse and numerous of all toxin types recovered (Fig. 2). There is evidence for at least six gene duplication events prior to the divergence of octopus, cuttlefish and squid (Fig. 2). Cuttlefish and squid S_1 proteases had basic pI values, including having the most basic sequences [*S. australis* SP1-SepT-4 (9.39)]. In contrast, in octopoids, the values ranged widely from strongly acidic to basic. Such extreme variations occurred within even a single species and were phylogenetically interleaved, e.g. *A. aculeatus* SP1-Abd-14 [8.07] and *A. aculeatus* SP1-Abd-7 [5.02] being phylogenetically sister sequences to each other. Although 10 cysteines were always conserved in the S_1 proteases, there were considerable variation amongst both prolines and charged residues. Carboxypeptidease encoding transcripts were recovered included from all three coleoid lineages. The carboxypeptidase sequences feature a highly conserved enzymatic core region, with an increase in variability as either terminal is approached, including an uneven number of cysteines (Fig. 3). A number of recovered transcripts featured motifs characteristic of GON_4 domains, alignments show a number of regions of high homology with major deletions in between, these deletions are highly variable in both length and location even within the same species (Fig. 4). However, despite the domain deletions, all GON_4 were basic, with pI values ranging from 8.28 to 9.18.

The greatest degree of variability was displayed by the multiple copies were recovered from all three coleoid lineages of a peptide type sequenced by us in an earlier publication which was phylogenetically unresolvable and simply referred to as ‘orphan 4’ (Fry et al. 2009). We were able to identify them as highly modified versions of the pacifastin peptide family. The pacifastin peptides were revealed to be encoded by a tri-product precursor, with each peptide post-translationally liberated from the others. The sequence previously obtained by us from *H. maculosa* (B6Z1Z0) was shown to contain only the first two domains whilst conversely the *S. pharaonis* sequence from this study was unique in having a fourth domain inserted between ancestral domains two and three (Fig. 5). Cysteines were highly conserved across the domains except domain 1 of *O. cyanea* c259 and domain 2 of *A. aculeatus* c6 and *H. maculosa* (B6Z1Z0). In contrast to the ancestral types, which are strongly basic in all domains (with the exception of the first domains of Q8MYK4, Q8MYK3 from *Schistocerca gregaria*), the coleoid sequences are shown to be highly variable in pI (isoelectric point) (Table 2). The acidic domain 1 forms are not monophyletic so this indicates convergent derivations. Domain 2 was also

Table 3 Molecular evolution of coleoid venom-encoding genes

		SLAC[a]	FEL[b]	REL[c]	Integrative analyses	MEME Sites[d]	PAML[e]	
					SLAC + FEL + REL + MEME		M8	M2a
	$\omega > 1$[f]	0	2	0	3			
CAP	$\omega < 1$[g]	22	63	All	63	1	0	0
	$\omega =$	0.34	–	0.34	–		0.22	0.37
Pacifastin (All Domains)	$\omega > 1$[f]	0	1	3	5			
	$\omega < 1$[g]	19	34	22	37	3	0	0
	$\omega =$	0.47	–	0.58	–		0.36	0.53
Pacifastin								
Domain I $\omega = 0.15$					Domain II $\omega = 0.19$		Domain III $\omega = 0.18$	
	$\omega > 1$[f]	0	0	0	1			
PLA2	$\omega < 1$[g]	4	20	0	20	1	0	0
	$\omega =$	0.56	–	0.74	–		0.52	0.60
	$\omega > 1$[f]	0	0	1	26		1	
Serine Protease	$\omega < 1$[g]	119	139	120	151	26	(0 + 1)	0
	$\omega =$	0.35	–	0.35	–		0.29	0.52

CAP Cysteine-rich secretory proteins, Antigen 5, and Pathogenesis-related 1 proteins, *PLA2* Phospholipase A2

