https://doi.org/10.1093/molbev/msw106

Article

The defensins consist of two independent, convergent protein superfamilies

Thomas M A Shafee*, Fung T Lay, Mark D Hulett, Marilyn A Anderson

Department of Biochemistry and Genetics, La Trobe Institute for Molecular Science, La Trobe University, Melbourne, Victoria 3086 Australia

* Corresponding author: T.Shafee@latrobe.edu.au

https://doi.org/10.1093/molbev/msw106

Abstract

The defensin and defensin-like proteins are an extensive group of small, cationic, disulphide-rich proteins found in animals, plants and fungi and mostly perform roles in host defence. The term defensin was originally used for small mammalian proteins found in neutrophils and was subsequently applied to insect proteins and plant γ-thionins based on their perceived sequence and structural similarity. Defensins are often described as ancient innate immunity molecules and classified as a single superfamily and both sequence alignments and phylogenies have been constructed.

Here, we present evidence that the defensins have not all evolved from a single ancestor. Instead, they consist of two analogous superfamilies, and extensive convergent evolution is the source of their similarities. Evidence of common origin necessarily gets weaker for distantly related genes, as is the case for defensins, which are both divergent and small. We show that similarities that have been used as evidence for common origin are all expected by chance in short, constrained, disulphide-rich proteins. Differences in tertiary structure, secondary structure order and disulphide bond connectivity indicate convergence as the likely source of the similarity. We refer to the two evolutionarily independent groups as the *cis*-defensins and *trans*-defensins based on the orientation of the most conserved pair of disulphides.

https://doi.org/10.1093/molbev/msw106

Introduction

Defensins are a prominent group of host defence (primarily) peptides that are ubiquitously expressed in most living eukaryotic taxa. They are highly sequence-diverse, but are generally small, cationic and cysteine-rich. The best-characterised families are the plant, invertebrate and vertebrate α- and vertebrate β-defensins.

The term 'defensin' was originally coined in 1985 for rabbit α-defensins in the context of their immune defence role (Ganz et al. 1985; Selsted et al. 1985) and was subsequently applied to other similar protein groups. The first invertebrate defensins were isolated four years later from the haemolymph (equivalent of blood) of bacterially-challenged black blowfly larvae, *Phormia terranovae* (Lambert et al. 1989). These 40-amino acid antibacterial proteins (insect defensin A and B) were named defensins because they shared sequence and functional similarities with mammalian defensins (Lambert et al. 1989). The term "plant defensin" was introduced in 1995 to describe proteins formerly known as γ-thionins when it was noticed that the primary and tertiary structure of antifungal proteins from radish seeds (Rs-AFP1 and Rs-AFP2) were more similar to known insect and mammalian defensins than to the plant thionins (Broekaert et al. 1995; Terras et al. 1995). In 2005, the first fungal defensin (plectasin) was isolated from *Pseudoplectania nigrella* (Mygind et al. 2005). Subsequently, defensins have been identified throughout the vertebrates, arthropods, fungi and spermatophyte plants, as well as in molluscs, cnidarians, annelids and nematodes (Hughes 1999; Froy and Gurevitz 2003; van der Weerden and Anderson 2013; Wu et al. 2014).

Despite little amino acid sequence identity, defensins share remarkably similar tertiary structures that typically feature a triple-stranded antiparallel β-sheet packed against an α-helix and constrained by intramolecular disulphide bridges. Some defensins have highly minimised or elaborated variants of this structure such as the 18-residue cyclic θ-defensins from primates (Li et al. 2014), or the 70-80 residue annelid macins (Jung et al. 2009). The varied sequence and length of the displayed inter-cysteine loops typically determines the protein's function (e.g. NaD1), although some functions depend on residues in the core (e.g. charybdotoxin) (Lay et al. 2012; Banerjee et al. 2013; Poon et al. 2014; Zhu et al. 2014; Baxter et al. 2015). Although innate immunity is the most commonly described function, structurally similar defensin-like proteins have toxin, or signalling activities (Lay and Anderson 2005; Fry et al. 2010; van der Weerden and Anderson 2013).

The extremely divergent sequences, structures, disulphide connectivities and functions of the defensins obscure whether they are all homologous, diverging from a common ancestor, or analogous, converging from multiple independent origins. However, it is notoriously difficult in cysteine-rich proteins to prove when a single fold has originated from convergent evolution or when distinct folds have emerged from extreme divergence (Cheek et al. 2006). It has been proposed that all defensins evolved from a common ancestral precursor before the plant, fungal and invertebrate kingdoms diverged (Hughes 1999; Thevissen et al. 2004; Lehrer 2007; Silva et al. 2014). Some works have also suggested that plant, insect and vertebrate β-defensins are more closely related than the vertebrate α- and β-defensins due to structural similarities (Hughes 1999; Thomma et al. 2002; Thevissen et al. 2004). Moreover, sequence alignments and phylogenies for these defensins have been constructed (Lambert et al. 1989; Broekaert et al. 1995; Hughes 1999; Hoover et al. 2001; Rosa et al. 2011; Semple and Dorin 2012; De Coninck et al. 2013;

Silva et al. 2014; Tassanakajon et al. 2015), primarily based on the alignment of their abundant cysteine residues.

Here, we present evidence that the defensins consist of two evolutionarily independent superfamilies, with extensive structural and functional similarities having arisen by convergent evolution. Sequence evidence of relatedness is statistically insufficient for defensins since short, divergent, cysteine-rich sequences easily exhibit chance resemblance. Analysis of tertiary structure similarity and secondary structure orientation supports the existence of two independent origins. The first superfamily consists of cysteine-stabilised αβ proteins from plants, fungi, and invertebrates. The second consists of the vertebrate α-, β-, θ-, and invertebrate big defensins. These two independent defensin superfamilies show a remarkable number of analogous features in their structure, function and evolution.

https://doi.org/10.1093/molbev/msw106

Results

In order to analyse the relationships between defensin sequences and structures, 2713 defensin and defensin-like sequences were gathered from the non-redundant protein database and structures were gathered from the protein databank. In total, 26 distinct disulphide connectivities were identified, 15 of which are structurally characterised, and 11 which are not (**table 1**). Some motifs are restricted to a single taxon, for example the α-defensins occur only in vertebrates, and the α-toxins only in chelicerates. Conversely, the C8 and C6 defensin motifs occur broadly across multiple animal, plant and fungal taxa (**supplementary table S1**). Additionally, three C-terminal cysteine variants were identified, notable because their location in β-strand constrains the direction of the disulphide bond. A sequence similarity network is sufficient to detect homology between many of these disulphide classes (**fig. 1A**). The network connects classes, such as the C8, C6, petunia, mollusc, plant fusion, fungal, α-toxin, endotoxin and maurotoxins. However, structural similarity is required to detect more distant divergent evolution.

Structural Alignments Detect Two Distinct Groups

Structural similarity is a key method for determination of distant evolutionary relatedness between proteins, since tertiary structure evolves more slowly than either DNA or protein sequence (Orengo and Thornton 2005). Statistically significant structural similarity can therefore be used to support probable homology (Orengo and Thornton 2005). Structural similarity was assessed for defensin structures exemplifying different disulphide connectivities from **table 1** (in addition to platypus and snake toxins with a β-defensin fold). This analysis was performed specifically for mature domains only, since defensin prodomains have no structural information and have been gained or lost on multiple occasions. Pairwise structural alignment revealed that the structures fall into two separate groups with statistically significant within-group similarity to support homology (Z-score). However, between-group similarity is below the threshold of chance similarity (supplementary **fig. S1**). A structural similarity network therefore connects the defensins into two distinct groups, linking together disulphide classes that are otherwise in separate sequence similarity network clusters (**fig. 1B,C**). An exception is the θ-defensins, which are so short that achieving statistically significant similarity to support homology is impossible. The orientation of structural alignments was also consistent within each group, but varied widely when attempted between the groups, further indicating that any inter-group structural similarity is the effect of chance. The structural superposition depends only on residue Cα locations, and so is also insensitive to disulphide connectivity (and indeed insensitive to side chain identity). The structural superposition of disulphide connectivity within each group is indicative that these disulphides are genuinely homologous, since they perform the same structural role.

Structural similarity, therefore, gives no evidence of a single origin for all defensins. Conversely, although it is possible that 'false positive' structural similarities result from multiple origins within each group, there is no direct contradictory evidence suggesting that this is the case. Therefore the maximum-likelihood interpretation is that each group is related by common descent, and the structural evidence of homology supports the existence of two evolutionarily independent groups.

https://doi.org/10.1093/molbev/msw106

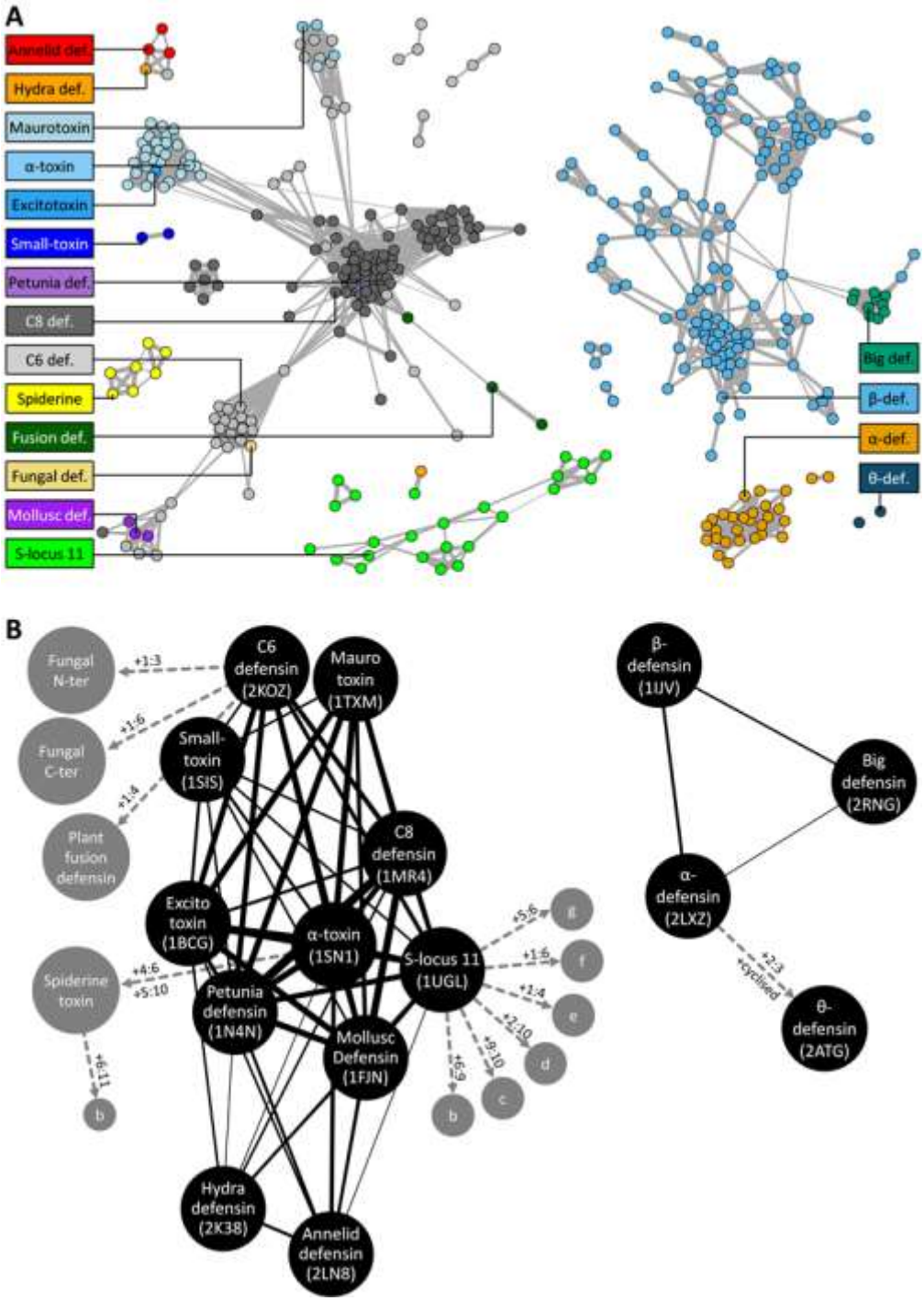

Fig. 1. Networks of sequence and structure relatedness split the defensins into two groups.

(A) A sequence similarity network of defensin sequences generates multiple clusters. Circles indicate sequences, coloured by disulphide class. Grey lines indicate probable sequence homology (widths relative to relatedness E-value). (B) A structural similarity network separates the defensins into two groups. Black circles indicate disulphide classes with solved structures, grey circles represent disulphide classes for which the structure is not yet known. Black lines indicate probable structural homology (widths proportional to relatedness Z-score). Dashed arrows indicate evidence of common origin from cysteine motif (and gene organisation in the case of θ-defensin). Numbers adjacent to arrows indicate additional, unique disulphide bonds of that class.

The Groups Differ by Incompatible Secondary Structure and Disulphide Topology

The failure of structural alignments between the two groups of proteins, despite their superficially similar structures, is a consequence of the orientation and order of their secondary structural elements. From the two separate groups, the C8 defensins and β-defensins have the most similar secondary structure composition (**fig. 2A,B**). However, the analogous secondary structure elements are present in a different order in the primary sequence (**fig. 2C,D**). Attempting to align the tertiary structures by secondary structure is therefore clearly incompatible with published sequence alignments.

The different order and orientation of the secondary structure elements precludes conversion from one fold to the other by simple rearrangements, such as circular permutation, due to their incompatible linkage topologies (**supplementary fig. S2A**). However, two hypothetical multi-step routes do exist for conversion from one fold to the other. The first involves deletion of the N-terminal β-strand along with two disulphides, addition of an internal downward β-strand and connecting upward loop, inversion of the β-sheet twist chirality, and finally deletion of another disulphide and addition of two new disulphides (**supplementary fig. S2B i-iii**). In this case, the only structurally equivalent disulphide would be disulphide 4:7 in the C8 defensin and 1:5 in the β-defensin with other disulphides connecting non-analogous secondary structure elements. The second possibility involves addition of an extra α-helix to the opposite face, conversion of the original, internal α-helix to a loop, deletion of two disulphides, addition of one, and alteration of another (**supplementary fig. S2B iv-vii**). In this case, the only structurally equivalent disulphide would be disulphide 4:7 in the C8 defensin, 3:6 in the β-defensin. Although both series of mutations are possible to link the two folds, there is currently insufficient supporting evidence to assert if either did occur. Additionally, similarly extensive modification would allow conversion to a number of other, distinct CRP topologies.

Evolution of disulphides and secondary structure certainly does occur. For example, compared to the C6 defensins, the C8 defensins have N- and C-terminal extensions with an additional disulphide, and the C10 petunia defensins replace several non-covalent interactions with a further disulphide (Janssen et al. 2003). However larger changes to disulphide connectivity do not appear to be well tolerated (Tanabe et al. 2007; Ojeda et al. 2014). The changes in disulphide connectivity required to convert from one group to the other are similarly extensive to the changes in secondary structure (**supplementary fig. S2B**), and there are currently no structures that suggest that such a transition occurred.

Consequently, although it cannot be ruled out that the two topologies are related by extensive divergent evolution, we currently favour the conservative interpretation that they result from independent origins.

https://doi.org/10.1093/molbev/msw106

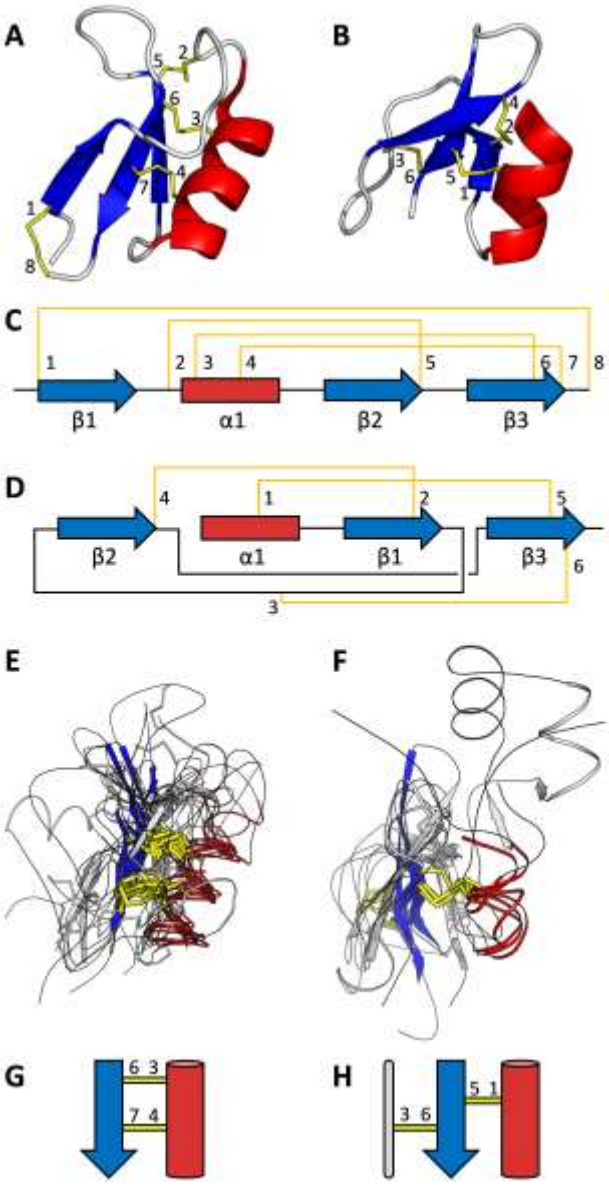

Fig. 2. Disulphide connectivities in the two defensin groups.
(A) The larger group of defensins are typified by the plant C8 defensin NaD1 (PDB:1MR4). (B) The second group of defensins are typified by the human β-defensin HBD1 (PDB:IJV). Aligned analogous secondary structure elements are in a different primary structure order in (C) C8 defensins from the first group, and (D) β-defensins from the second group. (E) A structural alignment of the eleven different disulphide connectivities from the first defensin group and (G) their conserved *cis* oriented disulphides from the C-terminal β-strand that orient in the same direction to bond to the same α-helix. (F) A structural alignment of the four trans-defensins with different disulphide connectivities and snake and platypus toxins with a β-defensin connectivity and (H) their conserved pair of *trans* oriented disulphides from the C-terminal β-strand that orient in opposite directions and so bond to different secondary structure elements. α-helices indicated in red, β-strands in blue, and disulphides in yellow. Secondary structure and cysteines are numbered by their sequence order. Structures were aligned by combinatorial extension.

https://doi.org/10.1093/molbev/msw106

Sequence Alignments Between the Groups can be Surpassed by Random Sequences

Alignment of protein sequences between the two groups should be impossible given the conflicting arrangement of secondary structure elements. However, published alignments often align sequence regions with entirely different secondary structure, orientation and spatial position in the structures. The ability to generate alignments of unrelated sequences arises from their abundance of cysteine residues.