Note

[a] Single Likelihood Ancestor Counting

[b] Fixed-effects likelihood

[c] Random-effects likelihood

[d] Sites detected as experiencing episodic diversifying selection (0.05 significance) by the Mixed Effects Model Evolution (MEME)

[e] Positively selected sites detected using the Bayes Empirical Bayes approach implemented in M8 and M2a. Sites detected at 0.99 and 0.95 significance are indicated in the parenthesis

[f] Number of positively selected sites at 0.05 significance (for SLAC, FEL) or 50 Bayes factor (for REL)

[g] Number of negatively selected sites at 0.05 significance (for SLAC, FEL) or 50 Bayes factor (for REL)

ω mean dN/dS

shown to be variable but to a lesser degree than Domain 1. Domain 3, however, remains in the ancestral basic state. In addition, *S. pharaonis* Paci-Sepia-1 had a fourth domain inserted after domains 2 and 3, which had a pI of 9.13 and a domain molecular weight of 4 kDa. The molecular weights of the coleoid forms (calculated between the first and last cysteines as the N- and C-terminal tails are variable in cleavage points) were consistently approximately 4 kDa for each domain. In contrast, the ancestral forms (also calculated between first and last cysteines) were approximately 3.1 kDa. The differences, however, may reflect a taxonomical trend, rather than a structural state impacting upon function. In contrast to the variability of the highly cysteine cross-linked scaffolds, the globular enzymes chitinase and hyaluronidase conversely showed extreme conservation.

In order to understand the molecular evolution of coleoid toxins, we first employed the conservative one-ratio model (ORM) which estimates a single omega value for the entire phylogenetic tree. It estimated an omega of 0.16, 0.15, 0.42 and 0.22 for the coleoid CAP, pacifastin, PLA$_2$ and serine proteases, respectively, suggesting strong evolutionary conservation (Supplementary Tables 1–4). ORM can only detect positive selection if the average over all the sites along the lineage is significantly greater than one and hence fails to identify sites that are affected by episodic adaptations. We further employed the site-specific selection analyses. Site model 8 highlighted the strong influence of negative selection on the coleoid toxins and computed omega values that ranged widely (Table 3): 0.22, 0.36, 0.52 and 0.29 for the coleoid CAP, pacifastin, PLA$_2$ and serine proteases, respectively. However, this model identified a single codon in the serine protease as evolving under the influence of positive selection. Single likelihood ancestral counting (SLAC), fixed-effects likelihood (FEL), random-effects likelihood (REL) and the evolutionary fingerprint analyses also supported the lack of variation in the coleoid toxins (Table 3 and Supplementary Fig. 1). The aforementioned models for identifying positive selection work best whilst detecting pervasive selection pressures. However, a large proportion of positively selected sites are often subjected to transient or episodic adaptations. When the majority of lineages in a phylogenetic tree follow the regime of negative selection, they mask the signal of positive selection that

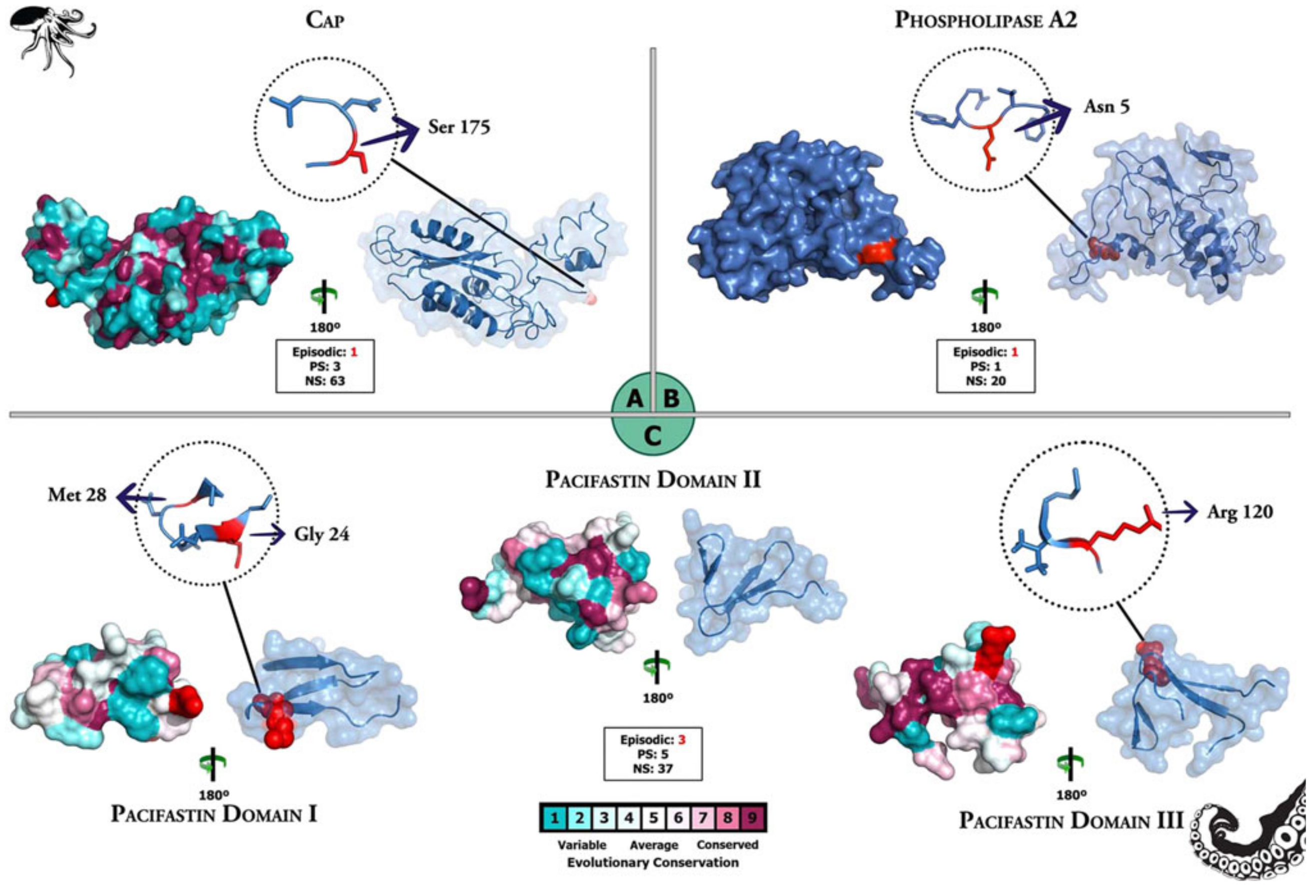


Fig. 6 This figure depicts the molecular evolution of different coleoid venom components. The homology models show the episodically adaptive sites in red (HyPhy, MEME approach). The total number of positively and negatively selected sites detected by HyPhy integrative approach (SLAC and FEL: 0.05 significance; REL 50 Bayes factor) are also indicated. Due to the lack of sequence information ($n = 4$), the molecular evolution of coleoid PLA gene could not be mapped onto its homology model, and hence it is shown in *gray*

might be only influencing a small number of lineages. Hence, these analyses may fail to identify positive selection in such scenarios. To overcome this drawback, we employed the advanced mixed effects model evolution (MEME) (Kosakovsky Pond et al. 2011) which uses the fixed-effects likelihood (FEL) along the sites and random-effects likelihood (REL) across the branches to detect episodic diversifying selection and is capable of not only identifying the episodic adaptations but also the pervasive selection pressures. MEME identified 26 and 3 sites in coleoid serine proteases and pacifastins, respectively which were influenced by the diversifying selection pressures, whilst identifying a single site in PLA_2 and CAP (Table 3 and Fig. 6). Assessment of selection pressures by partitioning different pacifastin domains revealed that they have been extremely constrained by negative selection (domain 1 $\omega = 0.15$; domain 2 $\omega = 0.19$ and domain 3 $\omega = 0.18$).