To demonstrate the ease with which unrelated sequences can be erroneously aligned to the defensins, we calculated the probability of finding defensin-like motifs within residue-biased random 50-mer sequences. Residue frequencies calculated for the defensins are strongly biased in favour of cysteines, as well as being somewhat enriched in positive residues and depleted in aliphatic hydrophobic residues compared to average proteins (**fig. 3A**). Broad cysteine motifs were used that encompass 95% of the members in each of the α-defensin, β-defensin and insect C6 defensin families (**table 2**). The C6 defensin motif was restricted to sequences from insects to create a motif with comparably broad inter-cysteine ranges to those of the α- and β-defensins.

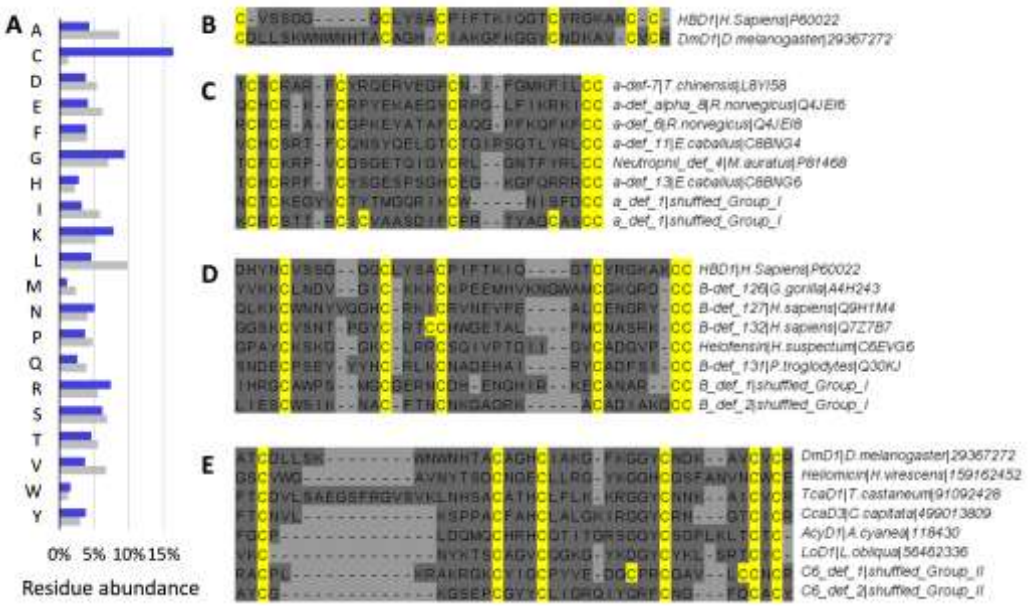

Fig. 3. Unrelated cysteine-rich sequences can be easily aligned.
Relative residue abundance in defensins (blue) compared to the average observed for all proteins in the non-redundant protein database (grey). (B) An alignment of a human β-defensin and Drosophila melanogaster C6 defensin adapted from (Hughes 1999). (C-E) Alignments of the α-defensins, β-defensins, and insect C6 defensins with random sequences. For each alignment, six true defensins from the family were aligned with two sequences that match the disulphide connectivity, but were generated from randomised sequence with defensin residue abundance. Cysteines in yellow, gaps in light grey, all other residues in dark grey.

Firstly, the probability of finding matches to defensin motifs was calculated algebraically. This calculation predicts how often any short, cysteine-rich random protein sequence would be expected to match a defensin motif (**table 2**). Secondly, this prediction was tested heuristically by generating 10,000 random sequences with the relative residue

abundances present in defensins (**fig. 3A**), and searching within them for defensin cysteine motifs (**fig. 3C-E**). These two methods confirm that sequences with defensin-like disulphide connectivities can be generated easily by chance (**fig. 3**). Indeed, an insect C6 defensin has a >2% chance of aligning to a randomised vertebrate defensin better than it aligns to an unrandomised sequence (**fig. 3B**). Consequently, it is impossible for similarities between such short, divergent sequences to statistically support a single origin for the defensins. It is therefore necessary to instead examine features of the proteins that evolve more slowly in order to detect any evolutionary relationships.

Assigning Members to Superfamilies

Given the statistical support for two separate superfamilies of defensins, it becomes necessary to define criteria for assigning sequences to either one. Several lines of evidence indicate relatedness within each superfamily. Within the two defensin groups, the most conserved feature is the orientation and connectivity of the disulphide bonds from the C-terminal β-strand, which is oriented parallel to the α-helix (**fig. 2E,F**). Since the name 'defensin' is well-established in both of these evolutionarily independent groups, we call them the '*cis*-defensin' and '*trans*-defensin' superfamilies (**fig. 2G,H**). In the *cis*-defensins, the CxC spacing causes both disulphides to bond to the same cysteine-stabilised α-helix (3:6 and 4:7 in C8 defensins, **fig. 2G**). In the *trans*-defensins, the CC cysteine spacing constrains the disulphides to orient in opposite directions and bond to different secondary structure elements (1:6 and 3:5 in β-defensins, **fig. 2H**).

All of the *cis*-defensins with solved structures show statistically significant structural similarity (**fig. 1B**). The main differences between their structures are the lengths of inter-cysteine loops and the locations of disulphides other than the conserved pair that define the superfamily (**fig. 4A**). The only exception to this is maurotoxin, which has an aberrant disulphide pairing (**supplementary fig. S3**) but is included in the *cis*-defensins based on its clear structural homology to the mollusc defensin structure (Blanc et al. 1997). Proteins of currently unknown structure (e.g. plant fusion defensins and spiderines) can also be assigned to the *cis*-defensins based solely on their sequence. This is on the basis that they relate to known *cis*-defensins by the addition or removal of a cysteine pair, positioned such that the new disulphide bond falls in a plausible region of the structure (Sachkova et al. 2014).

The four major families of the *trans*-defensins have more distinct structures (**fig. 4B**). Although α- and β-defensins have different disulphide bonding, evidence for their common ancestry by gene duplication comes from their sequential genomic location and gene organisation (Liu et al. 1997). Similarly, the θ-defensins have a unique disulphide bond scheme and cyclic backbone but are clearly descended from the *trans*-defensins as their mature domain is processed from within an α-defensin (Li et al. 2014).

https://doi.org/10.1093/molbev/msw106

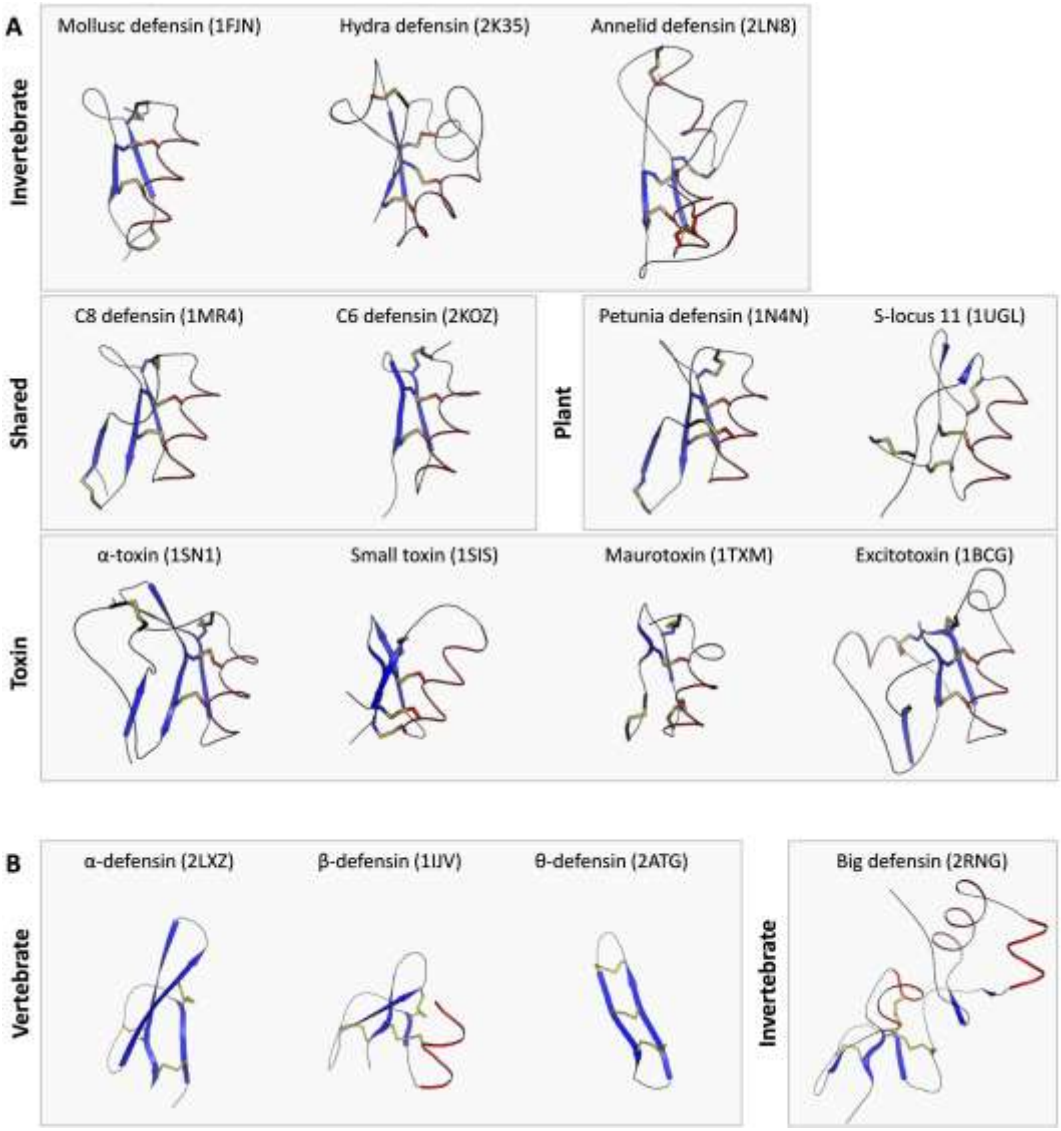

Fig. 4. Structures of *cis*- and *trans*-defensin disulphide connectivities.
Representative examples of (A) 11 distinct cis-defensins and (B) four trans-defensins with distinct disulphide connectivities for which structures have been solved. Within these disulphide frameworks, inter-cysteine loop length and orientation varies widely. PDB identifiers indicated in brackets, α in red, β-strands in blue and disulphides in yellow.