Mutation of the surface chemistry is one of the prominent characteristics of venom evolution. Estimation of the accessible surface area ratio or the surface accessibility of amino acid side chains revealed that 85 % of the episodically adaptive sites in serine proteases were exposed, whilst only 15 % were buried (excluding the 6 sites that could not be assigned to buried or exposed class), suggesting that most mutations are focussed on the molecular surface (Fig. 7 and Supplementary Table 5). The mutation of the surface chemistry would not only increase the range of receptors these toxins can target but could also help in evading the host immune response. The remaining proportion of sites could not be significantly assigned to either buried or exposed class (Fig. 7 and Supplementary Table 5).

Discussion

The relative timing of recruitment of coleoid toxins has remained unclear due to the only limited sampling that has previously been undertaken combined with the lack of wide recognition that all coleoids share a single common venomous ancestor. Several of these toxins previously only known from one lineage were shown to be in fact basal when more lineages were characterised. We anticipate that

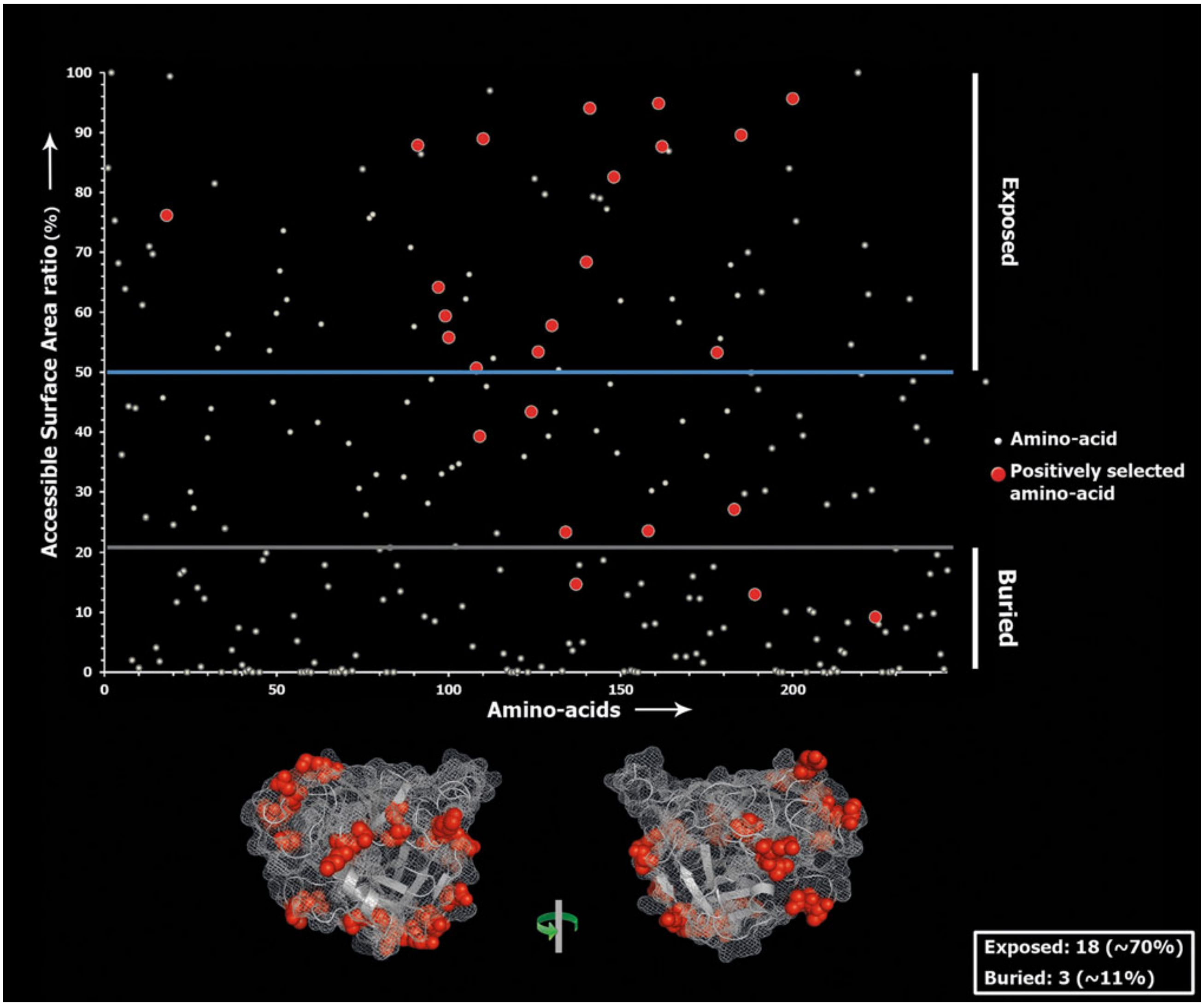


Fig. 7 In this figure, a plot of amino acid positions (*x*-axis) against accessible surface area (ASA) ratio (*y*-axis) indicating the locations of amino acids (exposed or buried) in the crystal structure of coleoid serine protease is presented. Amino acids with an ASA ratio of more than or equal to 50 % are considered to be exposed to the surrounding solvent whilst those with a ratio lesser than 20 % are considered to be buried. Three-dimensional structure of coleoid serine protease is also presented and the episodically adaptive sites (HyPhy, MEME approach) with buried and exposed side chains are indicated by *brown* and *blue* labels, respectively. The sites which could not be assigned to the aforementioned classes are indicated with *white* labels