Separating the two superfamilies enables the generation of two sequence alignments that align homologous cysteines that occur in equivalent positions in related structures (**fig. 5**). These alignments illustrate the range of cysteine motifs within each superfamily and the variation of inter-cysteine loop lengths. The *cis*-defensins are on average longer and more diverse in length than the *trans*-defensins, but both superfamilies have similar distribution of net charges and hydrophobicities resulting from their similar amino acid biases (**supplementary table S2**). The *cis*-defensins are also more widely distributed throughout the eukaryotes, with the *trans*-defensins found only in animals, and mostly in vertebrates (**supplementary table S1**).

https://doi.org/10.1093/molbev/msw106

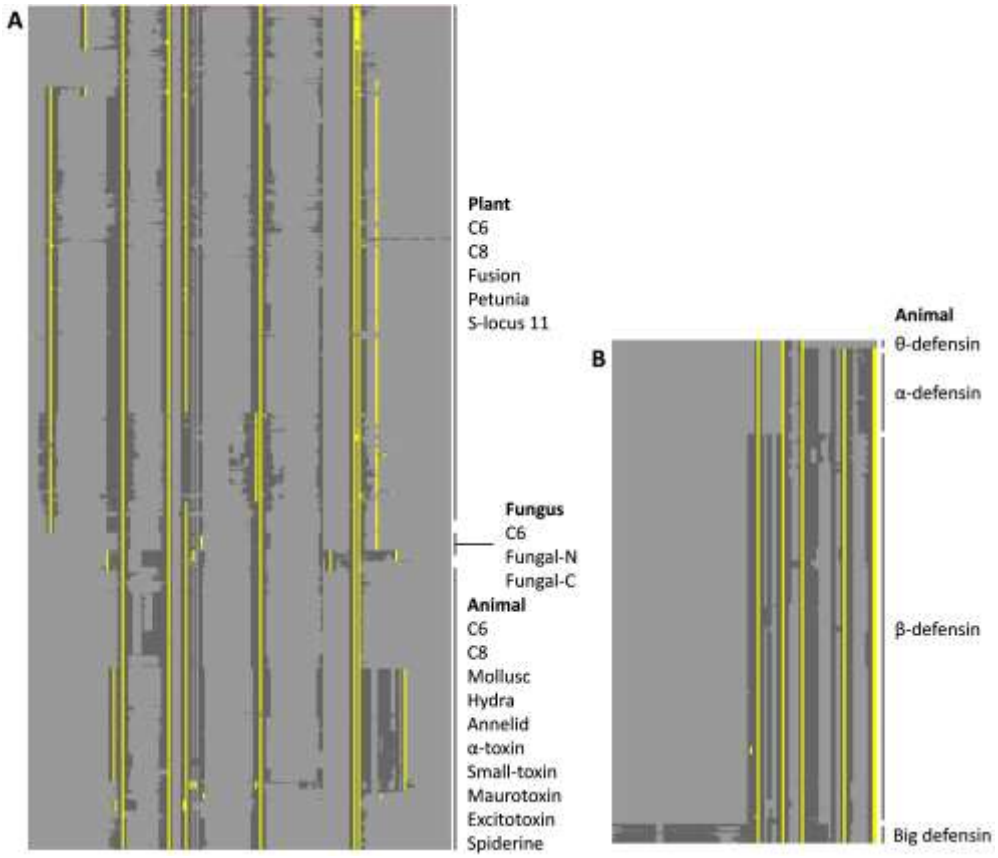

Fig. 5. Defensin superfamily sequence alignments.
Alignments based on structures and sequences of (A) 1820 members of the cis-defensin arranged by kingdom and (B) 839 of the trans-defensin superfamilies. Sequences were aligned by first barcoding homologous cysteines based on known structures, to force homologous loops to align. Dark grey indicates sequence, light grey indicates gaps, and yellow indicates cysteine

Other Convergent Structures

Structural similarity searches using the plant defensin NaD1 as a query demonstrate how commonly the fold occurs in other unrelated proteins. Amongst these are the macrophage receptor with collagenous structure (MARCO) (Ojala et al. 2007), stomagen hormone (Ohki et al. 2011; Takata et al. 2013) and bubble protein of unknown function (Olsen et al. 2004; Seibold et al. 2011). These proteins structurally align to NaD1 with root mean squared deviations below 5 Å and contain some analogous disulphides (**supplementary fig. S4A**), but the similarity is below that expected by chance (**supplementary fig. S4B**). Additionally, differing gene architectures and disulphide connectivities mark them as unlikely to share a common ancestor.

There are also several molecules of contested relatedness to defensins. Nematodes and sponges express antibacterial factors (e.g. AsABF) that share some similarities in length, cysteine spacing, and charge (Zhang and Kato 2003; Froy 2005; Wiens et al. 2011; Tarr 2012), but their structures differ greatly from known defensins. Similarly, bacterial defensin-like proteins have also been reported (e.g. AdDLP), but these have only two disulphides and no other sequence similarity (Zhu 2007). These convergent protein families further exemplify how the arrangement of the disulphide-stabilised β-sheet and α-helix is an optimal fold that has been reached independently a number of times.

https://doi.org/10.1093/molbev/msw106

Discussion

Analysis of the primary, secondary and tertiary structures of the defensins revealed that the maximum likelihood scenario is the existence of two independent superfamilies that are no more related to one another than to any other cysteine-rich proteins. The term 'defensin' is therefore not an evolutionary classification, and we suggest the names *cis*-defensin and *trans*-defensin for the two evolutionary superfamilies. Protein sequence alignments and tertiary structure similarity have been used previously as evidence of relatedness between defensins from vertebrates, invertebrates and plants (Lambert et al. 1989; Broekaert et al. 1995; Hughes 1999; Hoover et al. 2001; Rosa et al. 2011; Semple and Dorin 2012; De Coninck et al. 2013; Silva et al. 2014; Tassanakajon et al. 2015). However, these approaches lack sufficient information to statistically support a single origin for all defensins. If the two superfamilies do share a discrete common ancestor, it is lost to evolutionary history and is beyond the limit of detection by current techniques.

The larger group, *cis*-defensins, contains plant, fungal and the majority of invertebrate defensins. The smaller group, *trans*-defensins, contains vertebrate defensins and invertebrate big defensins. Assigning a member to either superfamily is most accurately done via its tertiary structure and conserved disulphides, requiring statistically significant similarity and the presence of the characteristic disulphide topology. In the absence of structural information, a sequence can be assigned if it matches one of the defined cysteine motifs of an already assigned member. New cysteine motifs that are inevitably identified (e.g. in the S-locus 11 proteins, (Watanabe et al. 2000)), can be assigned when the new cysteines are merely a modification to the previously known motifs, and fall within the disulphide bonding range in currently known structures.

Evolutionary convergence occurs when genetic architecture allows only a limited set of ideal adaptive responses to a selection pressure (Doolittle 1994). This differs from the more common case of divergent evolution, in which many different solutions are valid responses to a selection pressure. Convergent traits are found in nature at all levels of biology – from organism physiology (Vopalensky and Kozmik 2009) and behaviour (Woodard et al. 2011) to enzyme active site geometry (Buller and Townsend 2013) and mechanism (Bork et al. 1993).

Evolutionary constraints on protein tertiary structures, and the robustness to mutation of protein folds, cause them to be more conserved than sequence (Orengo and Thornton 2005). This same constraint can also cause structural convergence (Lupas et al. 2001; Pearson and Sierk 2005). This effect is negligible for large structures. However, for small structures such as the defensins it can lead to false positives when assessing relatedness (**supplementary note 1**). Indeed, the stabilisation of an α-helix by a pair of disulphide bonds to a β-strand has evolved repeatedly, with families displaying all four possible orientations of helix and strand (Tamaoki et al. 1998). Similarly, the sequence similarities between the two defensin superfamilies stem from their short length, divergent sequence, insertion and deletion rate, and skewed residue abundances. Sequence alignments and phylogenies containing members from both groups are therefore misleading, as the lack of homology gives spurious results.

The *cis*- and *trans*-defensin superfamilies therefore represent a remarkable example of the evolvability of a compact, disulphide-stabilised core with displayed cationic loops, leading to extensive evolutionary convergence of sequence, structure and function.

https://doi.org/10.1093/molbev/msw106

Methods

Structure and Sequence Gathering

Proteins with significant structural similarity to defensins were gathered using DALI (Holm and Rosenstrom 2010). The initial query structures used were the prototypical plant defensin NaD1 (PDB:1MR4) and human defensin HBD1 (PDB:1IJV) (Hoover et al. 2001; Lay et al. 2003). Unique proteins whose structures had Z-scores >2 were collected and used as queries in turn until no new structures were identified. The θ-defensin, retrocyclin-2 (PDB:2ATG), was added based on genetic evidence of its relatedness to α-defensins (Nguyen et al. 2003).

Additional defensin and defensin-like sequences were gathered via BLAST searches against the non-redundant protein database using the structurally characterised protein sequences identified above as queries (E-value cutoff <0.005). Additionally, any sequences in Genbank annotated as defensins were collected. Cysteine-rich protein sequences with spurious sequence similarity were removed (e.g. metallothionins and crambins) if the region of identified sequence similarity was less than 25 nt long, had fewer than four cysteines, or had known structural dissimilarity (sequences with associated structures queries using the DALI webserver).

Cysteine Motif and Disulphide Connectivity Identification

Sequences were aligned using the CysBar webserver (Shafee et al. 2016). Briefly, homologous cysteines were identified from the alignment of structures. Homologous cysteines were barcoded, the resulting sequences aligned with ClustalΩ (Sievers et al. 2011), and the barcoded columns restored to their original sequence to generate the final alignment. Sequences were then clustered by the presence of additional cysteines between the homologous cysteines and lengths of the inter-cysteine loops to define the separate cysteine motifs and disulphide connectivities. Disulphide connectivities with three or more members were assigned their own family. If possible, disulphide connectivities with only one or two examples were assigned as modified versions of larger groups if possible (typically with an extra or missing cysteine pair). Unique disulphide connectivities that were not minor modifications of other groups, but that nonetheless conformed to a defensin-like motif, were classified as 'misc' and excluded from further analyses.

Pairwise Structural Alignment Matrix

Structurally characterised disulphide connectivities (11 *cis*-defensins, 4 *trans*-defensins) were analysed using the Protein Structure Comparison, Knowledge, Similarity and Information (*ProCKSI.net*) server (Barthel et al. 2007). The snake and platypus *trans*-defensin toxins (PDBs:1B8W and 4GV5) were additionally included as the most divergent sequences of the β-defensin disulphide connectivity. Pairwise structural alignment of residue Cα atoms was performed by combinatorial extension (Shindyalov and Bourne 1998), with structural similarity Z-scores (standard deviations from the mean) used to infer statistically significant probability of homology. A dendrogram of structural similarity was generated by neighbour-joining hierarchical clustering of the pairwise Z-score matrix to identify the groups for which homology was statistically supported.

Sequence and Structure Similarity Clustering

Networks were generated using the igraph [R] package (Csardi and Nepusz 2006). For sequence similarity, an all-against-all BLAST comparison of a subset of 500 defensin sequences with a cut-off p value of 0.005. An equivalent network of structural similarity was plotted for the set of disulphide class structures with a cut-off p value of 0.005.

Probabilities of Finding Defensin-like Motifs in Unrelated Sequences

The relative abundance of residues for the 11-Nov-2015 release of the UniProtKB/TrEMBL database (54540801 sequences) (Uniprot 2015) was compared to that of the gathered *cis*- and *trans*-defensin sequences (1820 *cis*-defensins, 893 *trans*-defensins).

For each of the α-defensins, β-defensins and insect C6 defensins, the natural variation in inter-cysteine distances within cysteine motifs was used to design a regular expression (sequence motif search pattern) to match 95% of sequences in each family (**table 2**). The probability of finding substrings conforming to these regular expressions in random text strings with biased letter occurrence was performed as described by (Sewell and Durbin 1995). Briefly, the probability of each residue being a cysteine was based on cysteine abundance in the existing defensin sequences; for the *cis*-defensins, *P*(Cys)=0.166 and for *trans*-defensins, *P*(Cys)=0.169. The probability of finding at least one regular expression match was calculated for each regular expression using equation 1 with parameters u=1, n=50 (expanded calculation detailed in supplementary data 1).

$$P(z,u) = \sum_{n,k\geq 0} p_{n.k} u^k z^n$$

Equation 1. Probability that a string contains a particular cysteine motif regular expression

Where z is any match to the regular expression, u is the number of match occurrences, n is the length of the searched string, k is each possible sequence defined by the regular expression ($k \in z$).

The algebraic prediction was tested heuristically by searching for the regular expressions in randomly generated sequences. Two libraries of 10,000 sequences of length 50 were generated with residue abundances based on 1820 *cis*-defensins, or 893 *trans*-defensins using the Sequence Manipulation Suite (Stothard 2000). These 50-residue strings were screened for sequences conforming to defensin disulphide connectivities using the aforementioned regular expressions. Representative sequences were aligned to a set of naturally occurring defensin sequences using ClustalΩ.

Sequence Alignment

Multiple sequence alignments were generated for the 1820 *cis*-defensin sequences and 893 *trans*-defensin sequences using the CysBar webserver and ClustalΩ. Homologous cysteines were barcoded to ensure their correct alignment (columns 11, 44, 65, 73, 107, 150, 152, 160 in the final *cis*-defensin alignment and columns 60, 70, 78, 95, 107, 108 in the final *trans*-defensin alignment). The CysBar webserver was also used to calculate the length, charge and hydrophobicity of each sequence.

https://doi.org/10.1093/molbev/msw106

Acknowledgements

This work was supported by the Australian Research Council (grant number 150104386) and Hexima Ltd.

References

Banerjee A, Lee A, Campbell E, MacKinnon R. 2013. Structure of a pore-blocking toxin in complex with a eukaryotic voltage-dependent K^+ channel. eLife 2:e00594.

Barthel D, Hirst JD, Blazewicz J, Burke EK, Krasnogor N. 2007. ProCKSI: a decision support system for Protein (Structure) Comparison, Knowledge, Similarity and Information. BMC Bioinformatics 8:416.

Baxter AA, Richter V, Lay FT, Poon IK, Adda CG, Veneer PK, Phan TK, Bleackley MR, Anderson MA, Kvansakul M, et al. 2015. The tomato defensin TPP3 binds phosphatidylinositol (4,5)-bisphosphate via a conserved dimeric cationic grip conformation to mediate cell lysis. Mol Cell Biol 35:1964-1978.

Blanc E, Sabatier JM, Kharrat R, Meunier S, el Ayeb M, Van Rietschoten J, Darbon H. 1997. Solution structure of maurotoxin, a scorpion toxin from *Scorpio maurus*, with high affinity for voltage-gated potassium channels. Proteins 29:321-333.

Bork P, Sander C, Valencia A. 1993. Convergent evolution of similar enzymatic function on different protein folds: the hexokinase, ribokinase, and galactokinase families of sugar kinases. Protein Sci 2:31-40.

Broekaert WF, Terras FR, Cammue BP, Osborn RW. 1995. Plant defensins: novel antimicrobial peptides as components of the host defense system. Plant Physiol 108:1353-1358.

Buller AR, Townsend CA. 2013. Intrinsic evolutionary constraints on protease structure, enzyme acylation, and the identity of the catalytic triad. Proc Natl Acad Sci U S A 110:E653-661.

Cheek S, Krishna SS, Grishin NV. 2006. Structural classification of small, disulfide-rich protein domains. J Mol Biol 359:215-237.

Csardi G, Nepusz T. 2006. The igraph software package for complex network research. InterJournal Complex Systems:1695.

De Coninck B, Cammue BPA, Thevissen K. 2013. Modes of antifungal action and in planta functions of plant defensins and defensin-like peptides. Fungal Biol Rev 26:109-120.

Doolittle RF. 1994. Convergent evolution - the need to be explicit. Trends Biochem Sci 19:15-18.

Froy O. 2005. Convergent evolution of invertebrate defensins and nematode antibacterial factors. Trends Microbiol 13:314-319.

Froy O, Gurevitz M. 2003. Arthropod and mollusk defensins - evolution by exon-shuffling. Trends Genet 19:684-687.

Fry BG, Roelants K, Winter K, Hodgson WC, Griesman L, Kwok HF, Scanlon D, Karas J, Shaw C, Wong L, et al. 2010. Novel venom proteins produced by differential domain-expression strategies in beaded lizards and gila monsters (genus *Heloderma*). Mol Biol Evol 27:395-407.

Ganz T, Selsted ME, Szklarek D, Harwig SSL, Daher K, Bainton DF, Lehrer RI. 1985. Defensins - natural peptide antibiotics of human-neutrophils. J Clin Invest 76:1427-1435.

Holm L, Rosenstrom P. 2010. Dali server: conservation mapping in 3D. Nucleic Acids Res 38:W545-549.

Hoover DM, Chertov O, Lubkowski J. 2001. The structure of human beta-defensin-1: new insights into structural properties of beta-defensins. J Biol Chem 276:39021-39026.

Hughes AL. 1999. Evolutionary diversification of the mammalian defensins. Cell Mol Life Sci 56:94-103.

Janssen BJC, Schirra HJ, Lay FT, Anderson MA, Craik DJ. 2003. Structure of *Petunia hybrida* defensin 1, a novel plant defensin with five disulfide bonds. Biochemistry 42:8214-8222.

Jung S, Dingley AJ, Augustin R, Anton-Erxleben F, Stanisak M, Gelhaus C, Gutsmann T, Hammer MU, Podschun R,

Bonvin AMJJ, et al. 2009. Hydramacin-1, structure and antibacterial activity of a protein from the basal metazoan hydra. J Biol Chem 284:1896-1905.