with further sampling efforts, several other toxin types will also show to have an earlier evolutionary origin than recognised based on current data. Our results also revealed that coleoid venoms are much more diverse than previously anticipated, rivalling in complexity with more intensively studied venoms such as those of snakes. It was also strongly suggested that at least some of these toxins are actively evolving under selection, with the cysteine-rich pacifastin and kallikrein types being particularly abundant and diverse. This is consistent with other venoms, where the components with the greatest cysteine content are the scaffolds most amenable to structural and functional mutations (Casewell et al. 2011; Chang and Duda 2012; Fry et al. 2003; Kordis and Gubensek 2000; Weinberger et al. 2010; Wong and Belov 2012). In contrast, the globular enzymes such as carboxypeptidase and hyaluronidase showed extremely little variation, consistent with globular enzymes from other venoms (Fry 2005) that have a three-dimensional structure driven by non-covalent interactions and a single amino acid change could decimate the correct folding.

Site-specific algorithms can only detect positive selection, when its influence at each site is constant throughout time. Thus, they assume that the diversifying selection affects the majority of lineages in a phylogenetic tree. However, very rarely do we encounter scenarios where there is a constant influence of positive selection across all the lineages. Mixed effects model of evolution (MEME), allows omega to vary not only from site to site but also from branch to branch at a site. Site-specific models suggested that most coleoid venom components were extremely negatively selected, showing very little variation. However, MEME identified certain sites in most coleoid venom components as under the influence of episodic diversifying selection (Table 3; Figs. 6, 7). To provide further support to these results, we assessed the selective influence on the 31 biochemical/structural amino acid properties using TreeSAAP (Woolley et al. 2003). All the sites detected by MEME as episodically adaptive were also identified as positively selected for one or more of the biochemical/structural amino acid properties, providing significant support to the nucleotide-level selection analyses (Supplementary Table 5). Episodic nature of selection has ensured that the molecular scaffold of most coleoid toxins remains extremely conserved over time, whilst allowing subtle accumulation of advantageous changes in certain regions of the toxin.