Lambert J, Keppi E, Dimarcq JL, Wicker C, Reichhart JM, Dunbar B, Lepage P, Van Dorsselaer A, Hoffmann J, Fothergill J, et al. 1989. Insect immunity: isolation from immune blood of the dipteran *Phormia terranovae* of two insect antibacterial peptides with sequence homology to rabbit lung macrophage bactericidal peptides. Proc Natl Acad Sci U S A 86:262-266.

Lay FT, Anderson MA. 2005. Defensins - components of the innate immune system in plants. Curr Protein Pept Sci 6:85-101.

Lay FT, Mills GD, Poon IK, Cowieson NP, Kirby N, Baxter AA, van der Weerden NL, Dogovski C, Perugini MA, Anderson MA, et al. 2012. Dimerization of plant defensin NaD1 enhances its antifungal activity. J Biol Chem 287:19961-19972.

Lay FT, Schirra HJ, Scanlon MJ, Anderson MA, Craik DJ. 2003. The three-dimensional solution structure of NaD1, a new floral defensin from *Nicotiana alata* and its application to a homology model of the crop defense protein alfAFP. J Mol Biol 325:175-188.

Lehrer RI. 2007. Multispecific myeloid defensins. Curr Opin Hematol 14:16-21.

Li DY, Zhang L, Yin HD, Xu HL, Trask JS, Smith DG, Li Y, Yang MY, Zhu Q. 2014. Evolution of primate alpha and theta defensins revealed by analysis of genomes. Mol Biol Rep 41:3859-3866.

Liu LD, Zhao CQ, Heng HHQ, Ganz T. 1997. The human beta-defensin-1 and alpha-defensins are encoded by adjacent genes: two peptide families with differing disulfide topology share a common ancestry. Genomics 43:316-320.

Lupas AN, Ponting CP, Russell RB. 2001. On the evolution of protein folds: Are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? J Struct Biol 134:191-203.

Mygind PH, Fischer RL, Schnorr KM, Hansen MT, Sonksen CP, Ludvigsen S, Raventos D, Buskov S, Christensen B, De Maria L, et al. 2005. Plectasin is a peptide antibiotic with therapeutic potential from a saprophytic fungus. Nature 437:975-980.

Nguyen TX, Cole AM, Lehrer RI. 2003. Evolution of primate theta-defensins: a serpentine path to a sweet tooth. Peptides 24:1647-1654.

Ohki S, Takeuchi M, Mori M. 2011. The NMR structure of stomagen reveals the basis of stomatal density regulation by plant peptide hormones. Nat Commun 2:512.

Ojala JRM, Pikkarainen T, Tuuttila A, Sandalova T, Tryggvason K. 2007. Crystal structure of the cysteine-rich domain of scavenger receptor MARCO reveals the presence of a basic and an acidic cluster that both contribute to ligand recognition. J Biol Chem 282:16654-16666.

Ojeda PG, Chan LY, Poth AG, Wang CK, Craik DJ. 2014. The role of disulfide bonds in structure and activity of chlorotoxin. Future Med Chem 6:1617-1628.

Olsen JG, Flensburg C, Olsen O, Seibold M, Bricogne G, Henriksen A. 2004. Solving the structure of the bubble protein using the anomalous sulfur signal from single-crystal in-house Cu K alpha diffraction data only. Acta Cryst D 60:618-618.

Orengo CA, Thornton JM. 2005. Protein families and their evolution - a structural perspective. Annu Rev Biochem 74:867-900.

Pearson WR, Sierk ML. 2005. The limits of protein sequence comparison? Curr Opin Struc Biol 15:254-260.

Poon IKH, Baxter AA, Lay FT, Mills GD, Adda CG, Payne JAE, Phan TK, Ryan GF, White JA, Veneer PK, et al. 2014.

Phosphoinositide-mediated oligomerization of a defensin induces cell lysis. eLlife 3:e01808.

Rosa RD, Santini A, Fievet J, Bulet P, Destoumieux-Garzon D, Bachere E. 2011. Big defensins, a diverse family of antimicrobial peptides that follows different patterns of expression in hemocytes of the oyster *Crassostrea gigas*. PLoS One 6:e25594.

Sachkova MY, Slavokhotova AA, Grishin EV, Vassilevski AA. 2014. Genes and evolution of two-domain toxins from lynx spider venom. FEBS Lett 588:740-745.

Seibold M, Wolschann P, Bodevin S, Olsen O. 2011. Properties of the bubble protein, a defensin and an abundant component of a fungal exudate. Peptides 32:1989-1995.

Selsted ME, Harwig SSL, Ganz T, Schilling JW, Lehrer RI. 1985. Primary structures of three human neutrophil defensins. J Clin Invest 76:1436-1439.

Semple F, Dorin JR. 2012. beta-Defensins: multifunctional modulators of infection, inflammation and more? J Innate Immun 4:337-348.

Sewell RF, Durbin R. 1995. Method for calculation of probability of matching a bounded regular expression in a random data string. J Comp Biol 2:25-31.

Shafee TMA, Robinson AJ, van der Weerden N, Anderson MA. 2016. Structural homology guided alignment of cysteine rich proteins. Springer Plus 5:27.

Shindyalov IN, Bourne PE. 1998. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 11:739-747.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J, et al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7:539.

Silva PM, Goncalves S, Santos NC. 2014. Defensins: antifungal lessons from eukaryotes. Front Microbiol doi: 10.3389/fmicb.2014.00097.

Stothard P. 2000. The Sequence Manipulation Suite: JavaScript programs for analyzing and formatting protein and DNA sequences. Biotechniques 28:1102-1104.

Takata N, Yokota K, Ohki S, Mori M, Taniguchi T, Kurita M. 2013. Evolutionary relationship and structural characterization of the *EPF/EPFL* gene family. PLoS One 8:e65183.

Tamaoki H, Miura R, Kusunoki M, Kyogoku Y, Kobayashi Y, Moroder L. 1998. Folding motifs induced and stabilized by distinct cystine frameworks. Protein Eng 11:649-659.

Tanabe H, Ayabe T, Maemoto A, Ishikawa C, Inaba Y, Sato R, Moriichi K, Okamoto K, Watari J, Kono T, et al. 2007. Denatured human alpha-defensin attenuates the bactericidal activity and the stability against enzymatic digestion. Biochem Biophys Res Commun 358:349-355.

Tarr DEK. 2012. Distribution and characteristics of ABFs, cecropins, nemapores, and lysozymes in nematodes. Dev Comp Immunol 36:502-520.

Tassanakajon A, Somboonwiwat K, Amparyup P. 2015. Sequence diversity and evolution of antimicrobial peptides in invertebrates. Dev Comp Immunol 48:324-341.

Terras FRG, Eggermont K, Kovaleva V, Raikhel NV, Osborn RW, Kester A, Rees SB, Torrekens S, Vanleuven F, Vanderleyden J, et al. 1995. Small cysteine-rich antifungal proteins from radish - their role in host defense. Plant Cell 7:573-588.

Thevissen K, Warnecke DC, Francois IE, Leipelt M, Heinz E, Ott C, Zahringer U, Thomma BP, Ferket KK, Cammue BP. 2004. Defensins from insects and plants interact with fungal glucosylceramides. J Biol Chem 279:3900-3905.

Thomma BP, Cammue BP, Thevissen K. 2002. Plant defensins. Planta 216:193-202.

Uniprot. 2015. Current release statistics. In: Uniprot.

van der Weerden NL, Anderson MA. 2013. Plant defensins: common fold, multiple functions. Fungal Biol Rev 26:121-131.

Vopalensky P, Kozmik Z. 2009. Eye evolution: common use and independent recruitment of genetic components. Philos Trans R Soc Lond B Biol Sci 364:2819-2832.

Watanabe M, Ito A, Takada Y, Ninomiya C, Kakizaki T, Takahata Y, Hatakeyama K, Hinata K, Suzuki G, Takasaki T, et al. 2000. Highly divergent sequences of the pollen self-incompatibility (S) gene in class-I S haplotypes of *Brassica campestris* (syn. rapa) L. FEBS Lett 473:139-144.

Wiens M, Schroder HC, Korzhev M, Wang XH, Batel R, Muller WEG. 2011. Inducible ASABF-type antimicrobial peptide from the sponge *Suberites domuncula*: microbicidal and hemolytic activity *in vitro* and toxic effect on molluscs *in vivo*. Mar Drugs 9:1969-1994.

Woodard SH, Fischman BJ, Venkat A, Hudson ME, Varala K, Cameron SA, Clark AG, Robinson GE. 2011. Genes involved in convergent evolution of eusociality in bees. Proc Natl Acad Sci U S A 108:7472-7477.

Wu J, Gao B, Zhu S. 2014. The fungal defensin family enlarged. Pharmaceuticals 7:866-880.

Zhang H, Kato Y. 2003. Common structural properties specifically found in the CS alpha beta-type antimicrobial peptides in nematodes and mollusks: evidence for the same evolutionary origin? Dev Comp Immunol 27:499-503.

Zhu S. 2007. Evidence for myxobacterial origin of eukaryotic defensins. Immunogenetics 59:949-954.