Serine proteases are known for their diversified biological activities. They are involved in immune responses, cellular differentiation, digestion, complement activation, haemostasis, etc. The presence of serine proteases in the salivary secretions of coleoids might suggest for a digestive and/or prey envenoming role through proteolysis. Using the nucleotide and complementary protein-level selection analyses, we detected as many as 26 episodically adaptive sites in coleoid serine proteases that were accumulating rapid mutations (Table 3 and Supplementary Table 5; Figs. 6 and 7). These variations could enable these aquatic predators to feed on a diverse variety of prey types. In snakes and bees, serine proteases are known to prevent blood coagulation in the bite-victim, through fibrin(ogen)olysis, enabling the rapid spread of other venom components through the blood stream. Serine proteases could thus perform a similar role in coleoids by preventing the coagulation of the blood and enhancing the effects of other venom components.

The recovery of novel protein scaffolds from the glands studied here reinforces how little is known about the protein composition of coleoid venoms. This is underscored by the number and diversity of novel scaffolds recovered despite the relatively limited sampling employed. More extensive sampling will no doubt recover novel isoforms of types identified to date as well as entirely new toxin classes. Of particular, focus for follow-up research should be the structure–function relationships, particularly for the small cysteine knotted peptides such as the pacifastins. It is hoped that these results will stimulate further investigation of these neglected glands and their secretory proteins in an increasingly diverse range of coleoid species.

Acknowledgments BGF was funded by the Australian Research Council, University of Queensland, Herman Slade Foundation and a Universitas 21 Edward Clarence Dyason Fellowship. BGF would also like to thank the University of Hong Kong's Swire Institute of Marine Science for hosting him during Hong Kong fieldwork. SAA was the recipient of postdoctoral fellowship (PDRF Phase II Batch-V) from Higher Education Commission (HEC Islamabad) Pakistan. KS was funded by the PhD grant (SFRH/BD/61959/2009) from Fundação para a Ciência e a Tecnologia (FCT). AA was funded by the project PTDC/AACAMB/, 121301/2010 (FCOMP-01-0124-FEDER-019490) from FCT. EABU acknowledges funding from the University of Queensland (International Postgraduate Research Scholarship, UQ Centennial Scholarship, and UQ Advantage Top-Up Scholarship) and the Norwegian State Education Loans Fund.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Anastasi A, Erspamer V (1962) Occurrence and some properties of eledoisin in extracts of posterior salivary glands of Eledone. Br J Pharmacol Chemother 19:326–333

Armon A, Graur D, Ben-Tal N (2001) ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. J Mol Biol 307:447–463

Cariello L, Zanetti L (1977) Alpha- and beta-cephalotoxin: two paralysing proteins from posterior salivary glands of *Octopus vulgaris*. Comp Biochem Physiol C 57:169–173

Casewell NR, Wagstaff SC, Harrison RA, Renjifo C, Wuster W (2011) Domain loss facilitates accelerated evolution and neofunctionalization of duplicate snake venom metalloproteinase toxin genes. Mol Biol Evol 28:2637–2649

Chang D, Duda TF Jr (2012) Extensive and continuous duplication facilitates rapid evolution and diversification of gene families. Mol Biol Evol 29:2019–2029

DeLano WL (2002) The PyMOL molecular graphics system. DeLano Scientific, San Carlos

Delport W, Poon AF, Frost SD, Kosakovsky Pond SL (2010) Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. Bioinformatics 26:2455–2457

Fraczkiewicz R, Braun W (1998) Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. J Comput Chem 19:319–333

Fry BG (2005) From genome to "venome": molecular origin and evolution of the snake venom proteome inferred from phylogenetic analysis of toxin sequences and related body proteins. Genome Res 15:403–420

Fry BG, Wüster W, Kini RM, Brusic V, Khan A, Venkataraman D, Rooney AP (2003) Molecular evolution and phylogeny of elapid snake venom three-finger toxins. J Mol Evol 57:110–129

Fry BG, Roelants K, Norman JA (2009) Tentacles of venom: toxic protein convergence in the kingdom animalia. J Mol Evol 68:311–321

Ghiretti F (1959) Cephalotoxin: the crab-paralysing agent of the posterior salivary glands of cephalopods. Nature 183:1192–1193

Ghiretti F (1960) Toxicity of octopus saliva against crustacea. Ann N Y Acad Sci 90:726–741

Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11:725–736

Gotz S, Garcia-Gomez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talon M, Dopazo J, Conesa A (2008) High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res 36:3420–3435

Gotz S, Arnold R, Sebastian-Leon P, Martin-Rodriguez S, Tischler P, Jehl MA, Dopazo J, Rattei T, Conesa A (2011) B2G-FAR, a species-centered GO annotation repository. Bioinformatics 27: 919–924

Grisley MS (1993) Separation and partial characterization of salivary enzymes expressed during prey handling in the octopus eledone cirrhosa. Comp Biochem Physiol B 105:183–192

Grisley MS, Boyle PR (1987) Bioassay and proteolytic activity of digestive enzymes from octopus saliva. Comp Biochem Physiol B 88:1117–1123

Grisley MS, Boyle PR (1990) Chitinase, a new enzyme in octopus saliva. Comp Biochem Physiol B 95:311–316

Kanda A, Iwakoshi-Ukena E, Takuwa-Kuroda K, Minakata H (2003) Isolation and characterization of novel tachykinins from the posterior salivary gland of the common octopus *Octopus vulgaris*. Peptides 24:35–43

Kelley LA, Sternberg MJ (2009) Protein structure prediction on the Web: a case study using the Phyre server. Nat Protoc 4:363–371

Kordis D, Gubensek F (2000) Adaptive evolution of animal toxin multigene families. Gene 261:43–52

Kosakovsky Pond SL, Frost SDW, Muse SV (2005) HyPhy: hypothesis testing using phylogenies. Bioinformatics 21:676–679

Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD (2006) Automated phylogenetic detection of recombination using a genetic algorithm. Mol Biol Evol 23:1891–1901

Kosakovsky Pond SL, Murrell B, Fourment M, Frost SD, Delport W, Scheffler K (2011) A random effects branch-site model for detecting episodic diversifying selection. Mol Biol Evol 28: 3033–3043

Lo Bianco S (1888) Notizie biologiche riguardanti specialmente il periodo di maturita sessuale degli animali del Golfo di Napoli. Mitth Zool Stat Neapel 8:385–440

Morishita T (1974) Participation in digestion by the proteolytic enzymes of the posterior salivary gland in octopus–II Isolation and purification of the proteolytic enzymes from the posterior salivary gland. Bull Jpn Soc Sci Fish 40:601–607

Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929–936

Posada D, Crandall KA (2002) The effect of recombination on the accuracy of phylogeny estimation. J Mol Evol 54:396–402

Romanini MG (1952) Osservazioni sulla ialuronidasi delle ghiandole salivari enteriorie posteriori degli Octopodi. Pubbl Staz Zool Napo 23:251–270

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574

Ueda A, Nagai H, Ishida M, Nagashima Y, Shiomi K (2008) Purification and molecular cloning of SE-cephalotoxin, a novel proteinaceous toxin from the posterior salivary gland of cuttlefish *Sepia esculenta*. Toxicon 52:574–581

Undheim EAB, Norman JA, Thoen HH, Fry BG (2010) Genetic identification of Southern Ocean octopod samples using mtCOI. CR Biol 333:395–404

Weinberger H, Moran Y, Gordon D, Turkov M, Kahn R, Gurevitz M (2010) Positions under positive selection—key for selectivity and potency of scorpion alpha-toxins. Mol Biol Evol 27:1025–1034

Wong ESW, Belov K (2012) Venom evolution through gene duplications. Gene 496:1–7

Woolley S, Johnson J, Smith MJ, Crandall KA, McClellan DA (2003) TreeSAAP: selection on amino acid properties using phylogenetic trees. Bioinformatics 19:671–672

Yang Z (1996) Maximum-likelihood models for combined analyses of multiple sequence data. J Mol Evol 42:587–596

Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15:568–573

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586–1591

Yang Z, Wong WSW, Nielsen R (2005) Bayes empirical Bayes inference of amino acid sites under positive selection. Mol Biol Evol 22:1107–1118