Zhu SY, Peigneur S, Gao B, Umetsu Y, Ohki S, Tytgat J. 2014. Experimental conversion of a defensin into a neurotoxin: implications for origin of toxic function. Mol Biol Evol 31:546-559.

https://doi.org/10.1093/molbev/msw106

Supplementary Information

SUPPLEMENTARY FIGURES

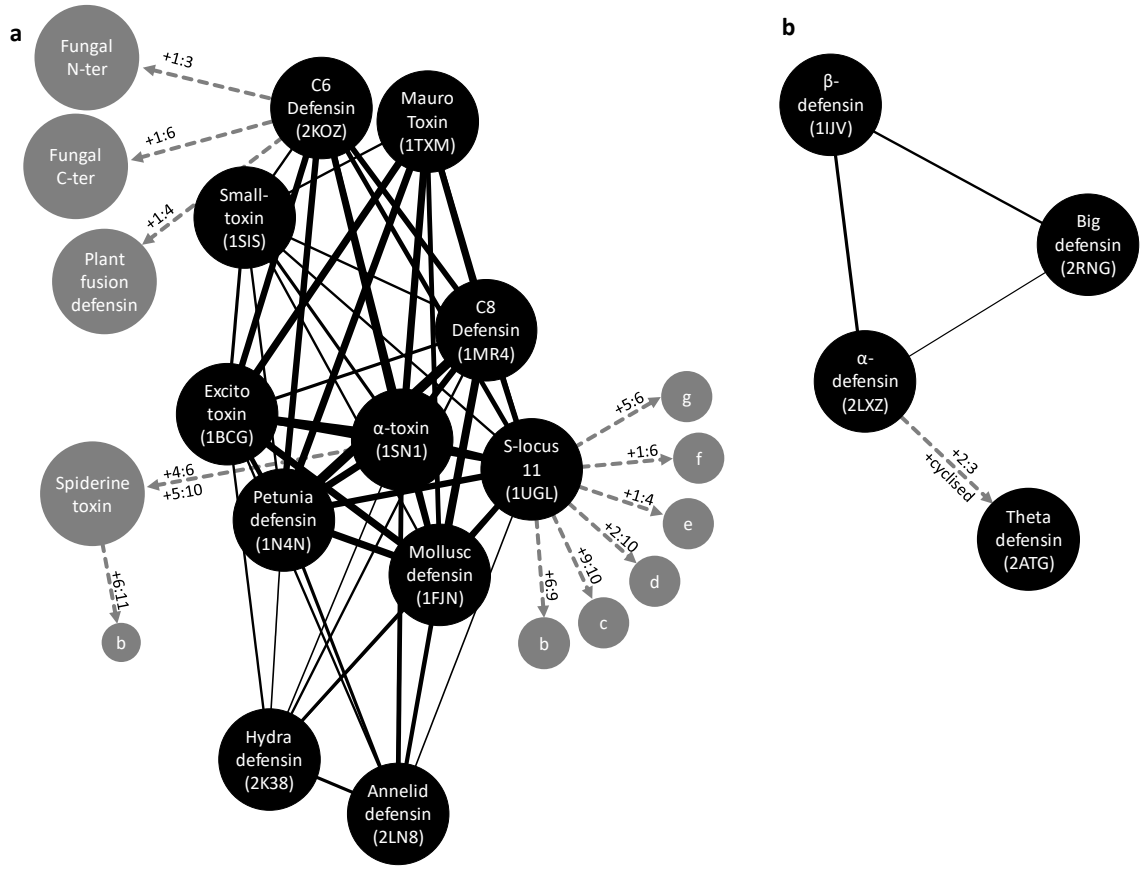

Figure S1 | Network of structural relatedness within each defensin superfamily

A network of (**a**) *cis*-defensin and (**b**) *trans*-defensin structures. Black circles indicate disulphide connectivities with solved structures, grey circles represent disulphide connectivities for which the structure is not yet known. Black lines indicate structural homology (widths proportional to structural relatedness Z-score). Dashed arrows indicate evidence of common origin from cysteine motif (and gene organisation in the case of θ-defensin). Numbers adjacent to arrows indicate additional, unique disulphide bonds of that connectivity family.

https://doi.org/10.1093/molbev/msw106

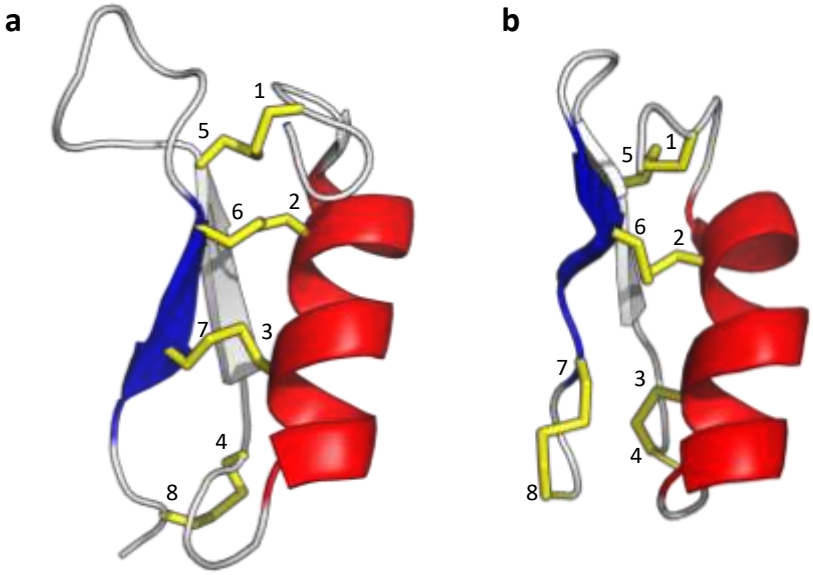

Figure S2 | Aberrant disulphide pairing in maurotoxin compared to mollusc defensins

(**a**) The mollusc defensins pair the disulphides that reside on their α-helix and C-terminal β-strand in the 2:6, 3:7 format that is typical of the *trans*-defensins (PDB:1FJN). (**b**) Maurotoxin, conversely, pairs cysteines C3 and C7 in a unique 3:4, 7:8 format (PDB:1TXM). Coloured with α-helix in red, C-terminal β-strand in blue, disulphides in yellow.

https://doi.org/10.1093/molbev/msw106

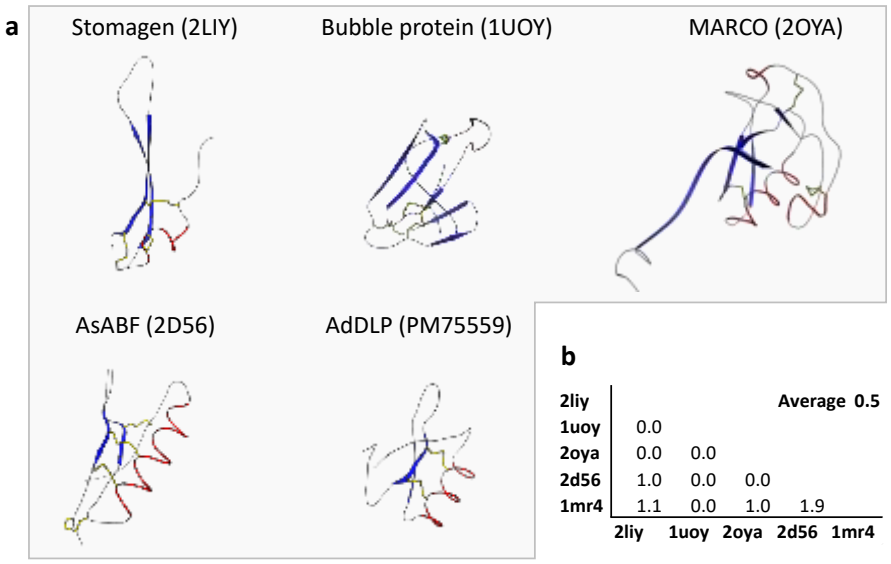

Figure S3 | Proteins showing structural convergence with *cis*-defensins

(**a**) Structures of proteins with convergent features to *cis*-defensins. (**b**) Pairwise matrix of Combinatorial Extension alignment probability Z-scores with averages. The bacterial AdDLP model (PMDB:75559) could not be analysed by the ProCKSI server.

https://doi.org/10.1093/molbev/msw106

Table S1 | Taxonomic distribution of defensin disulphide connectivities

	DESCRIPTION	Taxonomic distribution
15 structurally characterised	C8 defensin	Angiosperm, Gymnosperm, Mollusc, Arthropod
	C6 defensin	Angiosperm, Gymnosperm, Ascomycote, Basidomycote, Glomeromycote, Zygomycote, Mollusc, Insect, Cnidarian, Nematode
	Petunia defensin	Angiosperm
	Mollusc defensin	Mollusc, Arthropod
	Hydra defensin	Cnidarian, Annelid, Mollusc
	Annelid defensin	Annelid, Mollusc
	α-toxin	Arthropod
	Maurotoxin	Arthropod
	Excitotoxin	Arthropod
	Small-toxin	Arthropod
	S-locus 11	Angiosperm
	α-defensin	Vertebrate
	β-defensin	Vertebrate
	θ-defensin	Vertebrate, Arthropod
	Big defensin	Arthropod, Mollusc, Cephalochordate
11 unknown structure	Plant fusion defensin	Angiosperm
	Fungal N-ter defensin	Ascomycote
	Fungal C-ter defensin	Ascomycote
	S-locus 11 variant b	Angiosperm
	S-locus 11 variant c	Angiosperm
	S-locus 11 variant d	Angiosperm
	S-locus 11 variant e	Angiosperm
	S-locus 11 variant f	Angiosperm
	S-locus 11 variant g	Angiosperm
	Spiderine-toxin	Arthropod
	Spiderine-toxin b	Arthropod

https://doi.org/10.1093/molbev/msw106

Table S2 | Defensin protein properties

Distributions for the length, hydropathy, and charge of 1820 *cis*-defensins and 839 *trans*-defensins (mean±SD).

	cis-	***trans-***
Length	47 ± 8	35 ± 10
Hydropathy	-1.1 ± 0.4	-1.0 ± 0.6
Charge	3.3 ± 3.2	3.5 ± 2.3

https://doi.org/10.1093/molbev/msw106

SUPPLEMENTARY NOTE

Note S1 | Bayesian inference of relatedness given similar protein structure

Structural similarity typically indicates common ancestry because the probability of chance similarity is low for average-lengthed proteins. However the defensins (and other cysteine-rich proteins) present an extreme case in which structures are small and highly constrained by their disulphide bonds. In this case, apparently similar structures can occur whilst still being below the threshold for significance.

$$\mathrm{p}(\text{related}|\text{structurally simiar}) = \mathrm{p}(\text{structurally similar}|\text{related}) \cdot \frac{\mathrm{p}(\text{related})}{\mathrm{p}(\text{structurally similar})}$$

The probability p(structurally similar|related) is very high due to conservation of structure in proteins that we know from other evidence to be related. Typically p(structurally similar) is low since most protein structures are large and complex. However, small and stable proteins have a high p(structurally similar) so p(relatedness|structurally similar) is lower than we would expect for an average protein (assuming that p(relatedness) changes little).

https://doi.org/10.1093/molbev/msw106

Table 1. Defensin cysteine motifs

	DESCRIPTION	MOTIF [1]	Connectivity [2]	#CYS	EXAMPLE	ACCESSION [3]
15 structurally characterised	C8 defensin	C-X_{10}-C-X_5-C-X_3-C-$X_{[9-10]}$-C-$X_{[6-8]}$-C-X-C-X_3-C	1:8, 2:5, 3:6, 4:7	8	NaD1	1MR4
	C6 defensin	C-$X_{[5-12]}$-C-X_3-C-$X_{[9-10]}$-C-$X_{[4-5]}$-C-X-C	1:4, 2:5, 3:6	6	NvD1	2KOZ
	Petunia defensin	C-X_3-C-X_5-C-X_5-C-X_2-CC-$X_{[10-11]}$-C-X_6-C-X-C-X_3-C	1:10, 2:5, 3:7, 4:8, 6:9	10	PhD1	1N4N
	Mollusc defensin	C-X_6-C-X_3-C-$X_{[4-5]}$-C-X_4-C-X_8-C-X-C-X_2-C	1:5, 2:6, 3:7, 4:8	8	MgD1	1FJN
	Hydra defensin	C-X_6-C-X_{14}-C-X_3-C-X_9-C-X_6-C-X_8-C-X-C	1:6, 2:5, 3:7, 4:8	8	Hydramacin1	2K35
	Annelid defensin	C-X_6-C-X_{14}-C-X_3-C-$X_{[1-2]}$-C-X_7-C-$X_{[6-7]}$-C-$X_{[5-9]}$-C-X-C-$X_{[11-16]}$-C	1:7, 2:6, 3:8, 4:9, 5:10	10	Theromacin	2LN8
	α-toxin	C-X_3-C-$X_{[5-6]}$-C-X_3-C-X_9-C-$X_{[6-9]}$-C-X-C-$X_{[14-15]}$-C	1:8, 2:5, 3:6, 4:7	8	BmK M1	1SN1
	Maurotoxin	C-X_5-C-X_3-C-X_5-C-X_4-C-X_4-C-X-C-X_2-C	1:5, 2:6, 3:4, 7:8	8	Maurotoxin	1TXM
	Excitotoxin	C-X_{10}-C-X_3-C-X_{10}-CC-X_3-C-X-C-X_{19}-C	1:4, 2:6, 3:7, 5:8	8	Bj-xtrIT	1BCG
	Small-toxin	C-X_2-C-X_{10}-C-X_2-CC-$X_{[5-6]}$-C-X_4-C-X-C	1:4, 2:6, 3:7, 4:8	8	Insectotoxin 15A	1SIS
	S-locus 11	C-$X_{[3-9]}$-C-$X_{[6-7]}$-C-$X_{[3-15]}$-C-$X_{[1-9]}$-C-$X_{[8-9]}$-C-X-C-X-$X_{[3-14]}$-C	1:8, 2:5, 3:6, 4:7	8	S8-SP11	1UGL
	α-defensin	X-C-X-C-X_4-C-X_9-C-X_9-C-X	1:6, 2:4, 3:5	6	HD5	2LXZ
	β-defensin	X_4-C-X_6-C-$X_{[3-4]}$-C-X_9-C-$X_{[5-6]}$-C-X	1:5, 2:4, 3:6	6	HBD1	1IJV
	θ-defensin	X-C-X-C-X_4-C	1^a:3^a, 2^a:2^b, 1^b:3^b	3+3	Retrocyclin2	2ATG
	Big defensin	$X_{[45-51]}$-C-X_6-C-X_3-C-X_{13}-C-X_4-C-X	1:5, 2:4, 3:6	6	TtBigDef	2RNG
11 unknown structure	Plant fusion defensin	C-$X_{[3-5]}$-C-$X_{[4-8]}$-C-X_3-C-$X_{[9-11]}$-C-$X_{[5-9]}$-CCC	1:4, 2:6, 3:7, 4:8	8	MtD36	357449491
	Fungal N-ter defensin	C-X_5-C-X_7-C-X_3-C-X_{10}-C-X_5-C-X_5-C-X-C	1:3, 2:6, 4:7, 5:8	8	Cglosin 1N	88178907
	Fungal C-ter defensin	CC-X_9-CC-X_3-C-$X_{[9-10]}$-C-X_5-C-X-C	1:6, 2:5, 3:7, 4:8	8	Cglosin 1C	88178907
	S-locus 11 b	C-X_9-C-X_7-C-X_{14}-C-X-C-X-C-X-C-X_8-C-X-C-X-C-X-C	1:10, 2:5, 3:7, 4:8, 6:9	10	BoS14	283131299
	S-locus 11 c	C-X_9-C-X_9-C-X_{16}-C-X-C-X-X_9-C-X-C-X_3-C-X_2-CC	1:8, 2:5, 3:6, 4:7, 9:10	10	PtS2	550331862
	S-locus 11 d	C-X-C-X_8-C-X_7-C-X_{15}-C-X-C-X-X_8-C-X-C-X-X_4-C-X_6-C	1:9, 2:10, 3:6, 4:7, 5:8	10	BrS14	90819164
	S-locus 11 e	C-X_9-C-X_6-C-X_7-C-X_6-C-X-C-X-$X_{[8-11]}$-C-X-C	1:4, 2:6, 3:7, 5:8	8	BoS7	283131295
	S-locus 11 f	C-$X_{[5-9]}$-C-X_7-C-$X_{[12-17]}$-C-X-C-X-$X_{[1-2]}$-C-$X_{[6-10]}$-C-X-C	1:6, 2:5, 3:7, 4:8	8	EsS2	557114862
	S-locus 11 g	C-$X_{[9-10]}$-C-$X_{[7-8]}$-C-$X_{[13-17]}$-C-X-C-X-$X_{[9-12]}$-C-X-C-X-$X_{[3-4]}$-C	1:8, 2:4, 3:7, 5:6	8	AtS32	254763280
	Spiderine-toxin	C-X_6-C-X_3-C-X-CC-X_4-C-X-C-X-X_{11}-C-X-C-X_{10}-C	1:7, 2:8, 3:9, 4:6, 5:10	10	Oxotoxin-Ol1b	148877261
	Spiderine-toxin b	C-X_6-C-X_3-C-X-CC-X-C-X-X_3-C-X-C-X-$X_{[10-12]}$-C-X-C-X-X_7-C-X_6-C	1:8, 2:9, 3:10, 4:7, 5:12, 6:11	12	Ctenitoxin-Pn1a	145572742
3 C	+CCC	…-CCC			AtPDF1.3	15225238

https://doi.org/10.1093/molbev/msw106

	+CxCC	…-C-X-CC	DLP96	332659178
	+CxCxC	…-C-X-C-X-C	Fabatin-2	3913646

https://doi.org/10.1093/molbev/msw106

Table 1. Defensin cysteine motifs

[1] Numbers in square brackets indicate the interquartile range of inter-cysteine distances (full ranges are skewed by outliers). [2] Unconfirmed disulphide connectivities are underlined. The θ-defensins consist of cyclised dimers (a+b). [3] PDB identifiers are provided for structurally characterised examples, Genbank identifiers are provided for examples of unknown structures and variant C-terminal cysteine motifs.

Table 2. The occurrence of defensin-like motifs in random 50-mer sequences

	CYSTEINE MOTIF [1]	EXPECTED [2]	OBSERVED [3]
α-defensins	X-C-X-C-$X_{[3,5]}$-C-$X_{[9]}$-C-$X_{[6,10]}$-CC	0.57%	0.42%
β-defensins	$X_{[4]}$-C-$X_{[5,7]}$-C-$X_{[3,5]}$-C-$X_{[8,11]}$-C-$X_{[4,6]}$-CC	1.99%	1.44%
C6 defensins [4]	C-$X_{[6,15]}$-C-$X_{[3]}$-C-$X_{[9,10]}$-C-$X_{[4,7]}$-C-X-C	2.12%	1.69%

[1] For each cysteine motif, a cysteine motif was used to search 50-residue sequence strings generated by randomising defensins from the opposite group. [2] The expected occurrence is the probability of any 50-residue string containing a match to the motif. [3] The observed occurrence is the occurrence of the motif in 10,000 50-residue strings generated with residue abundances based on the defensins. [4] The C6 defensin motif was restricted to insect C6 defensins only.